Abstract

Nonverbal Communication in Socially Assistive Human-Robot Interaction

Henny Admoni

2016

Socially assistive robots provide assistance to human users through interactions that are inherently social. Socially assistive robots include robot tutors that instruct students through personalized one-on-one lessons [197], robot therapy assistants that help mediate social interactions between children with developmental disorders and adult therapists [216], and robot coaches that motivate children to make healthy eating choices [222].

To succeed in their role of social assistance, these robots must be capable of natural communication with people. Natural communication is multimodal, with both verbal channels (i.e., speech) and nonverbal channels (e.g., eye gaze, gestures, and other behaviors).

This dissertation focuses on *enabling human-robot communication by building models for understanding human nonverbal behavior and generating robot nonverbal behavior in socially assistive domains.* It investigates how to computationally model eye gaze and other nonverbal behaviors so that these behaviors can be used by socially assistive robots to improve human-robot collaboration.

Developing effective nonverbal communication for robots engages a number of disciplines including autonomous control, machine learning, computer vision, design, and cognitive psychology. This dissertation contributes across all of these disciplines, providing a greater understanding of the computational and human requirements for successful human-robot interactions.

To focus nonverbal communication models on the features that most strongly

influence human-robot interactions, I first conducted a series of studies that draw out human responses to specific robot nonverbal behaviors. These carefully controlled laboratory-based studies investigate how robot eye gaze compares to human eye gaze in eliciting reflexive attention shifts from human viewers; how different features of robot gaze behavior promote the perception of a robot's attention toward a viewer; whether people use robot eye gaze to support verbal object references and how they resolve conflicts in this multimodal communication; and what is the role of eye gaze and gesture in guiding behavior during human-robot collaboration.

Based on this understanding of nonverbal communication between people and robots, I develop a set of models for understanding and generating nonverbal behavior in human-robot interactions. The first model uses a data-driven approach based in the domain of tutoring. It is trained on examples from human-human behavior, in which a teacher instructs a student about a map-based board game. This model can predict the context of a communication from a new observation of nonverbal behavior, as well as suggest appropriate nonverbal behaviors to support a desired context.

The second model takes a scene-based approach to generate nonverbal behavior for a socially assistive robot. This model is context independent and does not rely on a priori collection and annotation of human examples, as the first model does. Instead, it calculates how a user will perceive a visual scene from their own perspective based on cognitive psychology principles, and it then selects the best robot nonverbal behavior to direct the user's attention based on this predicted perception. The model can be flexibly applied to a range of scenes and a variety of robots with different physical capabilities. I show that this second model performs well in both a targeted evaluation and in a naturalistic human-robot collaborative interaction.

Nonverbal Communication in Socially Assistive Human-Robot Interaction

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

by
Henny Admoni

Dissertation Director: Brian Scassellati

May 2016

Contents

List of Figures

List of Tables

Acknowledgments

PhD dissertations are never produced in isolation. I have benefited tremendously from many mentors, guides, and colleagues during the process of researching and writing my dissertation. There are far too many to enumerate here, but I will attempt to explicitly thank some of them who deserve special mention.

I owe a great debt of gratitude to my advisor and friend, Brian Scassellati (Scaz). Scaz's insight, intellect, and tireless efforts on behalf of his students have made my success possible, and they stand as an example to which I continue to aspire.

All of my co-authors have contributed to the research described in this dissertation, as well as to my personal development, and I am grateful to them all: Scaz, Anca Dragan, David Feil-Seifer, Maja Matarić, Siddhartha Srinivasa, Caroline Bank, Christopher Datsikas, Bradley Hayes, Ahsan Nawroj, Joshua Tan, Mariya Toneva, Daniel Ullman, and Thomas Weng.

Beyond my immediate collaborators, I am thankful for everyone who was a part of the Yale Social Robotics Lab during my time there. You have all taught me something I didn't know before, and I am a better researcher, student, and mentor for my interactions with you. I am specifically grateful to Larissa Hall for her tireless administrative support and much-needed reality checks.

I was a visitor at the Personal Robotics Lab at Carnegie Mellon University for six weeks during the summer of 2013. Thank you to Sidd, Anca, and the rest of PRL for hosting me. The time I spent with you inspired and energized my research.

A few other labs served as secondary homes during my dissertation research. Thank you to the students, postdocs, and PIs of the Yale GRAB lab, USC Interaction lab, and MIT Personal Robots group for the collaboration, and for expanding my academic horizons.

My committee (Greg Trafton, Holly Rushmeier, and Drew McDermott) shared their deep knowledge with me and helped shape this final presentation of my research. Thank you especially to Greg Trafton for his amazing career advice and research guidance since even before he joined my committee. Thank you as well to Fred Shic, who gave me valuable advice as someone who had been through the process.

The Yale Computer Science Department has been my academic home for over six years, and I am so grateful for the support and encouragement from the members of this incredibly smart group of people.

My New Haven friends helped me both within and outside of research. There are far too many people to mention, but if we have shared a meal, a drink, or a climbing night during my time in grad school, please know that it was important to me.

As ever, my family as stood beside me as my strongest and most tireless supporters. Thank you to Marcel, Sara, Sha-har, and Netta Admoni for loving me even when I didn't call.

Several organizations provided financial support for my research. I have been supported by a National Science Foundation Graduate Research Fellowship, the Google Anita Borg Memorial Scholarship, and the Palantir Women in Technology Fellowship. I am also grateful for grant support from the National Science Foundation (0835767, 0968538, 113907), the Office of Naval Research (N00014-12-1-0822), the DARPA Computer Science Futures II program, Microsoft, and the Sloan Foundation.

1

Introduction

The field of Human-Robot Interaction (HRI) strives to enable easy, intuitive interactions between people and robots. Socially Assistive Robotics (SAR), a subfield of HRI, focuses specifically on interactions that help people in social, rather than in physical, interactions [238]. For example, SAR research investigates how robots can act as therapy assistants for children with autism [216], perform one-on-one language tutoring [154], and teach children about nutrition and healthy eating [222].

Such interactions require natural communication. Although verbal communication tends to be primary in human-human interactions, nonverbal behaviors such as eye gaze [27] and gestures [164] can convey mental state, augment verbal communication, and reinforce what is being said [99]. Eye gaze is a particularly important nonverbal signal—compared with pointing, body posture, and other behaviors—because evidence from psychology suggests that eyes are a cognitively special stimulus, with unique "hard-wired" pathways in the brain dedicated to their interpretation [82].

However, producing effective nonverbal communication with robots is still an open problem. Human nonverbal behavior is a complex and dynamic phenomenon. It is highly dependent on context, including the task at hand and the interaction partners. Computational models of nonverbal behavior must accurately represent human

behavior, either by learning appropriate behaviors from demonstration or building models that account for the relevant parts of human cognition, such as visual perception of the scene. Furthermore, the effects of producing nonverbal behavior with *robots*—whose physical appearance and capabilities are different from humans'—have not been fully understood.

This dissertation investigates how eye gaze and other nonverbal behaviors can be used by socially assistive robots to improve human interactions. To understand the effects of nonverbal behavior in human-robot interactions, we first present four studies investigating nonverbal behavior's role in different aspects of interaction, from establishing mutual attention to performing collaborative action. Based on these HRI studies, we develop two complementary models for generating robot nonverbal behavior in a social collaboration.

Though many studies have incorporated nonverbal behavior in HRI, the work described here involves a careful, controlled analysis of how nonverbal behaviors influence specific elements of socially assistive human-robot interactions. The work is interdisciplinary, spanning the fields of Computer Science (artificial intelligence, computational modeling, and robotics) and Cognitive Science (psychophysics and cognitive psychology).

This dissertation begins with a comprehensive report about the current state of the art in eye gaze for social robots (Chapter 2). This review chapter organizes the abundant literature on eye gaze in human-robot and human-agent interaction into three broad categories of research. It also describes terms and concepts used in HRI research, including a discussion of how HRI studies are conducted and evaluated (Section 2.2). This terminology will be used to discuss subsequent chapters in this dissertation.

Chapters 3–6 describe well-controlled HRI studies that target specific aspects of nonverbal communication from robots. Chapter 3 begins this sequence at the finest

level of behavioral analysis, investigating millisecond-level, reflexive responses to robot eye gaze as compared to human eye gaze. The experiment described in this chapter, an extension of a well-studied psychophysical task, reveals that robot eye gaze may not be cognitively processed in the same way as human eye gaze. This provides the foundation for a theme of this dissertation: that successful robot nonverbal behavior depends on the expression of agency. In other words, people must have a reason to attribute meaning to a robot's nonverbal behaviors. This contribution is explored in Section 10.1.1.

In the following chapters, we attempt to identify the specific features of robot nonverbal behavior that lead to successful social human-robot interaction. Establishing mutual attention is first step in coordinated interaction, so Chapter 4 investigates how a robot can use eye gaze to convey attention to an interaction partner. The experiment described in this chapter uses a method that is common in psychophysics but novel to HRI, identifying a target object amidst similar distractors, to isolate the impact of different features of a robot's gaze behaviors on people's perception of the robot's attention. In order to implement the visual search task with real-world robots, we developed a low-cost programmable robotics platform built from modified children's toys. This new robot design is also described in Chapter 4.

After identifying how a robot's attention can best be directed to its interaction partner, we move on to another important task in human-robot interactions: directing a partner's attention to objects in the environment. Chapter 5 describes a study that uses multimodal communication (both speech and gaze) to reference objects. The study also explores what happens when the multimodal behaviors are in conflict, as in the case where a robot names one object and looks at another. The results show that people can use eye gaze to inform object selection but are not hindered by incompatible multimodal behavior. This study underscores the importance of deixis in human-robot collaboration and provides a basis for modeling deictic behaviors

(e.g., looking and pointing to perform object references), which we focus on in the later modeling chapters.

Chapter 6 introduces a different kind of multimodality: two nonverbal behaviors, gaze and gesture, that communicate complementary information. This chapter focuses on handovers, a useful capability in human-robot collaboration. The chapter describes an experiment in which a robot performed gaze cues about desired object locations while handing objects to a human partner, who then selected where to place the objects. The study finds that people can use a robot's eye gaze to inform their choices, but only when an orthogonal nonverbal behavior—the robot's gesture—indicates that the gaze is meaningful.

Chapters 7 and 8 describe models of nonverbal behavior in both human-human and human-robot interactions. First, Chapter 7 describes a data-driven computational model of eye gaze and gestures between a teacher and student in a human-human tutoring interaction. This model, which is trained from empirical data collected in a naturalistic laboratory-based interaction, can predict the communicative context of a nonverbal action (such as question asking, providing a fact, or performing a demonstration), as well as suggest a new nonverbal behavior to match a desired context.

However, the data-driven model is dependent on the collection and annotation of human-human interaction data, which is time consuming and restricts the model to the example domain. Chapter 8 addresses this by introducing a scenario-independent and robot-agnostic generative model of robot nonverbal behavior for human-robot collaborations. This model focuses specifically on deictic gaze and gestures that support verbal object references. It is flexible enough to be applied to a variety of scenarios that require object references, and to a variety of robots with different nonverbal behavior capabilities. Two evaluations are described in this chapter, showing that the model successfully predicts the best gaze or gesture to support a spoken object

reference.

We test the generative behavior model from Chapter 8 in a naturalistic human-robot interaction experiment, described in Chapter 9. The experiment evaluates whether a robot's nonverbal referential behaviors improve people's performance on a construction task in a human-robot collaboration. Results show that nonverbal behaviors are more beneficial for more difficult tasks, improving building accuracy and decreasing time to completion on challenging constructions.

The final two chapters reflect on the original research in this dissertation. Chapter 10 discusses the contributions and impact of this research, identifying some key themes and suggesting open questions for future investigation. Chapter 11 concisely summarizes the dissertation's contributions.

This dissertation contributes to the understanding and development of socially assistive robots. In this dissertation, we:

- Conduct a series of well-controlled studies on the effects of gaze and other nonverbal behaviors in directing attention, establishing mutual attention, referencing objects, and performing object handovers;

- Build a data-driven computational model of human nonverbal behavior in a teaching scenario based on empirical data collected in a naturalistic laboratory study; and

- Develop a complementary model for generating robot nonverbal behavior that is scene-independent and robot-agnostic, and evaluate this robot behavior model in several ways including two in-person human-robot interaction studies.

2

Current State of the Art in Social Gaze for Robots*

This chapter provides background information on HRI and organizes the broad literature on eye gaze in robotics. It begins with a section on concepts and terminology that are important for understanding eye gaze research in the field of HRI, and which will be used throughout this thesis. It also highlights the diversity of research in HRI, including the range of robot platforms, the different approaches of incorporating eye gaze into interactions, and the variety of application domains.

2.1 Introduction

The earliest research into communicative gaze was led by the virtual agent community in the 1990s [60, 243, e.g.]. Virtual agents were imbued with eye gaze as a means for capturing attention, maintaining engagement, and increasing conversational fluidity with human users [59]. Roboticists began introducing meaningful eye gaze into their

*A version of this work is in submission [6]

systems in the late 1990s, in robots such as Cog [213] and Kismet [49].

Modern-day approaches to incorporating eye gaze into human-robot interactions vary widely; research investigating the effects of social eye gaze on human-robot interactions spans the fields of robotics, virtual agents, artificial intelligence, and psychology. Some researchers use robots as stimuli to understand the limits of human perception. Others try to understand the effects of robot gaze by manipulating features of robot appearance and behavior and measuring their influence on human responses. Still others focus on the underlying technologies required for establishing convincing social eye gaze.

This chapter presents the current state of research on social eye gaze in human-robot interaction. To address the large variety of research included in this topic, we divide the corpus of work on gaze in HRI into three broad categories of research. The categories are distinguished both by their goals and by their methods. These categories are:

Human-focused: This research aims to understand how people respond to robots. The emphasis is on human behavior, with the robot serving as a stimulus to provoke a measurable response. This research generally involves well-controlled, laboratory-based studies.

Design-focused: This research focuses on the appearance or behavior of the robot as it affects human responses. Design-focused papers tend to manipulate one feature of gaze at a time (such as the length of fixation) to reveal people's response to that feature, and include both laboratory-based and field-based evaluations.

Technology-focused: This research aims to build computational tools for generating robot eye gaze in human-robot interactions. Though the technologies may be evaluated with human users, this work generally focuses on mathematical or

technical contributions, rather than the effects of the system on the interaction.

The focus of this chapter is *social eye gaze*, any gaze that can be interpreted as communicative by an observer. Social eye gaze includes eye movements that are intentionally expressive, such as gaze aversions that are designed to communicate thoughtfulness. Social eye gaze also includes eye movements that serve a purpose that is not *explicitly* communicative, such as orienting a robot's field of view on an object of interest, as long as these movements are part of an interaction where they might be perceived by other people. Social eye gaze does not include eye movements that are not typically perceived by others during social interactions, such as gaze actions that happen in isolation, viewpoint-stabilization actions like the vestibulo-ocular reflex, or visual processing routines that do not involve changing the camera's point of focus.

Throughout this thesis, we refer to various types of eye gaze using established terminology:

- *Mutual gaze* is often referred to colloquially as "eye contact;" it is eye gaze that is directed from one agent to another's eyes or face, and vice versa. Face-directed gaze without reciprocity is not mutual gaze.

- *Referential gaze* or *deictic gaze* is gaze directed at an object or location in space. Such gaze sometimes occurs in conjunction with verbal references to an object, though it need not accompany speech.

- *Joint attention* involves sharing attentional focus on a common object [169]. It can have several phases, beginning with mutual gaze to establish attention, proceeding to referential gaze to draw attention to the object of interest, and cycling back to mutual gaze to ensure that the experience is shared.

- *Gaze aversions* are shifts of gaze away from the main direction of gaze, which is typically a partner's face. Gaze aversions can occur in any direction, though

some evidence suggests the purpose of the aversion influences the direction of the shift [24].

The type of eye gazes a robot will use in a human-robot interaction will depend on the context and goals of the interaction. Eye gaze can reveal a social robot's mental states, including its knowledge and goals [85]. Gaze can be used by robots to demonstrate their engagement with and attention to a user [238]. Robot eye gaze can increase the fluidity of conversation [163] or direct a user's attention to relevant information in a tutoring setting [128]. However, a tutoring robot may want to express attention to and engagement with a user by performing frequent mutual gaze, while a collaborative assembly factory robot may prioritize task-focused gaze that enables joint attention and object reference.

The remainder of this chapter is organized around the three research categories established earlier: human-focused, design-focused, and technology-focused. First, Section 2.2 provides background about concepts and terminology that are common throughout the diverse studies described in this chapter. The review of current research begins in Section 2.3 with an introduction to gaze in human-human interactions, focusing on findings that are relevant to eye gaze for human-robot interactions. This section introduces insights from psychology that influence the development of gaze for robotics. Section 2.4 discusses human-focused research on gaze in HRI, including human capabilities and limitations when interacting with robots that use gaze communication. Section 2.5 describes design-focused research, specifically how a robot's physical appearance and behavior can be manipulated to elicit effective eye gaze communication within human-robot interactions. Section 2.6 presents technology-focused research, covering the various systems and frameworks for developing robot eye gaze.

2.2 Background

This section describes some common themes found throughout the research on social eye gaze for HRI. In identifying the commonalities, this section also highlights the diversity in this body of work; many different approaches, domains, metrics, and technologies make up the state of the art in social eye gaze for HRI.

2.2.1 Robot appearance

Eye gaze research in HRI is conducted using robots with a wide range of variability in appearance and capability. These platforms range from simple cartoon-like robots to extremely human-like robots and virtual agents.

The differences in gaze capabilities are related to the high cost of implementing eye movements in robots. Each movement along an axis, also known as a *degree of freedom*, must be produced by some motor or other actuator. Adding capabilities means adding actuators, some of which must be quite small (to fit into the robot's head) and powerful (to perform rapid movements like saccades). These requirements drive up a robot's cost, complexity, and fragility. Most social robots attempt to minimize these costs by choosing not to implement some biological capabilities.

Figure 2.1 illustrates the spectrum of biological realism in robot eye gaze. This spectrum is a rough indicator of the range of human-likeness in eyes, in terms of appearance and capability. The extreme right end of the realism spectrum contains humans. Moving leftward on the spectrum indicates descending levels of biological realism, with fewer human-like capabilities such as pupil dilation, ocular torsion, and saccades.

Just to the left of humans on the spectrum are virtual agents, which have the potential for extremely high levels of biological realism. By nature of being animated, virtual agents can mimic human eye capabilities with greater precision than physical

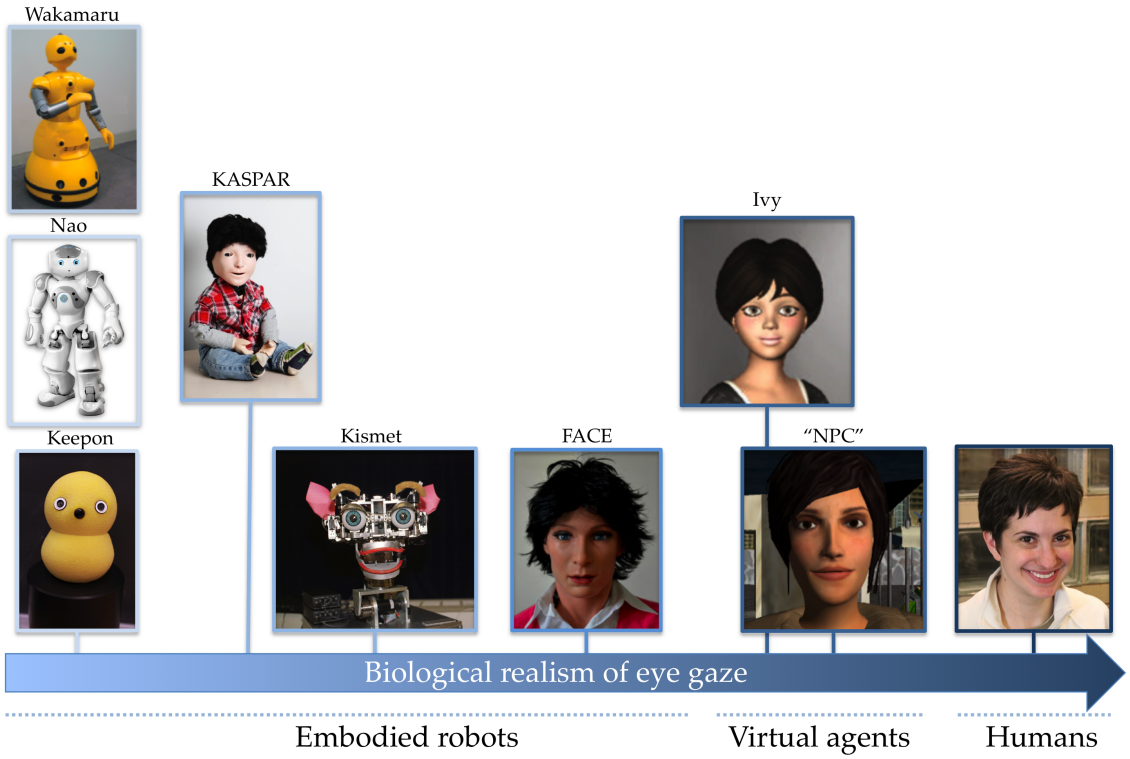


Figure 2.1: Robots and virtual agents with a range of appearances and capabilities are used for gaze research in HRI. This spectrum roughly sketches the range of biological realism with examples drawn from research cited in this review: Wakamaru [236], Nao [15], Keepon (author photograph), KASPAR (courtesy of the Adaptive Systems Research Group, University of Hertfordshire, UK), Kismet [48], FACE [270], Ivy [20], and an NPC [182].

robots, though computationally encoding biologically realistic gaze behavior is an active area of research [206]. While some virtual agents are implemented with complex, biologically faithful models of muscle movement that control eye motion, others use motion generators that are less consistent with the underlying biology [206], so there is a range of possible realism within the virtual agent literature. In Figure 2.1, the virtual agent referred to as "NPC" uses a biologically-based model to animate its saccades, blinks, and gaze shifts [182]. In contrast, the virtual agent called Ivy uses timings of gaze aversions drawn from video-coded observations of human conversation [20].

Moving leftward the spectrum of biological realism, different capabilities are lost.

The most realistic physical robots, for example, do not implement pupil dilation, even though this is an indicator of mental state (such as cognitive effort) in humans [119]. The biologically realistic robot FACE uses a human-like gaze model based on motion capture data from human examples to control the speed and magnitude of eye movements [270].

Less biologically realistic robots retain gaze capabilities but have simpler appearances and gaze control models. Kismet has an independent pan and joint tilt degrees of freedom for each eye, two degrees of freedom for each eyebrow, and independent eyelids, enabling expressive behavior like winking [46]. Still less biologically realistic robots, such as KASPAR [69], have eyes that do not move independently of each other, eliminating the capability to perform lower-level components of biological gaze, such as vergence.

At the leftmost end of the realism spectrum are robots that have fixed eyes. These robots, such as Keepon [144], Nao [15], and Wakamaru [236], are incapable of eye movements that are separate from head orientation, such as the kind people perform when orienting to a lateral visual target [87]. These robots rely on head turns to indicate gaze direction. While this mechanism can be communicative on a gross level, there is evidence that head pose is an inadequate indicator of human gaze direction in human-robot interactions [136].

The variability in appearance and capability of robot eyes is important to note when discussing research on robot eye gaze. Because studies are conducted with different robots, their results may not directly transfer from one robot to another. Each study described in this chapter should be considered in the context of the robot or virtual agent it employs.

2.2.2 Embodiment and virtual agents

Much of the work on social eye gaze emerged from the virtual agents community in the 1990s [60, 243]. This work led the way for embodied gaze research in robotics, and the virtual agent community continues to make advances in the design and understanding of social gaze for intelligent agents [206]. For this reason, virtual agents are presented alongside physically embodied robot systems in this paper. However, there are some notable differences between the two fields.

Virtual agents can provide fine control over the appearance and timing of gaze behaviors, such as subtle eyelid, eyebrow, and eye ball movements. These kinds of fine movements are difficult to achieve with physical motors on embodied robots: it is difficult to assemble enough fast motors in a small enough space to create the level of expressivity in the human face. Though some hyper-realistic humanoid robots—such as Geminoid [208] and FACE [270]—strive to achieve human-like face actuation, most do not achieve the level of facial expressiveness available in animated characters. Therefore, virtual agents provide a platform with which to study the effects of well-controlled, subtly expressive motions of social eye gaze.

There is disagreement, however, on whether physically embodied systems elicit different responses than animated agents or even video representations of those same physical systems. Some researchers have found that physically embodied systems improve interactions over virtual systems. Children spend more time looking at a robot tutor that is physically embodied than at a virtual representation of that robot [135], and adults retain lessons about a cognitive puzzle better when they've been tutored by a physically embodied robot than by a video representation of that robot [153]. People also fulfill unusual requests from a robot more frequently when that robot is physically embodied than when it is tele-present [31]. Physically embodied agents are rated more positively [196, 259] and attributed greater social presence [151] than their virtual or tele-present counterparts.

However, not all research has supported the benefit of physical embodiment over virtual presence. In a tutoring interaction involving sorting, children fail to show differences in learning from embodied and virtual robots [135]. In an interaction with a health care robot, people remembered less information provided by a physically co-located robot than information provided by a virtual representation of that robot [196].

Research on embodiment to date has not specifically focused on the effect of embodied social gaze (see Section 10.2.3 for how this question might be addressed). Whether or not embodiment affects an interaction, research on both virtual agents and physically embodied robots is important for understanding social gaze for intelligent agents, and both the virtual agents and robotics communities have made important contributions to our understanding of eye gaze in human-agent interaction.

2.2.3 Study locations and controls

Human-robot interactions can be evaluated both inside and outside of the laboratory. Laboratory-based and field-based studies have complementary benefits and limitations, making both important in the investigation of eye gaze in HRI. Based on the location of the study, researchers can control the environment and potential confounding variables to a greater or lesser degree. The trade-off for increased control is a decrease in the generalizability of the research findings to real-world settings.

Laboratories provide well-controlled environments in which to perform highly repeatable, consistent experiments. The laboratory can be outfitted with sensors to capture a variety of experimental data, including cameras for video, motion capture systems to detect body positions, and eye trackers for precise gaze analysis. Laboratory-based studies are particularly well-suited to research that systematically manipulates a variable to understand its effect on an interaction, because they can exclude potential confounding factors by rigidly controlling the environment. However,

laboratory-based studies are limited in their ecological validity, because the controlled and restricted environment does not necessarily represent how robots will operate in the real world.

Field-based studies involve placing robots in naturalistic environments, such as shopping malls, hospitals, and building atriums. Interactions tend to be more free-form because the circumstances of the interactions cannot be precisely predicted or controlled. Data collection is often more limited than in laboratory-based studies, and tends to be more observational than empirical. However, these types of studies can more accurately reveal people's interactions with robots "in the wild."

There is a spectrum of study types between these two extremes. For example, laboratories can be augmented with furniture or people to manufacture a more realistic setting. Sensors can be arranged in the field for additional data collection, sometimes at the cost of slightly more inhibited interactions. This dissertation cites references across this spectrum of study types, from carefully-controlled laboratory research to long-term deployments in unpredictable human environments.

2.2.4 Evaluation metrics

When evaluating the effects of gaze on human-robot interactions, both objective and subjective metrics can provide useful information. Which evaluation metric is used depends on the interaction task and the research goals. This section provides an overview of the many objective and subjective measures used in research on gaze in HRI, with some specific examples of each.

Objective Measures

Objective metrics quantify observable behavior. The behaviors that these metrics quantify can range in scale from millisecond-level actions to hours-long performances. The unifying factor is that objective measures address data that can be observed and

quantified.

Precise measurements can reveal low-level (and not necessarily conscious) responses to robot gaze. For example, measuring millisecond-level response times to a robot's directional gaze [2] or recording tiny eye saccades with an eye tracker [269] can reveal underlying differences between people's responses to robots and humans.

Larger-scale measurements can quantify a robot's effect on longer-term human behavior. For example, how well a robot's referential gaze facilitates understanding of object references can be measured by how long it takes a user to select the correct object [3, 44, 47]. The effectiveness of a robot tutor's gaze behaviors can be revealed by the amount of information a user is able to recall from the interaction [23, 236]. Information recall can also act as a proxy for attention: if participants pay more attention, they can recall more information, so measuring recall reveals how much attention different robot gaze behaviors elicit from people [113, 172].

Some objective measures involve post-hoc interpretation of human behavior, often accomplished through video coding. This process entails careful analysis of a recorded interaction to evaluate users' responses to a robot's gaze behaviors, in terms of pre-defined items like engagement behaviors [132], the conversational function of utterances [20], or the use of body language [116]. Because these post-hoc interpretations may be subject to the coder's perceptions and biases, these interpretations are often coded by two or more individuals, with correlations confirmed by statistics like Cohen's κ-coefficient [67].

Objective evaluations can also be applied to the robot systems themselves. For example, the success of a robot gaze system can be measured by whether a robot can predict the correct speaker [248, 254] or influence human users into certain conversational roles [175].

Subjective Measures

Subjective measures can provide insight into user experiences that may not be outwardly observable. Subjective measurements typically involve collecting user perceptions and opinions through surveys and interviews.

The most common type of subjective measure for studies investigating social eye gaze in HRI is a survey or questionnaire, often provided to users at the end of an experiment [20, 65, 116, 223, 248, e.g.]. Survey questions are often formulated as Likert scales, through which participants reveal their perceptions and opinions by indicating their strength of agreement or disagreement with selected statements. For example, to evaluate how well gaze behaviors make a robot seem like a positive interaction partner, these scales measure characteristics like intelligence, animacy, and likability [34]. Subjective measures can also include direct evaluations of a robot's behavior. For example, to evaluate how well a robot can convey emotions by changing its eye and facial expressions, a user might be asked to identify what emotion the robot is conveying for various expressions [156].

Interviews are another tool for eliciting subjective feedback from users. Interviews can reveal, for example, children's subjective impressions of a robot tutor [207]. Interviews can also reveal whether users consciously observed the manipulations in the study [4, 271]. Though not specifically evaluating user responses to robot behavior, these questions allow researchers to identify whether the effects of the experimental manipulation were perceived or not.

Objective and subjective measures provide complementary approaches for evaluating the effects of robot gaze in human-robot interactions. The field of HRI uses a diverse set of measures, and understanding the role of these different types of metrics is important for interpreting the research in the field.

2.3 Gaze in Human-Human Interactions

Gaze is important to human-human interactions because it is closely tied to what people are thinking and doing. People use their observations of others' eye gaze to guide everything from conversation [140] to speech [28] to attention [92]. In this section, we draw out specific research findings from psychology that have a direct impact on the design of social eye gaze for human-robot interaction. The findings described in this section are applied to research that is described throughout this dissertation. The studies in this section are aligned into three general topics:

- How people use eye gaze for conversation and speech (highlighted in Sections 2.5.1 and 2.5.2)
- How people use eye gaze when they refer to and manipulate objects (highlighted in Sections 2.5.3 and 2.5.4)
- Methods for testing people's responses to eyes and faces (highlighted in Section 2.4.2)

2.3.1 Gaze for conversation and speech

People generally look at what they are attending to. For example, in conversations, gaze predicts the target of conversational attention. When someone is listening, the person they are looking at is likely the person being listened to (88% of the time) [254]. Similarly, when someone is speaking, they are often looking at the target of their speech (77% of the time) [254]. Other studies confirm these numbers, with gaze directed at conversational partners approximately 80% of the time [58].

During conversation, eye gaze signals when a speaker wants to maintain or relinquish the floor, indicates cognitive effort, and balances attention with intimacy [24]. Researchers have extracted very specific timings for the gaze cues that are part of

conversation [20, 183]. For example, speakers establish mutual gaze approximately 2.4 seconds before relinquishing the floor. Intimacy modulating gaze aversions tend to be short (between 1 and 2 seconds), while gaze aversions that signal cognitive effort (such as looking away while beginning a response to a question) are longer, at about 3.5 seconds [20].

Pairs of people tend to hand off control of the conversation via gaze cues. For example, the "reference-action sequence"—in which an instructor refers to an object and then a worker acts on that object—can be divided into five cyclically repeating phases, each with their own distinct gaze behaviors: pre-reference, reference, post-reference, action, and post-action [19]. A worker's gaze tends to follow the instructor's gaze in the early and late phases, while the instructor's gaze tends to follow the worker's behaviors during the middle phases (post-reference and action) while the worker performs the task [19].

In addition to managing interpersonal interactions, gaze also relates directly to the syntax of speech. People often look away from their partner when beginning the *theme* of the sentence (which indicates what the sentence is about) and look toward their partner when beginning the *rheme* of the sentence (which provides information or exposition about the theme) [60].

The conversation topic also influences gaze. When guiding a tour, people look between the exhibit and their audience, among other nonverbal behaviors, and these behaviors elicit engagement responses from the audience [264]. People show less mutual gaze when their conversation involves high levels of intimate self-disclosure [131]. Two partners' nonverbal behaviors, including their eye gazes, can be used to extract the context of an utterance during an interaction, such as conveying a fact or answering a question [8] (Figure 2.2).

Gaze durations during conversation are affected by people's personalities, as well. Extroverts spend more time looking at their partner than introverts [21]. People

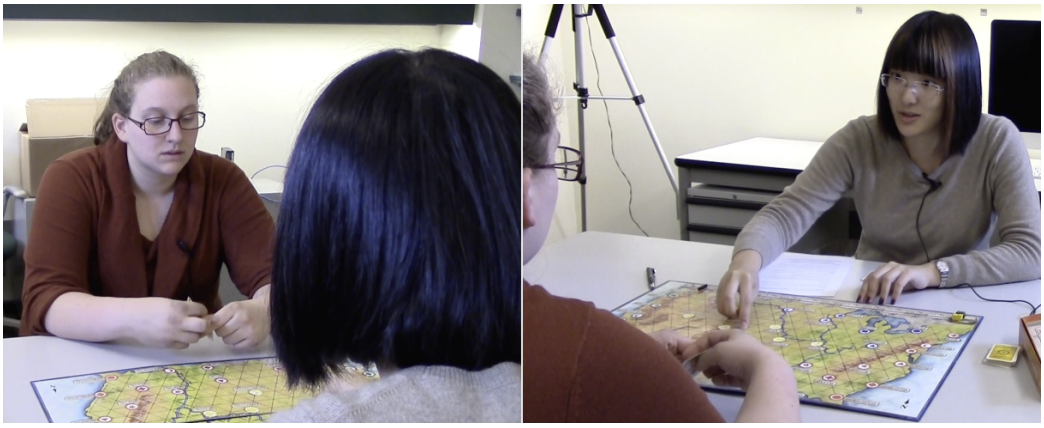


Figure 2.2: The gaze dynamics of student (left) and teacher can reveal the context of an interaction. Here, the teacher is conveying a spatial reference, performing face-directed gaze at the student while the student observes the deictic gesture [8].

are also more likely to speak when their conversational partner looks at them more often [253]. However, the amount of mutual gaze between conversational partners depends on the interpersonal dynamics between the partners, not only on each partner's individual traits [52, 58].

2.3.2 Gaze for object reference and manipulation

Eye gaze is an important part of communicating about the environment. When referring to objects or locations around them, people's gaze is closely tied to the content of their speech. Objects are typically fixated one second or less before they are named [100, 269], though this may be slightly longer when speakers must search for the object [19].

Teams of people use eye gaze as a subtle, non-intrusive channel of communication [219]. When partners refer to objects or locations in the environment, people use their partner's eye gaze to predict their partner's next verbal object reference, and can more quickly respond to that reference [44]. In contrast, when access to a partner's eye gaze is restricted, people are slower at responding to their partner's referential communication [44]. If there is ambiguity in the object reference, gaze is a strong

and flexible cue for eliminating uncertainty about referential expressions [104, 232]. Chapter 5 describes novel research that shows the benefit of gaze in augmenting spoken object references and examines what happens when speech and gaze references are misaligned.

When people are manipulating objects, their eye gaze is similarly tied to their task and intended action. Eye gaze typically reaches the object of interest before any movement of the hands has started [148]. Though people fixate the same object while they act on it, eyes often shift to the next object in the task sequence before the action is completed on the current object [148]. These shifts of gaze to a new object often correspond to the start of a significant kinematic event on the current object; for example, gaze directed at an object to be grasped will shift away from that object just as the hand closes around it [127]. Objects not related to the task at hand are rarely fixated [106].

Gaze is also used to signal availability for interaction. Caregivers in a nursing home demonstrate their availability to their patients through broadly distributed gaze, and people naturally wait for caregivers to establish mutual gaze before requesting assistance [265]. When they must pass objects back and forth, object handovers between people rely on a receiver signaling readiness to receive an object by gazing at their partner [234, 235]. Interestingly, the giver in this interaction is not required to return the receiver's mutual gaze in order for the handover to occur successfully.

2.3.3 Methods for measuring people's responses to gaze

People are very highly tuned to others' gaze direction. Three-month-olds already shift their attention in the direction of an adult's gaze [112]. In adults, seeing someone's eyes directed laterally—even in a photograph—evokes rapid, reflexive attention shifts in the direction of the gaze [112, 149]. A series of experiments has tested this reflexive attention shift and found that it is resistant to conscious control [77, 80, 90]. In these

experiments, participants are shown a picture of a face gazing to one side. Even when are told that they should look in the opposite direction of the gaze, their attention is still drawn to the direction of the gaze in the first 500 milliseconds, an effect called *counterpredictive cueing*.

Counterpredictive cueing provides evidence that faces are special stimuli that are processed in unique cognitive pathways, because the effect is not seen in response to non-gaze directional cues such as arrows [90] or tongues [77]. An ambiguous image can even be manipulated to elicit the counterpredictive cueing effect or not, depending on whether the image is presented as a face or a car [203]. Functional MRI studies show that a single image activates different brain pathways depending on whether it is presented as eyes or as a non-social directional image [138], further strengthening the idea that eyes are processed differently than other cues.

Chapter 3 details a counterpredictive cueing experiment with robot faces as stimuli, which shows that robots are cognitively processed like non-social directional cues and not like human faces.

The counterpredictive cueing effect might be explained in part by people's strong tendency to have a theory of mind for another person, that is, a belief that the person has knowledge, goals, and intentions of their own. Functional MRI studies reveal a significant overlap in the brain areas that process theory of mind and those that process directional eye gaze [57]. In fact, observing someone signaling the presence of an object with referential gaze elicits the same neural response as observing someone physically reaching to grasp that object [193], indicating that people use gaze as a powerful indicator of others' intentions.

2.4 Human-Focused Research

The studies described in this section focus on learning about the characteristics and limits of human perception through human interactions with robots. These studies generally take place in well-controlled laboratory environments, where the limitations and features of human perception can be closely examined. Understanding how people perceive and respond to robot gaze—including what is effective and what is not—is the first step in developing gaze behaviors for robots.

2.4.1 Human response to robot social gaze

Before people can make use of a robot's social eye gaze, they must first perceive it. In multi-party conversations, people notice a robot's gaze when it looks at or near them, but not when it gazes at someone else nearby [120]. This suggests that the perception of robot gaze is egocentric—gaze is most frequently perceived when the robot is gazing directly at the viewer, and is less frequently perceived when the robot is gazing at someone else. People have stronger feelings of "being looked at" when a robot gazes at them using short, frequent glances rather than longer, less frequent stares [5].

People are also sensitive to robot eye gaze when that gaze is directed at objects or locations in the environment. For example, in object selection games, people can use referential gaze cues from a virtual agent [30] or a robot [175] to make predictions about which objects to select, even when they are not consciously aware of those cues. For a back-projected robot head, people can predict the target location of the robot's gaze almost as accurately that of a human's gaze, though accuracy suffers when the head is viewed from the side or when gaze involves just head orientation and not eye movement [13].

Such object-directed referential gaze has specific gaze timings that appear nat-

ural to people. Using an immersive virtual environment, researchers were able to empirically measure the timing of referential gaze during an interaction between a person and a virtual agent [192]. They found that the mean time a referential gaze dwelled on a referenced object was about 1.9 seconds, and that participants expected a responding gaze to be directed from their partner to the target object within about 2.5 seconds of their reference. These timing values can inform the production of gaze in future agent systems.

While people can successfully interpret robot eye gaze for object references, having a robot display mutual gaze also improves people's subjective evaluations of that robot. Mutual gaze from a stuffed animal companion robot leads to favorable evaluations of the robot [266]. When a robot is learning from human demonstration, displaying mutual gaze leads people to view the robot as more intentional than displaying random gaze; people spend more time teaching the robot, pay more attention to it, and speak more with it [122].

People's preconceived expectations for an agent's gaze influence how they respond to that gaze. In what they describe as a "non-verbal Turing test," researchers manipulated the amount of gaze following displayed by a virtual agent, and asked participants to evaluate whether the agent was being controlled by a human partner or by a computer program [191]. They found that ascriptions of humanness varied by whether the human partner was introduced as naïve to the task, as cooperative, or as competitive, suggesting that interpretations of the "humanness" of gaze behavior depend on the intent ascribed to the agent.

For robots that act as therapy assistants to children with autism spectrum disorder (ASD), gaze can be a particularly important cue because of the deficit in social gaze that is often part of this disorder [216]. Some children with ASD show spontaneous social gaze behaviors in response to robots, including increased eye gaze and shared attention during robot interactions as compared to human interactions [239]. These

findings lend support to the use of robots as therapy tools. However, there is large variability in responses, and other children do not demonstrate the same increase in gaze behavior [239]. Because gaze trajectories sometimes differ between people with ASD and those without, computationally recognizing and modeling people's gaze might be a way to diagnose and evaluate ASD [63, 215, 220].

2.4.2 Differences in human response to robot and human gaze

It is tempting to assume that perfectly matching robot gaze behaviors to human gaze behaviors will elicit identical responses from people, but this is not always the case. There are several studies that suggest that gaze from robots is interpreted differently than gaze from humans.

In general, it is difficult to compare robot gaze to human gaze directly, because while robot gaze can be infinitely controlled, human gaze tends to have small, unpredictable variations. However, one well-controlled study made this comparison using a trained actor who performed identical behaviors to a pre-programmed robot (Figure 2.3). While viewers' gaze patterns were overall similar between the human and robot conditions, fine-grained analysis reveals differences in people's responses to robot gaze and human gaze. For instance, people spend significantly more time looking to a robot partner's face than to a human partner's face when naming an object, indicating an apparent concern for ensuring that the robot is attending to the object in question [269].

Other fine-grained analysis reveals that robot gaze is not afforded the same special cognitive status as human gaze. Recall from Section 2.3.3 that people show a tendency to unavoidably shift their attention in the direction of another person's averted eye gaze, referred to as the reflexive cueing effect. This effect suggests that gaze is processed in a different neural pathway than other directional symbols like arrows. A

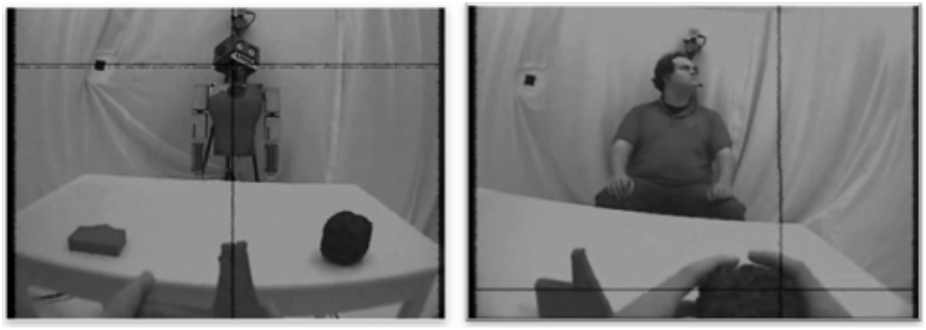


Figure 2.3: Comparing robot gaze to human gaze using eye tracking can reveal differences between people's responses to the two stimuli. People spend more time looking at a robot partner's face than a human partner's face when naming an object [269].

test of reflexive cueing using both highly anthropomorphic and highly stylized robots showed that robots failed to elicit reflexive cueing in people, suggesting that robots are cognitively processed more like arrows than like faces [2, 7].

Even infants will disregard robot gaze while treating human gaze as meaningful. When infants are shown videos of robots and humans looking at objects, they can follow the robot gaze as well as the human gaze. However, the infants look longer at, and show a preference for, objects gazed at by the human but not objects gazed at by the robot [184]. Only after infants observe a robot engage in a socially communicative exchange with an adult do they follow a robot's directional gaze [165]. This suggests that, even to infants, robot gaze is not automatically as meaningful as human gaze.

The examples in this section provide some evidence for differences in human responses to robot versus human gaze, but more investigation is warranted (see Section 10.2.2). For example, while Yu *et al.*'s study tried to carefully control the human actor's behavior to make it identical to the robot's behavior, there may still have been minute differences in the human performance. Additionally, differences in appearance between the robot and human actor might have played a role in eliciting different gaze responses from people. Admoni *et al.*'s study (detailed in Chapter 3) employed two particular robots with specific appearances, and the characteristics of these appearances may have affected why robots did not elicit reflexive cueing. In

the two studies with infants, the robots exhibit a lack of animacy cues, which might have influenced whether infants saw them as social agents. It may be that some cue tangential to social communication, and not the task of communication itself, led infants to respond differently to robots in Meltzoff *et al.*'s study.

2.5 Design-Focused Research

Researchers who take the design-focused approach develop the appearances and behaviors of robots to address certain goals such as demonstrating engagement or participating in joint attention. In this section, we describe how manipulations of robot gaze behavior can affect human-robot interaction both positively and negatively.

Gaze can serve many purposes, and the goal of eye gaze communication is often dictated by the task at hand. For example, a robot engaged in conversation might display user-focused mutual gaze, while a tour guide robot performing a presentation might want to direct gaze to an exhibit using referential gaze. For this reason, we group the articles in this section by the task, or context, of the interaction: conversation, narration, collaboration, manipulation, and expression.

2.5.1 Conversation

Conversation involves an alternation of speaking and listening. For example, robots for tutoring or entertainment must be able to maintain an engaging, natural conversational exchange with human partners. The main nonverbal challenges of conversation are managing attention and turn-taking between partners, selecting the correct gaze for the conversational content, and adopting the correct conversational roles.

Before beginning an interaction, a robot needs to gain the attention of its listeners. If the robot fails to successfully engage its intended partner, the listener can be unaware or uncertain about the robot's intent to communicate, even though they

may be interested in that communication. Robots can use mutual gaze to improve the success of initiating conversation [211]. Even a very simple illusion of gaze improves user attention. Having a virtual face on the flat-screen monitor of an embodied but non-anthropomorphic robot increases the number of users who stop when greeted by the robot [53]. Having the robot's head "look at" a person by turning toward the person's location has a similar effect on engagement, even without a virtual face, though the combination of a virtual face and person tracking lead to the greatest user engagement [53].

For robot tutors, acquiring, monitoring, and maintaining user engagement is particularly important because reduced engagement means reduced learning. Animated pedagogical agents can use gaze to regulate dialogue and direct student attention to relevant information [128]. When diminishing attention is detected, robots [236] and virtual tutors [75] can use verbal and nonverbal cues, including gaze, to restore the listener's attention. Reorienting student attention in response to diminished engagement increases information recall [236], specifically on questions that require deep reasoning [75].

Conversational fluidity is managed as much by the absence of mutual gaze as by its presence. Gaze aversions can be used to demonstrate cognitive effort, modulate intimacy, and mediate turn taking [20]. Using empirical timings for gaze aversions collected from lab-based observations of human-human conversation, researchers designed gaze aversion behaviors for virtual conversational partners. Virtual agents using gaze aversions for these conversational functions are more successful at regulating the conversational flow and elicit greater disclosure from people than agents that do not perform gaze aversions or perform gaze aversions at inappropriate times [20].

This gaze aversion model, when applied to embodied robots, yields a similar effect even though the robot (a Nao) uses head turns to signal gaze direction instead of articulated eyes [24]. In some cases, though, averting gaze may not be the most

effective way of mediating turn-taking. In an interaction that involved handing off speaking turns between a person and a Nao, flashing the eye LEDs to yield the speaking turn led to the fastest responses from people, while using gaze aversions actually led to slower responses than using no turn-taking cue at all [251].

People are also sensitive to the dynamic interplay between their own gaze and a robot's gaze. Robot gaze that is responsive to the user—that is, joint attention and mutual gaze that occur in response to human behavior—increases the self-reported "feeling of being looked at" over gaze that is independent of a user's behavior [267, 268].

The content of conversation influences what kind of gaze works best. In conversations about emotionally neutral topics, robots that make eye contact are seen as more sociable and intelligent than robots that avoid it, but this effect is reversed when the topic of conversation is embarrassing, with eye contact avoiding robots rated more highly [65]. In persuasive conversation, natural gaze behaviors improve a robot's persuasiveness [103], even more than using expressive vocalizations [64]. Gaze also seems to mitigate the effects of other nonverbal behaviors on persuasiveness: when performed with eye gaze, persuasive gestures improve a robot's overall persuasiveness, but when performed without eye gaze, persuasive gestures actually have the opposite effect, hindering a robot's persuasiveness [103].

In multi-party conversations, robot eye gaze can influence people to take on certain conversational roles. Several studies have found that a robot can use gaze behaviors to manipulate certain members of a group into taking conversational roles such as onlooker, active participant, or listener [139, 173, 174]. A robot's gaze behaviors are successful at influencing people to conform to the intended roles as much as 97% of the time [173]. A virtual agent's gaze can also influence which participant in a multi-party conversation takes the conversational floor next, with up to 86% effectiveness in releasing the floor to the intended speaker [42].

Seeing an agent perform sensible eye gaze during conversation improves people's perceptions of that agent. When a robot is a listener in a multi-party conversation, seeing the robot track the conversation with its gaze elicits higher evaluations of that robot's comprehension and naturalness than seeing the robot perform random gaze turns between speakers [143]. A robot that displays gaze focused on its human conversational partner, but occasionally responds to motion in the background, is evaluated as more natural, human-like, and attentive than a robot that exclusively focuses on the partner or that distributes its gaze randomly [225]. Virtual avatars that use turn taking gaze during conversations are evaluated as more natural and more pleasant, and their conversation is rated as more engaging, than avatars that use random gaze or no gaze in their communication [95]. In an immersive virtual reality setting, researchers confirmed that people have more positive subjective evaluations of an agent when it performs conversationally-driven gaze than when it performs random gaze, but that the effect depends on the agent's appearance. More realistic avatars benefit from appropriate conversational gaze, but low-realism avatars, such as stick figures, are adversely affected by human-like gaze behavior [96].

2.5.2 Narration

Unlike conversation, narration primarily involves a single speaker. There may be a single listener or an audience with multiple listeners. Contexts that involve narration include lecturing (as with robot tutors providing information about a topic), storytelling (as with entertainment robots), and presenting (as with robot tour guides that describe museum exhibits). Challenges in narration involve ensuring information recall and directing attention to external information sources such as exhibits in a museum.

The type of robot gaze performed during narration can influence how much information is remembered by listeners. Longer participant-directed gaze from a sto-

rytelling robot leads to better recall of story content [172]. In contrast, virtual agent tutors that display more gaze toward the subject matter than toward their listener generate better retention of information [23]. In general, however, socially communicative gaze is better for ensuring information recall than no gaze or gaze that is incongruous with communicative goals [115]. Gaze behaviors can also be combined with other socially supportive behaviors, such as natural gestures and empathetic facial expressions, to improve student performance in language learning from a robotic tutor [207].

Listener-directed robot gaze during tutoring and storytelling is correlated with positive perceptions of a robot. Subjective ratings of likability and other positive attributes are higher for robots that display more affiliative gaze (that is, gaze directed at the listener) than referential gaze [23]. Robots exhibiting gaze that correlates to the content of their communication are seen as more natural and competent [115], and longer gazes toward a listener yield greater feelings of likability [133]. Mutual gaze, when presented with other social behaviors like head nods and posture mimicking, greatly improves people's perceptions of rapport with a virtual agent [261]. Joint attention from a robot toward the topic of discussion is seen as more human-like than only mutual gaze [133].

However, there are cases in which listener-directed gaze negatively impacts people's perceptions of a robot or virtual agent. High levels of mutual gaze without other social behaviors can decrease rapport with a human user to the same levels as a virtual agent specifically designed to show boredom [261]. Additionally, the benefit of listener-directed gaze may be influenced by gender; when listening to a storytelling robot, men evaluate the robot more positively when it looks at them more frequently than at their partner, while women show the opposite effect [172].

Some presentations, such as guided tours, involve narration about material that is situated externally to the agent. Tour guide robots might present a new technology

to a user [223], provide directions to shoppers in a mall [211], or give location tours of indoor spaces [141]. A primary challenge for this kind of narration is to direct attention toward objects of interest in the environment, which can be accomplished using deictic gaze.

A tour guide's deictic gaze has a positive effect on listener engagement and attention. When a robot displays deictic gaze that reflects the subject of its speech, people display more nodding and mutual gaze, signaling increased engagement, than they do when the robot's deictic gaze occurs at random points in its speech [145, 264]. When a robot uses deictic gaze in addition to spoken object references, people are more engaged, spending more time interacting with the robot and displaying more coordinated gaze behaviors than when the robot simply speaks without supportive gaze [223]. When listening to a robot tour guide, listener gaze directed away from the robot is often congruent with the robot's topic of discussion [141], indicating that robots can successfully guide listener attention to desired locations.

Tour guides can affect a listener's experience by whether they look at the listener or at the display. A robot that orients its body (including its eyes) toward an exhibit can more easily engage its listeners than a robot that orients its eyes toward the audience, but people lose interest in the robot and its narrative more often when the robot looks at the exhibit and not at its audience [132]. Robots can influence people's experience of a tour by how often they direct gaze to each listener. When a robot "favors" a person by gazing at them longer than others in the group, that person reports greater feelings of likability toward the robot [133].

2.5.3 Collaboration

Collaboration requires communication of goals, knowledge, and intentions. For example, a robot that helps a user construct furniture needs to express its current goals and intended action to fluidly collaborate with a human partner. Gaze can be used

to reveal these mental states to a partner in unobtrusive ways. Collaboration often involves the physical environment, so in addition to gaze that reveals mental states, such interactions also require gaze that references objects and physical locations.

Revealing mental states through nonverbal communication (including eye gaze) makes cooperative task performance faster, with errors detected more quickly and handled more effectively than purely task-based nonverbal communication [47]. Indicating engagement and providing feedback through subtle gaze behaviors improves performance of a human-robot team [129]. Users also report understanding the robot better during their collaboration when it makes its mental models explicit [47]. Expressive eye gaze is one behavior (among many drawn from animation principles) that can make intentions and desires more explicit, for instance, by looking at a door handle when wanting to open a door [237]. Even when users are unaware of the intended communication, robots can "leak" their intentions through eye gaze, influencing human behavior in measurable ways [175].

A key element of collaboration is referencing objects in the environment. Joint attention from a companion robot effectively draws a user's attention to where the robot is looking [266]. Eye gaze can also act as a reinforcement of pointing gestures [212]. A robot can use eye gaze to support its speech in a cooperative object selection task, in which a human user needs to select an object referenced by the robot as quickly as possible [3, 44]. People can recognize and respond to predictive eye gaze that indicates spatial references, completing the task faster than if they had been relying on the robot's speech alone.

Errors in robot gaze hinder speech understanding, because people expect the robot's gaze to indicate what the robot intends to verbally reference [3, 113, 231, 232]. For tasks that involve a light cognitive load (for instance, selecting the object referred to by the robot as quickly as possible), people recover quickly from errors in robot eye gaze and show no difference between incongruent gaze and having no gaze cues

at all [3, 113]. Chapter 5 details the results of a study that supports this. However, in more cognitively demanding tasks (such as deciding whether a statement about the referenced object's visual features is true or false by comparing features of the referenced object to other visible objects), incongruencies in a robot's eye gaze and speech lead to diminished performance over having no gaze at all [231, 232].

To improve collaboration, users can teach skills to robots by performing demonstrations of those skills [26]. When learning from demonstrations in this way, robots can use gaze to establish joint attention and solicit feedback when uncertainty is high [159]. When a robot student responds to joint attention by following the human teacher's gaze, it better conveys the robot's internal states and knowledge, which leads to more efficient teaching: fewer errors, faster recovery from errors, and less repetition of learned information [117]. People also rate the robot as more natural and competent at its task when it engages in joint attention [117]. People are sensitive to the robot's mental state when they are teaching it, and will adjust their behavior (in terms of pauses, speed, and magnitude of motions) to account for the robot's visual attention [194]. When there are multiple robots to be taught, people are sensitive to each robot's gaze behavior; they look longer and are more engaged in teaching robots that actively seek mutual gaze than robots that passively follow the human's attention when it shifts to the other robots [263].

2.5.4 Manipulation

One of the primary benefits of robots as physically embodied systems is their capability to physically manipulate objects in their environments. Many robot manipulators are still isolated in factories or other carefully controlled settings, but robots are increasingly required to operate in environments inhabited by people [134]. For example, robot caregivers or office assistants can benefit from the ability to pick up, carry, and hand over objects to assist their users.

Social robot gaze is particularly important for object handovers because this type of manipulation depends on coordination with a partner. While people are generally capable of performing successful handovers without much thought, the process of handing an object to another individual employs a series of subtle but important nonverbal cues, including eye gaze (Figure 2.4). A decision tree built on empirical data of human-human handovers reveals that joint attention (attending to the same location or object) is important for coordinating a handover between two people, but that mutual gaze (where both people make eye contact with each other) is not [234, 235]. Robot-to-human handovers are improved when a robot monitors its partner's eye gaze for attention and engagement, only releasing the object when user's focus of attention has turned to that object [101]. In multi-party scenarios, robots can also use eye gaze to nonverbally select a member of the crowd to whom to hand an object [139].

Gaze improves the efficiency of handovers. During a handover, people begin reaching for an object earlier—signaling their confidence in the handover—when a robot continuously looks at the projected position in space where the handover will occur, than when it looks away from that location [168]. People reach for the object even earlier when the robot continually gazes at their faces than when it looks at the handover location [271]. Gazes that transition between the user's face and handover location do not improve how quickly reaching begins, though people report that these gazes communicate the handover timing more effectively than continuous gazes [271].

Occasionally, a robot will need to direct a person's object manipulation, for example, when requesting that a person move an object within the robot's reach. Social behaviors including gaze cues can help inform people about where and how the robot would like such assistance; for example, the robot can look at the location to which it wants the object moved [187]. This kind of social and referential gaze may be ignored during handovers unless the receiver has good reason to interpret the robot's

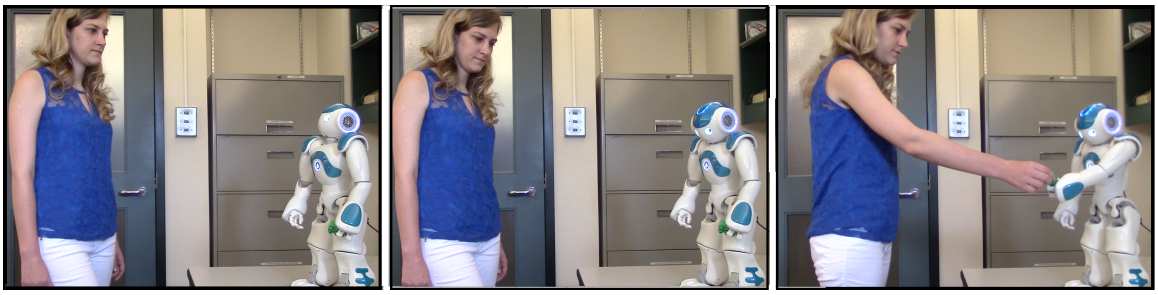


Figure 2.4: Gaze is an important part of successful handovers. In this sequence, a human and a robot first establish mutual gaze, then joint attention to the handover location, and finally complete the handover.

eye gaze as intentional [4]. By introducing an unexpected action—a delay in releasing the object—into the handover, robots can increase the amount of time spent looking at their gaze and user's compliance with gaze-based spatial references [4]. Chapter 6 details this finding.

2.5.5 Expression

Robots may benefit from the ability to express personality or emotion. For example, robots engaged in long-term interactions should have engaging personalities that keep the interactions from becoming stale; entertainment and companion robots may wish to express emotions that engage their users. Eye gaze is one way to express personality and emotion, though the challenge lies in generating the right kind of gaze to influence this subjective judgment.

Gaze behavior—in terms of where and for how long the robot gazes—can be used to express recognizable personalities and emotions. High levels of mutual gaze expresses feelings of trust [182] and extroversion [21]. Conversely, gaze aversions express feelings of distrust [182] and introversion [21]. In animated agents, eye movement can be used to express recognizable emotions such as joy, sadness, anger, fear, disgust, and surprise [156]. Systematically manipulating features of gaze (such as amount of gaze, duration of gaze, and the points of fixation during gaze aversion) yields consistent

impressions of dominance and friendliness in a robot [93].

Robot expressiveness is important because it influences how people respond to the robot. Matching the robot's behavior with a user's personality (as evaluated by a personality survey) leads to greater motivation to engage in a repetitive task and improves subjective perceptions of a robot during collaboration [21]. When people perceive a robot favorably, they show no difference in proxemic behaviors when the robot increases its amount of mutual gaze. However, for people who dislike the robot, an increase in the robot's mutual gaze causes them to physically distance themselves from the robot [170]. Interestingly, this effect of interpersonal dynamics does not extend to psychological distancing during conversation, as measured by people's willingness to answer a series of revealing questions [170].

2.6 Technology-Focused Research

There are many approaches to achieving communicative social gaze from robots and virtual agents. One approach models the underlying neurological or psychological processes, based on the idea that mimicking biology is an effective way to generate gaze that appears natural. Another approach is data-driven, basing gaze behaviors off of empirical measurements of gaze features—such as the timings, frequencies, and locations of gaze aversions—which are recorded during observations of human interactions. A third approach is to construct heuristic systems that are not grounded in biological or empirical observations, but (as with rules drawn from animation principles) still appear to generate expressive gaze. In this section, we review these approaches to implementing gaze in virtual and physical systems.

2.6.1 Biologically inspired systems

Biologically inspired gaze models attempt to replicate the underlying cognitive or neurobiological mechanisms that control gaze behavior in people. The systems in this category adhere to what we understand of the brain's function, though they operate at varying levels of detail. Some systems replicate the neuron-level receptive fields in the visual cortex to perform visual attention [126]. Others employ a developmental approach to learning joint attention and other gaze behaviors; developmental robotics mimics human cognitive growth by attempting to replicate the process of human ontogenetic development [161]. There is some evidence that biologically inspired models have higher accuracies than other gaze models in predicting human-like gaze fixations [43].

Many biologically inspired gaze models focus on directing attention to areas of interest in a visual scene by replicating the neurological response to those visual stimuli. These models generally have a similar structure: they compute the saliencies of several features in parallel, then combine these saliencies into a single saliency map [91]. Both low-level scene features (such as color, intensity, and orientation) and high-level contextual features (such as object or scene recognition) are used to create saliency maps [126]. Neurobiological models of gaze behavior that use bottom-up saliency maps can successfully track salient targets and perform visual search in demanding scenes [125], leading to realistic visual attention behavior in virtual avatars [123, 124]. Adding high-level contextual or motivational information to the low-level saliency cues enables robots to naturally direct their visual attention based on the current task and environment [48]. Behavior can be influenced by both visual and auditory saliency maps, as in an implementation of attention for the iCub robot [205]. Other models of overt visual attention produce head and eye turns using dynamical neural networks that respond to visual saliency [256].

Cognitive models attempt to replicate high-level human cognition, so they operate

at a level of abstraction above neuronal responses. Visual attention can be incorporated into cognitive models, such as ACT-R, to address how people's cognitive systems respond to environmental input [18]. For example, a computational cognitive architecture called ACT-R/E (ACT-R with Embodiment) performs conversational tracking by switching its visual attention to the speaker in a multi-party conversation [248]. By tightly integrating gaze behaviors with the underlying cognitive model controlling reasoning, dialogue management, and goals, the Rickel Gaze Model can generate real-time gaze shifts that reveal a virtual agent's internal processes [150].

Developmental models are also inspired by biology, though they attempt to replicate the higher-level cognitive process of learning rather than the underlying neuronal structure. For example, a computer vision model on a developmental robot uses saliency maps of the environment along with a probabilistic algorithm that estimates a teacher's gaze vectors to perform shared attention and gaze imitation; shared attention and imitation are foundational skills that bootstrap cognitive learning [110]. A robot can develop the ability to perform joint attention through demonstrations of attention to salient objects, much in the same way that infants acquire this capability by interacting with their adult caregivers [76, 176, 250]. These basic joint attention behaviors serve as the basis for learning more complex social communication skills in a humanoid robot [214].

Biologically inspired systems closely match actual neurological or biological processes, and their strength lies in this adherence to real cognitive processes. However, modeling underlying biological functions is not always practical due to computational constraints, or even constraints on what is known about human cognitive processes. Other methods described in the next sections have more abstract approaches to modeling gaze behavior.

2.6.2 Data-driven systems

The data-driven approach to generating robot eye gaze takes advantage of people's natural expressiveness by using quantified observations of human behavior to develop and train gaze systems. Though these systems use empirical behavioral data, they generally do not consider the underlying biological or cognitive mechanisms. The process of building data-driven systems generally follows three steps: first, observations of people using eye gaze in a desired scenario (such as in conversation) are collected. Second, a model of gaze behavior is developed from the gaze data in these observations, which are acquired either by manual coding or through automated feature extraction. Third, the behavior model is evaluated in a human-robot or human-agent interaction.

Data-driven researchers have recorded and analyzed human gaze behavior in a wide variety of scenarios. Conversational gaze has been recorded for pairs of previously unacquainted people speaking about movie preferences [20], free dialogue between two people with various existing relationships (including hierarchical work relationships and romantic relationships) [121], and in four-person conversations about controversial topics such as "should euthanasia be legitimized" [186]. Observations of gaze in tutoring scenarios have been collected for student-teacher pairs covering topics as varied as paper making [116], board games [8], and preparing canapés [200]. Gaze data during object manipulation have been collected for individuals constructing Lego objects [209].

Once the observational data are collected, they are annotated and processed to build a model of eye gaze within the specified interaction. Some models are built to generate robot behavior by extracting the features of gaze behaviors that achieve certain communicative functions. For example, researchers have extracted statistical information on timings and directions of gazes in dyadic conversations that achieve certain conversational functions, such as mediating turn-taking and regulating inti-

macy [20, 188]. Others have identified the direction and timing of gaze during head tilts and nods in conversation [157], or during a physical construction task in which assistance may be required [209]. Models of gaze for narration can also incorporate other communicative behaviors like gesture and speech [116].

After these models are developed, researchers must test their performance. For models that identify gaze behaviors to achieve certain communicative effects, researchers can incorporate these models into robot behavior generation systems and evaluate them in human-robot interactions. Data-driven gaze aversions in conversation lead to more disclosure from humans, better turn-taking regulation, and more positive subjective perceptions for virtual agents [20] and robots [24]. Gaze during head tilts and nods that is generated according to a data-driven model increases the naturalness of a conversational robot [157]. Robots that use a data-driven model to generate gazes and gestures during narration perform as well as robots that use pre-scripted behaviors [116].

In contrast with models that generate gaze behavior, other data-driven models are built to extract information about the interaction based on gaze information. For example, a probabilistic model for multi-party conversation can identify conversational regimes through gaze patterns among participants [186]. A computational model trained on physical task tutoring data can recognize the occurrence connection events that facilitate engagement between student and teacher [111, 200]. In a different tutoring interaction, a model of eye gaze and gesture trained with the k-nearest neighbor algorithm can predict the context of communication based on observations of nonverbal behaviors including gaze [8]. These models don't suggest gaze behaviors in particular scenarios, so they are not evaluated through human-robot interactions. Instead, these models are evaluated by comparing their accuracy to ground-truth data that is annotated by humans. Chapter 7 details the development of one such data-driven model.

The strength of a data-driven approach is that the empirical demonstrations used to train the models provide a principled approach to model development and ensure that the models correspond to actual human behavior. However, this need for empirical data has two main weaknesses. First, it is time consuming to collect and annotate human behavior examples. Second, the models that are developed tend to be tied to the domain from which they were collected, and are less flexible to be applied to alternate domains.

2.6.3 Heuristic systems

A third approach to developing gaze technology employs heuristics that lead to appropriate looking behavior, regardless of actual biological function or human behavior. These heuristics allow researchers to directly design gaze behaviors, using understanding of psychology or knowledge of multimodal behavior, without being tied to underlying biological realities or requiring a large corpus of observational data.

One heuristic for generating gaze behaviors is to link a robot's gaze to its speech. By representing each "communicative act" as comprised of a meaning (the information to transmit) and a signal (the nonverbal expression of that meaning), gaze can be closely integrated into the content of a robot's speech [195]. A tool that automatically extracts syntactic and semantic information from a typed sentence can use that information to generate appropriate gaze behavior for a conversational virtual agent [61].

Gaze generation based on speech may not even require semantic understanding of that speech. Some social contexts can be extracted exclusively from the timing and structure of speech; using this information, a robot can automatically generate natural gaze behaviors that support the intended context without needing to understand what is being said [171, 229]. Even loose coordination between a robot's head motions and the sentence structure of its intended speech leads to reasonable socially acceptable

gazes in tele-operated or Wizard of Oz settings [228].

Another heuristic for generating gaze behavior is to respond directly to a user's gaze. For example, one human-aware manipulation planner for robot-to-human handovers takes into account where people are looking to inform where the handover should take place, and then communicates the robot's intention to perform the handover by having the robot look at the object to be given [224]. A robot behavior system for collaboration is responsive to fine-grained, real-time human eye movements collected with a head-mounted eye tracker [263]. A virtual agent for conversation monitors a user's gaze to assess their level of interest, and responds with pre-specified gaze behaviors to elicit and maintain the user's engagement [190]. Gandalf, an embodied conversational agent that teaches people about the solar system, detects users' gaze acts and generates its own gaze in response to support its lesson [242, 244]. Systems that detect and respond to eye gaze can be used to shape behaviors, for instance to promote social skills like joint attention to children with ASD [37, 70].

Such responsive systems can also account for multimodal inputs that include auditory or gesture information in addition to gaze. Combining auditory cues like sound source localization with visual cues like face detection, one robot performs mutual gaze and joint attention with viewers while presenting a museum exhibit [38]. Another robot can take human gaze direction, deictic gestures, and mood into account to attend to and interact with multiple people simultaneously [226].

A major source of heuristics for gaze behaviors is the psychology literature. Models built with heuristics drawn from psychology do not attempt to precisely replicate known cognitive functions. Moreover, unlike data-driven models, which observe human behavior in the precise task to be performed by a robot, heuristics drawn from psychology are not specific to a single scenario.

For example, using approximate timings of face-directed and averted gaze from the psychology literature, as well as from informal observations, gaze behavior systems

can support real-time conversation with virtual agents [68], as well as expressions of emotion and responses to environmental distractions [102]. The Automated Visual Attending system uses rules drawn from psychology to generate attention behaviors in a virtual agent, in which goal-oriented intentional behaviors compete with involuntary attentional responses to stimuli [137]. A parametric computational model for animating gaze shifts of virtual agents that uses features informed by neurophysiology is successful at performing gaze shifts to peripheral targets [22]. Using a psychologically-based emotional model called the Geneva Emotion Wheel, a virtual agent expresses primary and secondary emotions by drawing pre-defined movement parameters for each emotional expression [155, 156]. In multi-party interactions, visual attention on a very realistic humanoid robot is driven by a context-dependent social gaze generation system that accounts for multimodal features such as proxemics, field of view, and verbal and nonverbal cues from the environment [270]. Chapter 8 details a heuristic model for generating deictic gaze and gesture in a human-robot collaboration.

Heuristic systems provide a framework for model development that applies knowledge from psychology but avoids the computational cost of biological modeling and the procedural cost of collecting examples for data-driven modeling. However, heuristics must be carefully chosen to accurately represent the cognitive factors at play in a given situation, or the model may not accurately capture the desired phenomena.

2.7 Summary

This chapter reviewed the state of the art in eye gaze for social robots. In doing so, it delineated three major approaches to research in the field of HRI—human-focused, design-focused, and technology-focused. It organized the large number of disparate research articles on gaze in HRI into these coherent categories. This chapter also established a grounding of terms and concepts that are used widely in HRI research,

and which will be used throughout this thesis. The remainder of this thesis describes novel contributions to the state of the art in eye gaze for social robots.

3

Robot Gaze and Reflexive Cueing*

Much of the research in social robotics assumes that psychology findings can be imported wholesale into human-robot interactions. For example, social eye gaze models for robots are often constructed based on observations of human behavior, under the assumption that human-robot interaction should mimic human-human interaction. In this chapter, we show that human eye gaze and robot eye gaze are actually processed in cognitively different ways. We extend a classic psychophysics experiment, which has shown that people reflexively attend in the direction of social eye gaze but not in the direction of socially irrelevant directional cues such as arrows. We ask whether robot gaze elicits the same reflexive cueing effect as human gaze. We consider two robots with varying levels of anthropomorphism to determine if differences in cueing effects are based on robot appearance. A millisecond-level analysis of human behavior indicates that both human and robot faces convey directional information, but that robots fail to elicit attentional cueing effects evoked by non-robot stimuli, and that no difference exists based on robot appearance.

*This work was originally published as:
Henny Admoni, Caroline Bank, Joshua Tan, and Mariya Toneva. Robot gaze does not reflexively cue human attention. In L. Carlson, C. Hölscher, and T. Shipley, editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci)*, pages 1983–1988, Austin, TX USA, 2011. Cognitive Science Society.

3.1 Introduction

Joint visual attention is an important aspect of typical social interactions. A single gaze communicates information—there are predators hiding behind that tree; a tasty source of food is over there; you are crossing into my territory—and supports social conventions such as conversational turn-taking or joint referencing. As robots become more integrated into daily human life, social interactions occur with increasing frequency between humans and robots, as well: robots assist nurses in hospitals, act as companions for the elderly and the disabled, and interact with children in therapy. In this chapter, we investigate whether people are responsive to joint attention cues from robots. Specifically, we focus on attentional shifts that occur in response to another person's eye gaze cues.

Evidence from psychophysics suggests that typical humans readily shift their attention in response to a directional cue, such as averted eyes or an arrow. In traditional *non-predictive* cueing experiments, subjects view a centrally-presented stimulus followed by a peripherally-presented visual probe, and press a keyboard key in response to the probe. Key press response times are theoretically correlated with attention: participants will respond more quickly to probes located in the direction to which they are already attending. Studies have found that when the stimulus contains directional information (such as a face with averted eyes, or an arrow pointing in one direction), people respond more quickly to probes at *cued* locations, in which the probe is on the same side as indicated by the stimulus, than to probes at *uncued* locations, even when they are told that the cue does not indicate probe location and should be ignored [77, 80, 88, 149]. Further studies confirm that the central cue, and not peripheral appearance of probes, causes this reflexive shift of attention [89]. At-

tention shifting via directional cue seems to be an early and reflexive skill for humans: children as young as three months old will attend more quickly to a peripheral probe on cued trials than on uncued trials when the cue is a human face [112].

When cues are *counterpredictive* of probe location, however, social stimuli such as faces and eyes elicit different patterns of behavior than other directional stimuli. In *counterpredictive* cueing paradigms, probes appear with significantly higher probability on the *opposite* side of that which is cued by centrally-located stimuli [80]. For example, when the centrally-located stimulus is directed toward the left, probes have a 75% chance of appearing to the right of center, and vice versa. In counterpredictive experiments, it is beneficial for participants to orient attention away from the cued direction; therefore, shorter response times to probes in the cued direction are attributed to reflexive or uncontrollable attention shifts. In contrast, shorter response times to probes in the uncued (but *predicted*) location are interpreted as *volitional* orienting of attention.

Counterpredictive experiments reveal that subjects reflexively orient in the direction of eyes [80] but volitionally orient away from the direction of arrows [90] or extended tongues [77]. A stimulus that is ambiguously social will elicit reflexive attention shifts when presented to participants as a social cue (a picture of eyes), but not when it is presented as a non-social cue (a picture of a car) [202]. Furthermore, the effect of this cue on reflexive attention persists if the cue is presented first as social and then as non-social, but not vice versa. This effect seems strongest for faces, but not necessarily unique to them: arrow cues have also been shown to trigger reflexive orienting, with magnitude of reflexive orienting toward arrows positively correlated with individuals' voluntary attention control [246], suggesting that dissimilarities in attention directed at eyes and arrows are differences of magnitude (strong versus weak), rather than of kind (reflexive versus volitional).

Eye-tracking and brain-imaging studies reveal similar results. People make more

erroneous eye saccades in the direction of a "distracter" cue they are told to ignore if that cue is a face, rather than an arrow [199]. Functional MRI studies show that the same cue activates different pathways depending on whether it is perceived as eyes or as a non-social directional image [138]. Attentional orienting to gaze cues and to arrow cues activates different cortical networks, with attentional orienting to arrow cues relying on a pathway associated with voluntary shifts of attention [108]. In a different fMRI study, however, the same cue activated the same extensive cortical network regardless of whether it was interpreted as an eye or an arrowhead, though the eye cue more strongly engaged some parts of this network [245].

Psychologists have suggested that shared attention is a precursor to developing a theory of mind for other people, and that lacking ability to interpret others' visual attention might indicate social disorders such as autism [33]. Children with autism fail to show preferential sensitivity to socially relevant cues such as human gaze: they demonstrate similar response times to both arrows and faces on a counterpredictive cueing task (whereas typically developing children are cued by faces but not by arrows) [218], and they avoid shifting their gaze in response to non-predictive gaze cues [203]. Participants' scores on the Autism-Spectrum Quotient have also been negatively correlated with cueing magnitude [36].

In summary, evidence suggests that for non-predictive cues, both social and non-social directional stimuli elicit reflexive attention shifts in cued directions, but that for counterpredictive cues, socially relevant stimuli (such as human eyes) continue to elicit reflexive attention shifts while non-social directional stimuli, such as arrows, exhibit weak or no reflexive attentional influence. The psychophysical methods used to isolate attention shifts for faces and arrows can be applied to novel stimuli to inform the field of human-robot interactions (HRI). HRI is interested in exploring how people perceive robots and understanding how designers can create robots that interact naturally with people. To date, there has been little research on the cognitive effects of

robots on human attention. As the presence of robots in day-to-day social situations increases, however, it becomes important to evaluate robots' cognitive influence to better understand the types of roles robots can perform and to improve the design of human-assistive robots.

Some evidence already suggests that robots can use gaze cues to "leak" information to humans. In conversations between robots and naïve human participants, robots were able to define participants' roles, such as addressee, bystander, or eavesdropper, through visual attention cues [174]. Another study found that robots can influence people's decisions in a game by shifting their eyes briefly to a target, even when participants do not report seeing those cues [175]. In the latter study, robot appearance influenced the effectiveness of gaze cues: Geminoid, a very human-like robot, was more effective at revealing intentions through gaze cues than Robovie, a robot with more abstract human features.

In this chapter, we ask: will robots be treated like humans or like arrows? That is, will robot gaze be interpreted by humans' cognitive systems as a social cue on par with human faces, with attendant reflexive shifts of attention in the gaze direction? Or will robots be perceived by humans as non-social entities, such as arrows or cars, allowing participants to override reflexive attention shifts in favor of volitional orienting toward predicted probe locations? Because robots are designed with varying levels of anthropomorphism, we use two robot stimuli, one from a very human-like robot called Zeno, and one from a less anthropomorphic robot named Keepon. Cueing effects from human faces have been found to be stronger for schematic faces than for real faces [107], suggesting that cueing information contained in schematic faces is overshadowed by the complexity of real faces. For this reason, we also use two types of human face stimuli: a photograph of a human face and a line drawing of a face. Finally, we use an arrow as a non-social but directional stimulus.

3.2 Methods

This experiment employs two commercially available robot platforms. Zeno is produced by Hanson Robotics as a realistic, expressive robot (Figure 3.1(c)). In addition to eyes and a nose, Zeno's face has human-like features such as eyebrows, lower eyelids, an expressive mouth, and hair. In contrast, Keepon's features are less human-like (Figure 3.1(d)) [144]. Keepon is a 20 cm tall robot made of two stacked yellow spheres of deformable rubber; its eyes are white circles overlapped by smaller, concentric black circles and its nose is a black circle. Keepon's deformable body and eyes with sclera suggest biological features, but its form and color (bright yellow) clearly indicate that it is robotic. The aim of selecting such different robots is to identify whether human-like features are necessary to evoke the same (purportedly social) response as to a human face.

Participants were 41 male and 29 female Yale University students between the ages of 18 and 34 (mean age 21.4). Each participant was assigned to a single stimulus condition (human, line drawing, Zeno, Keepon, or arrow) in an alternating fashion. Participants were recruited in person or with flyers around campus, and were rewarded with candy at the conclusion of the experiment.

3.2.1 Stimuli

All stimuli were displayed on a laptop screen positioned approximately 30cm away from the seated participant.

The human gaze stimulus is a head-and-neck photograph of a woman (Figure 3.1(a)). Her head subtends a visual angle of 6.2° horizontally. Each eye subtends 1.0° and the center of each eye is 1.2° to the right or left of center. This stimulus was chosen as a human (i.e, social) analogue to photographs of the robots.

The line drawing stimulus, re-created from [90], is a black-and-white line drawing

of a face with circular eyes and nose, and a line for the mouth (Figure 3.1(b)). The head subtends 7.5°; each eye subtends 1.0° and its center is 1.0° left or right of image center, where the nose is located. This stimulus has been previously shown to elicit both reflexive and volitional shifts of attention on a similar task [90].

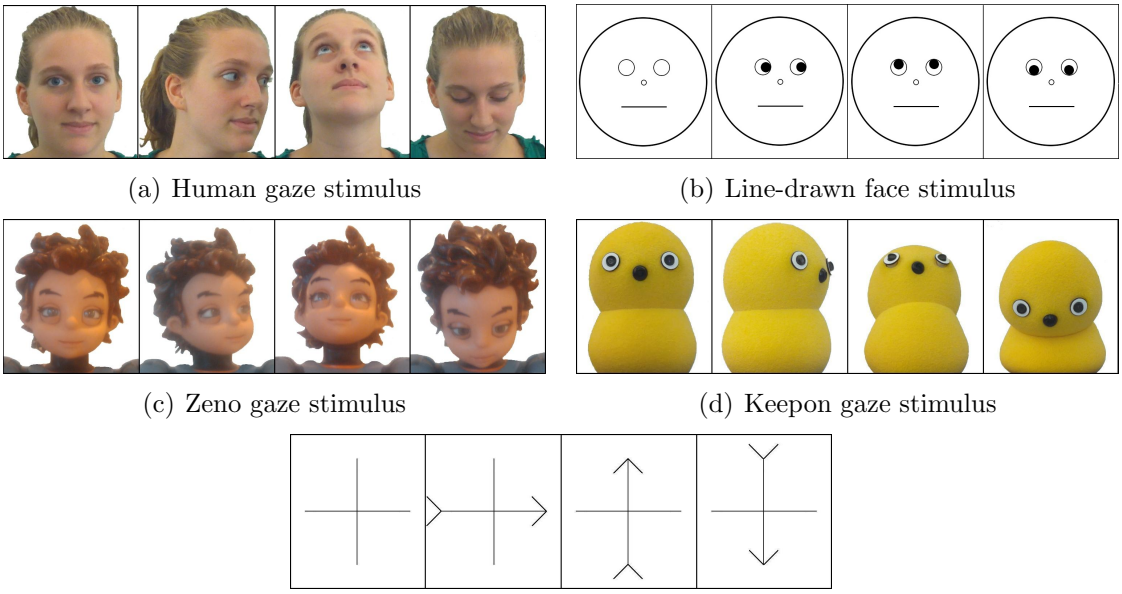
Zeno provides an example of a highly anthropomorphic robot (Figure 3.1(c)). The Zeno stimulus is a head-and-neck photograph of the robot, with face subtending 6.7° in width (7.8° including hair) and each 1.0° eye located 1.3° to the left and right of center.

Keepon represents the opposite end of the scale of anthropomorphism (Figure 3.1(d)). The Keepon stimulus is a full-body photograph of the robot, subtending 6.2° of the visual field, with each 1.0° eye located 1.75° to the left and right of center.

The arrow stimulus is 7.1° long and drawn over a 6.2° fixation cross; equal amounts of visual information are presented at the head and tail of the arrow, thereby avoiding the possibility that cueing results simply from additional features in the head direction (Figure 3.1(e)).

Each stimulus had left-, right-, up- and down-facing variants (see Figure 3.1). In a single trial of the cueing condition, the front-facing variant was presented for 500 milliseconds, followed by one of the other ("turned") variants. After a 400 millisecond stimulus onset asynchrony (SOA), or a 600 millisecond SOA in human and Zeno conditions, a probe letter, either "T" or "L," appeared on the screen in one of four positions relative to the image: above, below, to the left, or to the right. Each probe letter was 0.9° tall and wide, and was presented along the midline 4.8° from center. Cue and probe remained on screen until participants responded by pressing a keyboard key or until 2 seconds elapsed. (See Figure 3.2 for an example.)

Following Friesen *et al.*, for each trial of the cueing condition, the probe had a 75% chance of appearing on the opposite (*predicted*) side of where the cue directed, and a 25% chance of appearing in one of the other three locations (approximately



(a) Human gaze stimulus

(b) Line-drawn face stimulus

(c) Zeno gaze stimulus

(d) Keepon gaze stimulus

(e) Arrow stimulus

Figure 3.1: Each subject was randomly assigned to one of five stimulus conditions. This figure shows the front, right, up and down versions of each stimulus; left versions are mirrors of right-facing versions and are omitted here for brevity.

8% chance each)—on the same side as where the cue directed (*cued*), or orthogonal to the direction of the cue (*not-predicted-not-cued* or *NPNC*), as shown in Figure 3.3 [90].

Once participants responded to the probe or 2 seconds elapsed, all images were replaced by a prompt asking participants to press any key to proceed to the next trial.

3.2.2 Procedure

Participants were seated approximately 60 cm in front of a 29 cm by 18 cm laptop screen. The experimental procedure was explained to them: they were told which stimulus they would observe and the sequence of images they would see (as in Figure 3.2). Participants were told they would first observe a front-facing stimulus, replaced by a "turned" stimulus, then a probe letter ("T" or "L"). They were also informed that the probe was three times more likely to appear on the side opposite

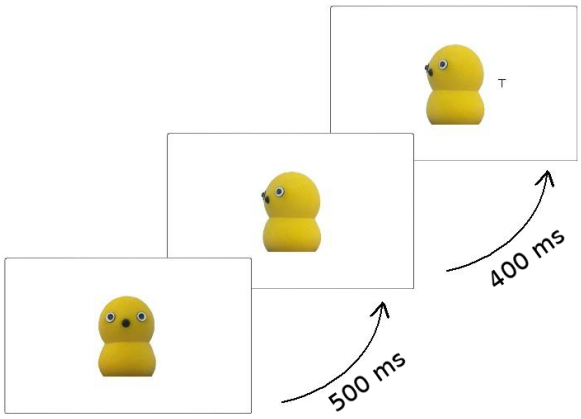


Figure 3.2: Time course for a single (predicted) trial of the Keepon gaze condition. Setup is similar for other stimuli and directions.

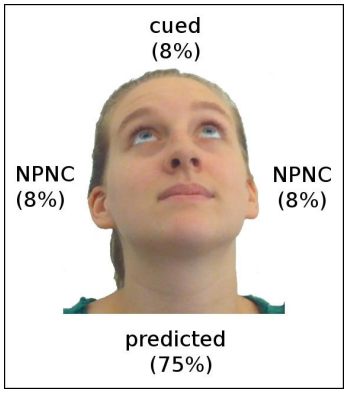


Figure 3.3: Three types of trials were presented: *cued*, in which probe and gaze are congruent; *predicted*, in which the probe is in the opposite direction to gaze; and *not-predicted-not-cued* or *NPNC*, in which the probe is on a different axis to gaze. Percentages indicate probability of occurrence.

where the gaze or symbol directed. Participants were asked to press the keyboard key of the letter appearing on the screen as quickly and accurately as possible. These instructions were also presented textually on the screen before the start of the experiment.

All participants saw 99 trials, consisting of 96 test trials and 3 additional practice trials drawn at random from the test trials and presented first. The set of test trials comprised 72 *predicted* trials (the probe appeared opposite where the cue indicated), 8 *cued* trials (the probe appeared on the side indicated by the cue), and 16 *NPNC* trials (the probe appeared on a different axis than the one directed by the cue), with

"T" and "L" presented equally frequently.

3.3 Results

Mean response times and standard deviations are listed by condition and trial type in Table 3.1. Figure 3.4 shows mean response times by stimulus condition and trial type.

Four participants were excluded for non-compliance (not following directions to respond as quickly as possible, or pressing keys at random as evidenced by high error rates); their data is not included in the analysis. Trials in which participants incorrectly identified probe letters, response times exceeded 1.5 seconds, or response times were less than 100 ms were treated as errors and excluded from analysis. The error rate was 3.9% over analyzed participants. In total, results from 70 participants across the five conditions were analyzed, as shown in Table 3.1.

A repeated measures analysis of variance with stimulus type (human, line drawing, Zeno, Keepon, and arrow) as the between-subjects variable and trial type (cued, predicted and NPNC) as the within-subjects variable revealed significant main effects for trial type ($F(2,130) = 19.819$, $p < 0.001$) though not for stimulus condition ($F(4,65) = 0.196$, $p = 0.939$). There was no interaction between stimulus type and trial type ($F(8, 130) = .673$, $p = 0.703$).

Because there was a significant main effect of trial type, we further analyzed the data within each stimulus condition with a repeated measures analysis of variance on trial type, which found significant main effects for trial type on most conditions (human: $F(2,28) = 3.675$, $p = 0.038$; line drawing: $F(2,30) = 4.328$, $p = 0.022$; Zeno: $F(2,26) = 3.409$, $p = 0.048$; Keepon: $F(2,26) = 13.558$, $p < 0.001$), and borderline significance main effects in the arrow condition ($F(2,22) = 2.672$, $p = 0.091$). In all conditions, pairwise comparisons reveal that each stimulus elicited significantly

Stimulus	Trial type	Avg. RT (ms)	SD	N
Human	cued	444	46	15
	predicted	428	54	
	NPNC	462	61	
Line	cued	458	73	16
	predicted	449	73	
	NPNC	474	70	
Zeno	cued	473	147	13
	predicted	452	108	
	NPNC	473	116	
Keepon	cued	464	65	14
	predicted	428	52	
	NPNC	469	55	
Arrow	cued	453	66	12
	predicted	433	44	
	NPNC	461	53	

Table 3.1: Average response time and standard deviation, in milliseconds. Each row represents a stimulus condition separated into trial types. The last column indicates how many participants were tested for each condition.

faster response times to predicted than to NPNC trials (human: mean difference = 33.921, sd = 8.764, $p = 0.002$; line drawing: mean difference = 24.892, sd = 5.902, $p = 0.001$; Zeno: mean difference = 24.515, sd = 8.335, $p = 0.011$; Keepon: mean difference = 39.878, sd = 9.410, $p = 0.001$; arrow: mean difference = 27.875, sd = 11.120, $p = 0.029$). Only in the robot conditions, however, were there significant or borderline-significant differences between predicted and cued trials as well (Zeno: mean difference = 23.746, sd = 12.712, $p = 0.084$; Keepon: mean difference = 36.698, sd = 8.613, $p = 0.001$).

3.4 Discussion

Results suggest that participants recognized the directional significance of all stimuli, but only responded to the cueing significance of non-robot stimuli (Figure 3.4).

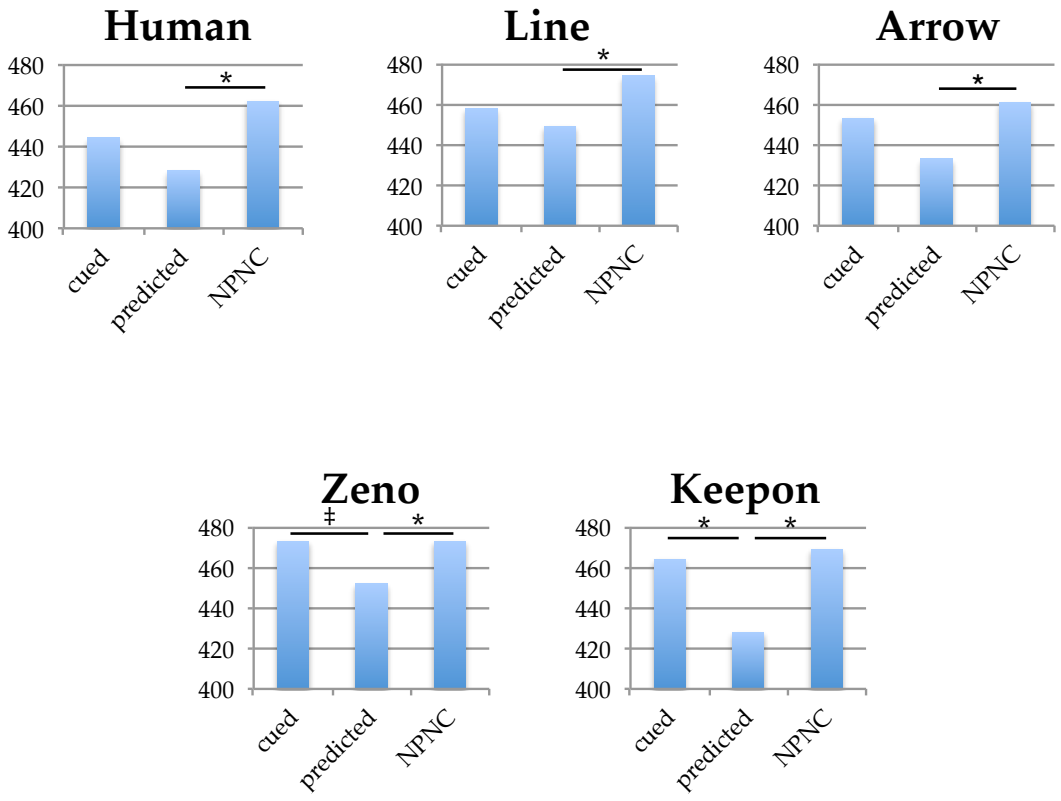


Figure 3.4: Mean response times in milliseconds for each trial type (cued, predicted and NPNC) by stimulus condition. A single asterisk indicates significant differences ($p < 0.05$), a double cross indicates borderline significant differences ($p < 0.10$).

The counterpredictive cueing task involved four possible locations for the probe to appear on each trial: a cued location, in the direction of gaze or pointing; a predicted location, opposite the cueing location, where participants were told probes would actually appear; and two not-predicted not-cued locations (NPNC), which are not cued but have the same probability of probes appearing at each of them as at the cued location. NPNC locations provide a good baseline because they involve an identical task (responding to a probe with a key press) but do not represent cued or predicted locations. In our results, participants were significantly faster at responding to probes at predicted locations than at NPNC locations for every stimulus, indicating that they recognized the direction indicated by the stimuli and used that information

to inform them of predicted probe position.

However, for the robot stimuli (Zeno and Keepon), response times were *also* statistically faster for predicted than for cued trials (borderline significance in the Zeno case, with $p = 0.084$). In other words, participants directed their attention significantly more toward predicted locations than toward cued locations, and thus show no evidence of having been cued by robot gaze. To participants in robot conditions, cued locations were attended to just as infrequently as NPNC locations that were neither cued nor predicted.

In contrast, response times were not significantly different between predicted and cued trials in the non-robot conditions (human face, line drawing of a face, and arrow). Participants in these conditions were not significantly more attentive to predicted than to cued locations, and in fact, Figure 3.4 shows that cued trial response times were, on average, greater than predicted trial response times but less than NPNC trial response times. This suggests that non-robot stimuli attracted participants' reflexive attention to cued locations despite the fact that they were no more motivated to look at cued locations than at NPNC locations.

Though they were able to extract directional information from robot gaze, participants in either robot condition were not susceptible to reflexively reorienting in the direction of robot gaze, as they were in the face or arrow conditions. In essence, participants seem to be ignoring the natural interpretation of robot gaze in favor of the counterpredictive interpretation, though they fail to do so with other directional cues. This behavior has been observed in children with autism, who are able to ignore non-predictive gaze cues, while their typically-developing peers are susceptible to reflexive cueing from non-predictive stimuli [203]. The fact that robots do not seem to cue reflexive attention, in a way that even non-social stimuli such as arrows do, suggests that robots are cognitively processed differently than common directional symbols or social entities.

Previous studies use a similar counterpredictive experimental design in which participants are asked to press a key when *any* probe appears [90, 246]. These studies use their detection task to analyze covert attention shifts, in which participants' eyes do not move (in fact, Friesen *et al.* tracked the eyes of several participants to ensure this was the case [90]). The task used in the current experiment required identifying the probe (either "T" or "L") by pressing the corresponding keyboard key, so results from this identification task are not directly comparable to results from previous detection-based experiments. It would be interesting, however, to analyze covert attention effects of various robots in detection tasks. Some robotics studies suggest that more anthropomorphic robots can convey social information—such as intention—to humans, suggesting that robot anthropomorphism affects covert attention [174, 175].

Attentional cueing is more pronounced with schematic drawings of faces than with real faces [107], so this study included both a photograph of a human face and a line drawing of a face as stimuli. Both faces elicited significantly faster responses to predicted versus NPNC trials, but not to predicted versus cued trials. Though the arrow stimulus also showed this effect statistically, differences between NPNC and cued trial response times are larger for the two social stimuli, with 17.183 ms average difference between cued and NPNC trials for the human face, and 16.140 ms average difference for the line drawing, compared with 7.548 ms average difference for the arrow.

Some stimuli were tested at 400 ms SOA (line drawing, arrow, and Keepon) while others were tested at a 600 ms SOA. This represents a methodological change undertaken partway through the experiment, in order to align more precisely with previous research. Both SOA times are within the threshold for "short" SOAs as described by Friesen *et al.*, and reflexive cueing effects have been found at up to 600 ms SOAs [90, 246]. Therefore, we believe these SOAs to be comparable.

This study provides some of the first insight into cognitive processing of robot

stimuli, using psychophysical techniques common in cognitive psychology but largely unused in the field of human-robot interaction (HRI). There is significant information to be gained from analyzing the cognitive effects of robots on human attention, both for cognitive scientists interested in which features cue attention, and robot designers interested in creating robots that engage in natural social interactions with people. Robot stimuli provide a "real life" testbed for cognitive attention experiments by allowing researchers to manipulate robotic features to test theories about what features cue reflexive attention. Robot designers can use this information in their designs, which would improve robot usability by allowing people to employ the same social cues with robots as they do in human-human interactions. The current study suggests that these two robots, Zeno and Keepon, are unable to cue human attention the way real faces, schematic faces, or even arrows do. These results should be further explored to identify what kinds of visual manipulations can make robots appear reflexively social.

3.5 Summary

Human eyes elicit strong attentional shifts in the direction of their gaze, even when this shift is detrimental to viewers' goals, while non-social directional cues such as arrows have demonstrated weaker attentional cueing effects. Little evidence currently exists for the cognitive effects of robot gaze cues, however. Using an established counterpredictive cueing experiment, we analyzed the attentional influence of two robots that vary in level of anthropomorphism, and compared our findings to attentional effects of human faces and arrows. Results indicate that human faces, robot faces, and arrows all conveyed directional information to participants, but that neither robot stimulus showed attentional cueing effects.

These findings confirm that common directional symbols (particularly human

faces) engage an automatic attention shift to the directed location despite top-down motivations to attend elsewhere. However, the findings also reveal that robot faces do not elicit this counterpredictive cueing effect. In other words, there is a difference in how robot eye gaze is cognitively processed in the moments after it is seen, when compared with human eye gaze.

Many HRI studies use what we call a *macro level* analysis: measurements of observable behavior are taken once for a trial or a whole interaction, which could last an hour or more. This analysis provides a holistic view of human behavior in a human-robot interaction. For example, macro level analysis has identified the general effectiveness of robot gaze in cueing conversational roles [173] and improving recall of stories [172].

In contrast, the work in this chapter is evaluated at a *micro level* of analysis. A micro level analysis brings us closer to evaluating actual cognitive responses by measuring people's rapid, short, or automatic responses. The metric used here—millisecond-level response times—analyzes human behavior on a much shorter time scale than typical HRI studies. On this time scale, we identified a difference between human and robot eye gaze processing.

This difference in micro level response to human and robot eyes emphasizes the need to understand the cognitive effects of robots by analyzing micro level behaviors in addition to the more commonly measured macro level behaviors. Section 10.1.2 further describes the difference between micro and macro levels of analysis.

This research is the first to apply psychophysical techniques to the analysis of cognitive effects of robot appearance. Further experimentation using these techniques might reveal what features influence natural social responses, and how robots can be designed to take advantage of existing cognitive biases.

4

Robot Gaze and the Perception of Attention*

Establishing mutual attention is the first step toward successful social interaction. Therefore, identifying how to best convey a robot's attention is important for the development of social robots. In this chapter, we try to understand the dynamics of robot eye gaze for conveying attention. Specifically, we investigate how the timing and frequency of user-directed gaze influences how people perceive a robot's attention. Participants viewed a group of My-Keepon robots executing random motions, occasionally fixating on various points in the room or directly on the participant. We varied type of gaze fixations within participants and robot group size between participants. Results show that people are more accurate at recognizing shorter, more frequent fixations than longer, less frequent ones, and that their performance improves as group size decreases. From these results, we conclude that multiple short

*This work was originally published as:
Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. Are you looking at me? Perception of robot attention is mediated by gaze type and group size. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 389–396, 2013.

gazes are preferable for indicating attention over one long gaze, and that the visual search for robot attention is susceptible to group size effects. This experiment used a novel (within HRI) multi-robot setup, for which we developed a set of inexpensive, easily replicable, programmable robots built on the basis of MyKeepon toys. In this chapter, we also describe the development of these robots, which have the potential to become a more widely used HRI platform.

4.1 Introduction

Eye gaze is a critical component of typical social interactions. We use gaze to indicate attention, whether toward a speaker or toward an object of mutual interest. However, subtle gaze timing can have a strong effect on realism and comfort in an interaction. Gaze fixations that are too short can be interpreted as shyness, avoidance, or disinterest. Gaze fixations that last too long can appear menacing or uncomfortable. With the development of real-world robotic systems comes a need to understand and use gaze cues effectively.

Human-human conversation partners frequently direct their gaze toward the person to whom they are listening or speaking [28, 254], using mutual gaze to signify attention. Robot gaze seems to be leveraged just as well as human gaze; for example, people use both human gaze [104] and robot gaze [232] to successfully disambiguate referential utterances.

In the current work, we seek to understand which features of a robot's gaze make that robot appear to be attending to someone. There are many components of gaze behavior: frequency, duration, and locations of fixations; scan paths taken to reach fixation points; congruency of fixations during mutual gaze and joint attention. Because making eye contact is a strong signifier of attention, in this study we fix the

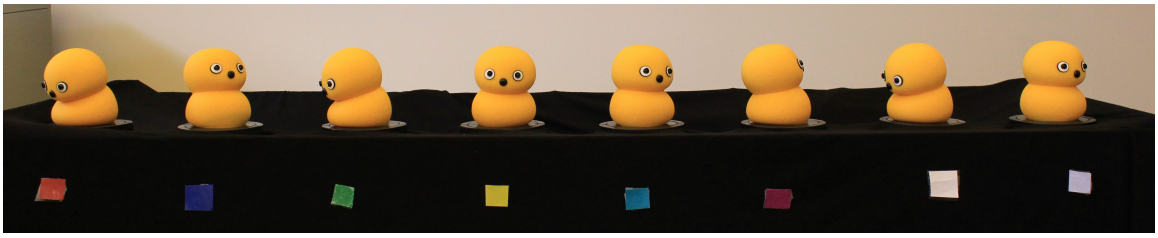


Figure 4.1: A photograph of a participant's view of the eight robot condition. The fourth robot from the left (with a yellow label) is fixating on the participant; the other robots are gazing elsewhere.

gaze location on the participant, and manipulate duration and frequency of fixations. We contrast gaze behaviors along a spectrum, from short, frequent glances to longer, less frequent stares, all directed at the user.

To quantitatively examine the effects of these two gaze types on the perception of attention, we measure the detection rate of attention fixations (i.e., fixations directed at the user) over three conditions on the spectrum of gaze types. To measure the detection rate of fixations, we present the target (the robot displaying attention fixations) among a number of distractors (identical robots displaying fixations not directed at the user, Figure 4.1). This visual search method is common in psychophysical studies [262], though it is (currently) uncommon for HRI experiments.

In addition to gaze type, we also manipulate group size to identify whether the number of distractors has an effect on participants' ability to recognize attention. The experimental procedure is described in Section 4.4.

Our search for an effective robot platform that could be used in a multi-robot experiment led us to create programmable research tools out of a readily available toy called MyKeepon. Section 4.3 describes the hardware and software modifications we used to create this novel research platform.

4.2 Related Work

Gaze recognition develops early and remains critical for non-verbal communication throughout life. Newborns [35] and older infants [51] show preferences for open eyes over closed ones. Adults are highly accurate at detecting another person's face-directed gaze during normal conversations [28]. In four-person conversations, researchers found an 88% probability that the person being looked at is the one being listened to, and a 77% probability that the person being looked at is the one being spoken to [254]. Eye gaze is also a useful cue in disambiguating referential expressions in dialogue. In an experiment where conversation partners verbally referenced objects on their displays, participants successfully used gaze cues to distinguish between competing referents before the linguistic point of disambiguation [104]. Interestingly, people tend to overestimate the amount of gaze directed at their own faces, mistaking a look over their shoulder for a gaze to their face [28].

In visual search tasks, where participants need to pick out one unique item from among a group of distractors, eyes gazing straight ahead are more easily detected among left-gaze and right-gaze distractors than either averted gaze image is from among the other two stimuli [257]. Mutual eye gaze also leads to faster processing, such as categorization of gender and access to semantic knowledge, than averted gaze [162].

Eye tracking studies reveal that gaze is affected by context. Head-mounted eye trackers show that gaze is task-driven, and that fixation locations are determined by the task at hand and learned over time [106]. Dual eye tracking has shown that the occurrence of mutual gaze, where two partners look at each other, depends on the dynamic interplay of behaviors and characteristics of both partners [52].

Functional MRI studies identify differences for processing different features of gaze. Gaze duration is processed in the medial prefrontal cortex, an area that is involved with more complex metacognitive processing, which is distinct from the

brain region processing gaze direction [50, 147]. In other words, gaze duration is a distinct feature which is processed independently of other gaze features. The intensity of brain activity in response to gaze shifts is modulated by context; fMRI studies show that activity in the superior temporal sulcus is affected by whether a virtual agent correctly or incorrectly shifts its gaze toward a target [50, 189].

As described in Chapter 2, a number of studies have tried to improve human-agent communication through appropriate agent gaze behavior, both in robotic systems [39, 117, 172, 173, 223, 248, 264] and in virtual intelligent agents [59, 102, 192]. For instance, researchers found that a robot that responds to and maintains joint attention improves task performance and receives higher ratings for competence and social interactivity than a robot that does not display joint attention behaviors [117]. Unlike eye gaze, however, people are sensitive to a robot's direct gaze but not to a nearby indirect gaze [120].

Using eye tracking, researchers found that participants follow a robot's gaze, even when the task does not require them to do so [232]. They also found that when a robot's gaze and utterances are congruent, participants can judge utterances more quickly than when gaze and utterances are incongruent. On the other hand, when examining millisecond-level psychophysical responses, robot gaze does not cue the same reflexive attention shifts that human gaze does, instead seeming to be susceptible to top-down control [2].

In this chapter, we are interested in how gaze frequency and duration affects the perception of attention. Some previous work attempts to specifically investigate these features of gaze during interactions. One such study found that a speaker looked at the face of an addressee between 25% and 56% of the time, depending on how many other people were involved in the conversation [173]. Researchers found that gaze switch timings consistent with human timings appeared more natural than gaze switches that occurred with every speech utterance [248]. Too much gaze was also a

problem, however: high levels of mutual attention without valence or responsiveness decreased rapport with a virtual agent [261].

Research in joint attention has also investigated gaze timings. One study found that a person's gaze dwelled on a referenced object for approximately 1.9 seconds on average, with no statistical difference in the amount of time spent looking when the referencer was human or a virtual agent [192]. Another such data-driven study of micro-level behaviors found that participants look at a communication partner's face (whether human or agent) within about 800 to 900 milliseconds after their partner's head movements and 600 to 700 milliseconds after naming an object for their partner to learn [269]. Participants spent longer fixating on a robot partner's face than a human partner's face, however.

4.3 Programming MyKeepon

In order to examine the effects of gaze duration on the perception of attention, we sought to use a robot platform with highly salient visual features (e.g., eyes) with an otherwise simple appearance. Keepon is a small, yellow, snowman-like robot with two eyes and a nose, but no other facial features. Originally designed for applications such as autism therapy, Keepon is a socially evocative robot that has been shown to elicit various social behaviors from children and adults [144]. The original research-grade robot is easy to control but expensive to buy, making it infeasible to use in our current study, which requires multiple robots. Fortunately, a version of Keepon is available as an inexpensive consumer-grade toy under the name MyKeepon from BeatBots LLC. In this section, we describe how we converted MyKeepon toys into programmable research tools. For more details and photographs of the process, please see our website at `http://hennyadmoni.com/keepon/`.

MyKeepon has four degrees of freedom (DOFs) using three DC motors. It can lean

forward and backward, lean left and right, rotate clockwise and counterclockwise on its base, and bob up and down. For this project, we number the motors arbitrarily: motor one controls rotation on the base, motor two controls left/right lean and bob, and motor three controls forward/back lean. Motor two's control is switched between its two DOFs using a small geared rocker mechanism; we found this mechanism difficult to control and therefore we only employ motors one and three in this experiment.

In the toy version of MyKeepon, motors are controlled by an internal circuit board. To take control of the robot's motors, we circumvented the internal board and soldered wires directly to the leads of each motor. We removed MyKeepon's internal control board along with the microphone, speakers and battery housing.

We use Arduino, an open-source hardware platform, as a control interface from computer to robot motors [25]. Each MyKeepon robot is attached to one Arduino Uno and one Arduino Motor Shield, which plugs into the Arduino Uno and is designed to run up to four DC motors. Each Arduino Uno is connected to the computer through its USB connector; when controlling multiple robots, we use a USB hub between the Uno boards and the computer.

The USB connection to the Arduino Uno allows us to send commands from the computer to the motors over a serial connection. To ensure replicability between participants, robot motions are pre-scripted, though they can be calculated and sent in real time. Each robot's motions are designed on the computer, then sent at the appropriate time to the Arduino Uno board attached to the robot. Commands are cached on the board until execution time, at which point the commands are played back sequentially, causing the motors (and the robot) to move. Figure 4.2 shows the hardware setup with control computer, USB hub, Arduino Uno and Motor Shield pairs, and MyKeepon robots. Though only three robots are shown in this figure, the setup is similar for any number of robots.

The simple DC motors in MyKeepon robots are less sophisticated than typical

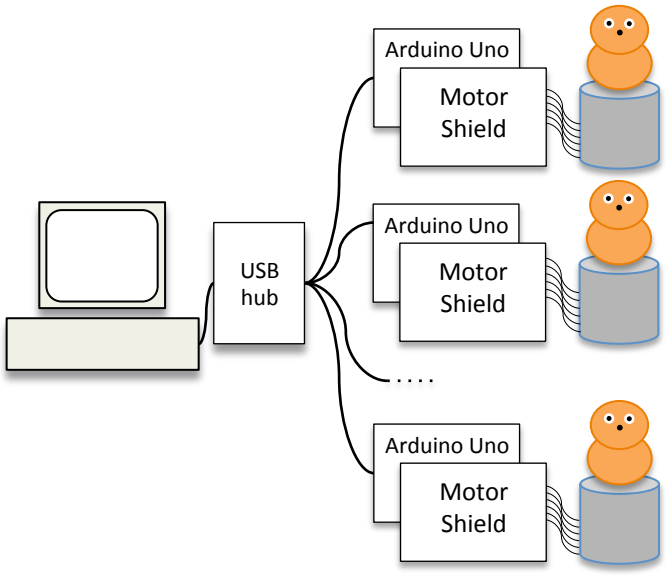


Figure 4.2: A diagram of the hardware setup. Robot motors were wired to an Arduino Motor Shield, which paired with an Arduino Uno to receive commands from a computer via USB.

high-precision motors used in research, specifically in the absence of encoders to report precise positioning. We compensated for this limitation with hand-tuning when necessary, but these motors are the major limitation for using MyKeepon as a research platform.

Several pieces of code design, transmit, and control robot motions, some running on the computer and others running on the Arduino boards. Computer-based code includes a movement generator that automatically designs robot motions given some criteria, such as direction and duration of gaze fixation. A Python script interfaces with code running on the Arduino by sending movement commands over a serial connection via USB. For instance, the Python command

```
move(keepon_ID, motor_num, time)
```

moves `motor_num` for `time` milliseconds on robot with ID number `keepon_ID`. The robot's ID number is hard coded on its Arduino board.

We use a publicly-available package called AFMotor to control the DC motors in MyKeepon from the Arduino. To communicate with this low-level control, we wrote

a state-based controller that listens for move commands arriving at the serial port and issues appropriate calls to AFMotor.

4.4 Experiment

We conducted an experiment using these programmable MyKeepons to evaluate the effects of gaze duration and group size on the perception of attention. The experiment was a mixed 3 (group size) x 4 (gaze duration) between- and within-subjects design. Participants viewed a group of robots (four, six, or eight, between-subjects) making simultaneous random motions. Between random movements, each robot occasionally fixated its gaze on various positions in the room for a given duration (zero, one, three, or six seconds, within-subjects); during these occasional fixations, a specific robot (the *target* for that trial) always fixated on or near the participant. All robots fixated for the same duration in a single trial and the total duration of fixation was held constant among trials; robots fixated six times on a one-second fixation trial, twice on a three-second fixation trial, and once on a six-second fixation trial. This inverse relationship between duration and frequency evokes the appearance of different gaze types, from frequent brief glances to longer stares. Each robot was the target in an approximately equal number of trials. After each trial, participants recorded which robot they thought was paying attention to them, as well as their confidence in that decision.

Our hypotheses are as follows:

H1 The type of gaze fixation affects accuracy: multiple short glances will be easier to detect than fewer longer fixations.

H2 The size of the group affects accuracy: more distractor robots will make it harder to detect the gaze of the target robot.

MyKeepon motor control is somewhat imprecise, so perfectly direct gaze toward participants is difficult to achieve. Each robot's movements were hand-calibrated to assure fixation toward the participant's location, though assuring that target robots directly oriented toward participants was challenging. In the experiment we report below, target robots fixated on or near participants on the target trials. Though robot fixations were not as precise as human fixations, this only served to make the task more challenging and to strengthen the results. Despite their imprecision, robot motors tend to be consistent, so whatever errors were present in target fixations likely existed for all participants.

4.4.1 Apparatus

MyKeepon robots were placed side-by-side in a containment apparatus which was covered in a black cloth (Figure 4.1). The apparatus was approximately 152cm wide by 61cm deep by 15cm tall. The robots were placed side-by-side with about 20cm from the center of one robot base to another. Figure 4.3 shows an overhead schematic of the experiment setup.

In the six robot condition, the two outermost robots were removed, and in the four robot condition the four outermost robots were removed, so that the robots present were always centered within the apparatus. Colored labels on the front of the box are used during the experiment to refer to robots. We chose not to use numbers in an attempt to avoid ordinal effects.

Participants were seated about 152cm away from and centered on the midpoint of the apparatus. At this distance, the total robot display subtended approximately 25° of the participant's visual field in the 4-robot condition (69 cm), 38° in the 6-robot condition (104 cm), and 49° in the 8-robot condition (140 cm). Each robot (9 cm across) subtended 3.4° of the visual field, and an individual robot eye (1.3 cm) subtended approximately 1° of the visual field. Although the size of an individual

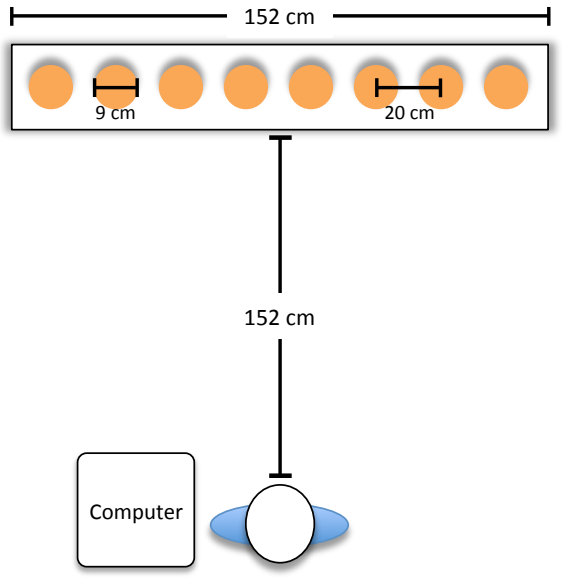


Figure 4.3: Overhead schematic of the experiment setup showing robots, containment apparatus, computer and participant. This figure is not to scale.

robot's eye is quite small, the eyes are fixed to its body, so the robot moves its entire body to orient its gaze. In every condition, participants needed to move their eyes and possibly their heads to foveate on every robot.

4.4.2 Procedure

Fifty-three participants (20 male, 33 female) took part in the experiment. The experiment took approximately 30 minutes, and participants were paid for their time. Each participant was randomly assigned to one group size condition (four, six, or eight robots).

Each participant viewed 30 trials, and each trial was comprised of 30 seconds of pre-scripted movement. In each trial, the robots exhibited automatically generated random motions in two DOFs, leaning forward or back and rotating clockwise or counterclockwise on their bases. The two DOFs could move simultaneously, causing the robots to appear to be looking around the room. At approximately equally spaced intervals, but not necessarily simultaneously, each robot stopped its motion for a set

amount of time before returning to performing random motions; we call this a *gaze fixation*, and the apparent location toward which the robot is oriented is its *fixation location*. For the target robot in a trial, the fixation location was always the participant; other robots oriented toward various points in the room during their fixations. For example, in a three-second fixation trial, the target robot fixated on the user and distractor robots fixated on various locations approximately every ten seconds, though the fixation periods did not necessarily overlap. Because behaviors between fixations were random, the robots had to move different distances from different positions to return to their fixation locations throughout the trial. Trial presentation order was randomized across participants.

Robots fixated for zero, one, three, or six seconds per trial. Zero-second fixations were a control, in which robots did not stop their random movements. One-second fixations were selected based on preliminary testing, which revealed that one-second fixations are brief enough to be difficult to identify in the eight robot condition. Six second fixations were chosen to be easily recognized, and three seconds was chosen as easily divisible to maintain total fixation duration in a trial. There were six zero-fixation trials and eight of each other fixation duration for a total of 30 trials.

Participants were seated next to a computer, which recorded their results and controlled the robots. Participants began a trial by clicking a "Start" button displayed on the computer monitor, which initiated the robots' pre-scripted movements for that trial. At the conclusion of each 30 second trial, a screen appeared on which participants selected which robot they believed was paying attention to them. They assigned a confidence value (from 0 to 100) to their choice by using a slider bar with whole-number increments. If they were able to make a decision sooner, participants could press the "Enter" key on the computer's keyboard to bring up the selection screen immediately, though the robots continued to move for the full 30-second trial. Before data recording began, participants engaged in two practice trials under exper-

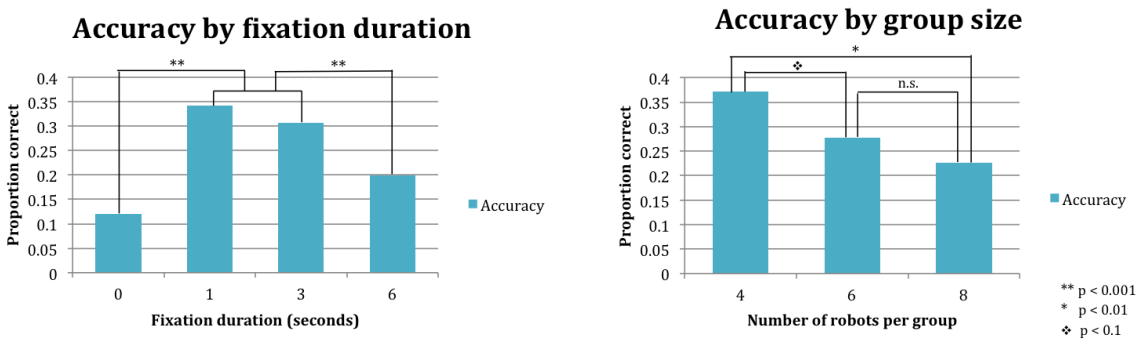


Figure 4.4: Proportion of accurate responses as a function of fixation duration (left) and group size (right). Significant differences are marked.

imenter supervision.

4.5 Results

Two participants were excluded from analysis due to technical malfunction. Additionally, four individual trials were excluded due to failure to respond or error in recording a response. We analyzed the results of 51 participants (25 in the eight-robot group, 19 in the six-robot group, and 7 in the four-robot group), for a total of 406 trials of one-second fixations, 408 trials of three-second fixations, 406 trials of six-second fixations and 306 trials of zero-second fixations across all robot group sizes. Figure 4.4 shows average accuracy split by fixation duration and by group size.

We conducted a mixed-model repeated measures ANOVA with fixation duration (0, 1, 3, or 6 seconds) as the within-subjects repeated variable and group size (4, 6, or 8 robots) as the between-subjects variable. There is significance for fixation time ($F(3, 144) = 17.503, p < 0.001$) and group size ($F(2, 48) = 5.105, p = 0.010$). There is also a significant interaction effect ($F(6, 144) = 3.554, p = 0.003$).

Pairwise comparisons of fixation duration reveal that one- and three-second fixations led to significantly higher accuracy than either zero- or six-second fixations ($p \leq 0.001$ in all cases). There was no statistical difference in accuracy between

zero- and six-second fixations, or between one- and three-second fixations. Given that zero-second fixations had accuracy of about chance for each group size (0.24 for four robots (chance is 0.25), 0.13 for six robots (chance is 0.17), 0.08 for eight robots (chance is 0.13)), six-second fixation accuracies were not statistically different than chance, though one- and three-second fixation accuracies were significantly better than chance.

Post-hoc analysis of group size using a Bonferroni correction found that accuracy in the four robot group was significantly better than accuracy in the eight robot group ($p = 0.007$) and marginally better than accuracy in the six robot group ($p = 0.054$). No statistical difference was found for accuracy between six and eight robot groups.

4.6 Discussion

H1 predicted that accuracy would improve as gaze fixation changed from long and infrequent to short and frequent. Results support this, with statistically significant differences in accuracy between short (one- and three-second) and long (six-second) fixations. Since total fixation time in a single trial was held constant, this suggests that multiple short fixations are better at conveying attention than fewer longer fixations. We predict that it is the transition from motion to gaze fixation, rather than the fixation itself, that cues the perception of attention. Therefore, people may be responding to "fixation events"—the transition between movement and fixation—rather than to active gaze. There were six times as many fixation events in the one-second condition than in the six-second condition, perhaps accounting for improved accuracy on shorter fixation trials. If our prediction is correct, this would be an interesting finding about human gaze processing.

H2 predicted that group size would have a negative effect on accuracy, which was also supported. Results show significant differences in accuracy rates between

the four-robot group and the other groups, with accuracy rates of 37%, 28%, and 26% over all fixation durations, respectively. This is consistent with findings from a visual search task, where more distractors led to a degradation in performance when detecting straight eye gaze [257]. These results suggest that eye gaze, while an important social factor, does not cause a "pop out" effect like some more basic stimuli.

MyKeepon robot motors are imprecise but consistent, so the target robot's gaze was offset by some small amount from a perfectly direct gaze on many trials. People overestimate the amount of gaze directed toward their face [28], but even so, this gaze offset may explain generally low accuracy rates. On the other hand, the motors appear to be consistent across many trials, so we are confident that stimuli were consistent across presentations. We present our results with the understanding that fixation errors are higher with robots as stimuli than they would be with humans.

Meaningful eye gaze consists of many features in addition to frequency and duration of fixations. For example, the velocity of a saccade and the scan path also reveal socially relevant information. Furthermore, different behaviors such as gaze following, joint attention, or attention maintenance may require the use of different features of gaze. The current experiment breaks down the complexity of social gaze by isolating frequency and duration in the context of indicating attention. A full exploration of gaze necessitates understanding all the features of gaze and their interplay, and is a rich avenue for future research.

One difference between this and most other HRI studies is the use of a multi-robot setup. Human-human and human-robot interactions outside of the laboratory do not occur in a vacuum. There are competing visual stimuli in many real-world tasks that draw attention away from a visual target. In multi-robot domains, gaze features combine not only within a single robot's behaviors but across robots, making for a complex visual scene. To make progress toward a more holistic understanding

of HRI, we must continue to explore visual attention under distracting and difficult conditions.

In Chapter 3, we presented the concept of micro and macro levels of analysis. Micro level analysis focuses on very short or very small changes in behavior (like millisecond-level response times, or eye saccades) that reveal cognitive processes. In contrast, macro level analysis focuses on behaviors that can be seen, often on the individual trial level or even the entire interaction level, which provide a more holistic view of human behavior. The current work falls somewhere on the spectrum between micro and macro levels of analysis. The measurement techniques (accuracy of identification) and methods (visual search of a target from among distractors) is drawn from psychophysics, and addresses underlying cognitive effects. However, the task (identifying attention through gaze) operates more on the macro scale.

As with many lab-based experiments, we must consider how transferable the findings of our study are to real-world interactions. The current work is a necessary step, but not a final point, on the path to understanding natural gaze. Given that, our work yields some suggestions for the design of robots and robot behaviors in HRI. Because frequent short glances were more easily recognized in this experiment, looking at a user to initiate or maintain an interaction may be most effective using short, frequent glances, rather than an extended stare. This is supported by research suggesting that an agent that maintains mutual gaze for an extended duration (without other social gestures) leads to strongly negative responses from users [261], but that gazes that are too short and frequent also hinder communication [248]. It would be interesting to identify whether this short-and-frequent gaze preference is also present in other gaze scenarios like joint attention, where gaze is directed toward an object of mutual interest, rather than at the user themselves. Because context plays a role in the control of eye gaze [50, 189], an experiment that tests gaze features during task performance (like providing driving directions) might reveal different features at work

in signifying attention.

4.7 Summary

Eye gaze is an important part of establishing attention at the beginning of social interactions. In human-robot interactions, robots can also use gaze to convey attention to a user, but features of the robot's gaze influence the effectiveness of such gaze behavior. Using a novel (in HRI) experiment design involving the identification of a target amidst distractors, we investigated how the type of gaze (short, frequent glances versus long, infrequent stares) impacted the perception of attention from a robot. Our study found that shorter and more frequent glances were more effective at conveying the robot's attention to the human user. Additionally, we investigated how the number of distractors affected people's accuracy at identifying the target robot. Results indicate that robot gaze is subject to group size effects, with accuracy decreasing as the number of distractor robots increases. This indicates that robot gaze is not a pop-out effect, but requires serial cognitive processing.

Our results have implications for robot designers who want to make their robots appear to be attending to users, as well as for psychologists who want to understand gaze in human-robot interactions. Based on the study conducted here, we suggest that robots use shorter and more frequent user-directed gaze behaviors to indicate their attention to a human interaction partner.

Attention-establishing gaze is just one type of gaze behavior robots can perform to improve the fluency of human-robot interactions. In the next chapter, we look at another type of gaze—referential gaze—which is used to refer to objects and locations in space. We combine this referential gaze behavior with verbal references, and we explore what happens when there are incongruencies in this multimodal communication.

5

Handling Errors in Multimodal Communication*

In human-robot collaboration, communication can occur through several modalities simultaneously: a person may be speaking while also augmenting their speech with eye gaze and other nonverbal behavior. In this chapter, we explore such multimodal behavior for human-robot referential communication. We investigate how a robot's referential gaze behavior—looking at an object—can augment spoken verbal references. We also explore how conflicts in multimodal communication (i.e., mismatches in gaze and verbal references) affect performance on a cooperative referential task. Participants play a selection game with a robot, in which the robot instructs them to select one object from among a group of available objects. We vary whether the robot's gaze is congruent with its speech, incongruent with its speech, or absent, and we measure participants' response times to the robot's instructions. Results indicate that congruent speech facilitates performance

*This work was originally published as:
Henny Admoni, Christopher Datsikas, and Brian Scassellati. Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci)*, pages 104–109, 2014.

but that incongruent speech does not hinder performance. Finally, we investigate participants' performance when their interaction partner is a human agent rather than a robot, and find the same results: in this type of activity, congruent gaze helps performance while incongruent gaze does not hurt it. We conclude that robot gaze may be a worthwhile investment in such situations, even when gaze behaviors may be unreliable.

5.1 Introduction

In typical human interactions, eye gaze supports and augments spoken communication [140]. People gaze almost exclusively at task-relevant information [106], and gaze is used to disambiguate statements about objects in the environment [104]. Similar mechanisms are also at play in human-robot interactions: task-relevant robot eye gaze can be used to improve the efficiency of collaborative action [44].

For example, imagine a human and robot collaboratively constructing a birdhouse. The robot can use its eye gaze to clarify an ambiguous speech reference, saying "Please pass the green block" while looking at a particular green building block to distinguish it from among other green blocks. This multi-modal communication makes the interaction more efficient by using multiple channels to convey information, requiring less investment in costly mechanisms like generating sufficiently descriptive speech, and improving the naturalness of the interaction [117].

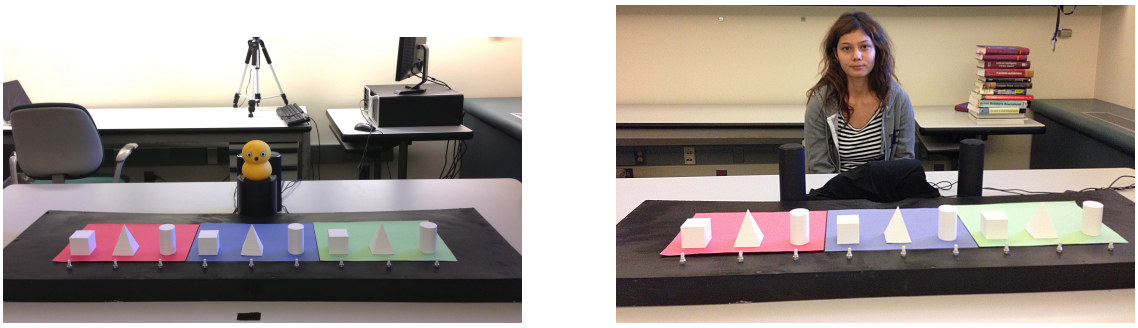
But robots are not perfect, and sometimes speech and gaze cues will conflict. Sensor errors, hardware malfunctions, and software bugs can cause mismatches between a robot's gaze and speech. In such cases, a human partner receives incorrect or contradictory information from the robot. The human might misinterpret the robot's speech or, at best, must hesitate to decide what the robot means, decreasing the collaboration's efficiency and increasing the human's cognitive load.

While a growing body of evidence shows that people can interpret robot gaze and speech, only a few studies to date have investigated the effects of speech-gaze conflicts. In this chapter, we investigate how speech-gaze conflicts are handled by human partners in collaborative, embodied human-robot interactions (Figure 5.1). We focus on object selection tasks in which a robot provides instructions to a human, because these scenarios are central to collaborative action. Misinterpreting communication in such scenarios can be costly.

We compare *congruent* gaze—in which the robot looks at the object it references in speech—and *incongruent* gaze—in which the robot looks at a different object—to a control condition in which the robot does not exhibit gaze cues. To quantitatively measure the effect of speech-gaze conflicts, we record the time between when the robot begins its instructions and when participants select an object. Response time serves as an approximation of task efficiency; faster responses mean less overall time taken for the task.

As a final manipulation, we also include a human agent condition, in which the robot is replaced by a person who performs the robot's role in the experiment. The human agent condition attempts to discover whether robot gaze is any more or less influential on human behavior than human gaze.

The results of this study provide evidence of the effectiveness of gaze in collaborative human-robot interactions. As described below, we find that congruent gaze facilitates performance in both robot and human conditions. Interestingly, we also find that incongruent gaze does not hinder performance in either the robot or the human conditions. In other words, in this task, people are able to recover quickly enough from speech-gaze conflicts that their performance is statistically no different than not having gaze at all. These results suggest that adding referential gaze may be a low-risk way to improve human performance in similar environments, even when the gaze system is unreliable.



(a) Robot agent condition

(b) Human agent condition

Figure 5.1: Participant view of the experiment. MyKeepon or a human actor provided verbal and gaze cues about which shape to select.

5.2 Related Work

Directional eye gaze seems to be a special stimulus, evoking reflexive attention shifts that are robust to top-down modulation [90]. Functional MRI studies reveal a significant overlap in the brain areas that process theory of mind and those that process eye gaze [57]. In fact, observing someone signaling the presence of an object with referential gaze elicits the same neural response as observing someone physically reaching to grasp that object [193], indicating that people use gaze as a powerful indicator of others' future behavior.

Where we look is closely coupled with what we say in human-human interactions. Objects or figures in the environment are typically fixated one second or less before they are named in conversation [100, 269]. When referencing objects, people use eye gaze as a strong and flexible cue for eliminating ambiguity [104]. When access to a partner's eye gaze is restricted, for instance because the partner is wearing sunglasses, people are slower at responding to their partner's referential communication [44].

As in human-human interactions, eye gaze is an important part of human-robot interactions. Robot eye gaze can influence whether people join a conversation or feel excluded from it [174], can influence people to favor certain objects over others [175], and can facilitate cooperative behaviors like object handoffs between humans and

robots [235]. Exhibiting joint attention, a type of social gaze, increases ratings of a robot's competency and naturalness [113].

More specifically, studies of human-robot interaction have shown that robot gaze can be used to clarify speech. If a robot gazes toward an object while naming it, people select the object more quickly than if the robot names the object without looking at it [44, 113]. With both robots [113, 172] and virtual agents [23], gazing at task-relevant objects during teaching—for instance, looking at a map while describing political boundaries—increases peoples' retention of information. In most of the literature about referential gaze in HRI, however, robot gaze is *congruent* with speech.

Some researchers have investigated the effects of speech and gaze conflicts in HRI. In a video-based study by Staudte and Crocker [232], participants evaluated the correctness of a robot's statements about objects in front of it (for instance, "the cylinder is bigger than the pyramid that is pink"). When the robot's gaze was congruent with its speech, response times were shorter than a no-gaze control; when gaze was incongruent, response times were longer than the control. This suggests that people relied on gaze to facilitate sentence processing, and that incongruent gaze hinders comprehension.

Unlike our experiment, however, Staudte and Crocker's task involved sentence evaluation rather than object selection, which requires a different cognitive skill set. Furthermore, their study was conducted with video stimuli instead of embodied robots. While virtual robots increase the ease of use and replicability of stimuli, they may not have as strong an influence on human behavior as physically embodied robots [31].

In contrast, research using an object selection task and an embodied robot finds no difference in response times between no gaze and incongruent gaze conditions, though results support the benefit of congruent gaze [113]. However, this study used a between-subjects design in which the robot exhibited only one type of gaze

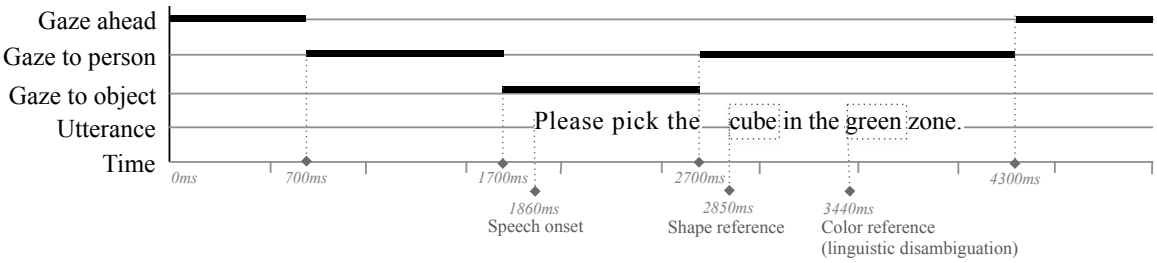


Figure 5.2: A visual representation of the speech and gaze during a typical trial. This figure shows a congruent trial: the agent both verbally and physically indicates the green cube. In an incongruent trial, the spoken word "green" is replaced with one of the other two zone colors. In a no-gaze trial, the agent gazes straight ahead.

(congruent or incongruent) to each participant. Participants could acclimate to the robot's gaze strategy, which does not address situations where gaze is usually helpful but occasionally incorrect.

The current work is inspired by these studies, and builds upon them by investigating conditions in which speech and gaze are incongruent rather than only congruent [44], using a physically embodied robot rather than a video [232], and introducing uncertainty about the robot's reliability to avoid habituation to one particular condition [113].

5.3 Experiment 1

This experiment is designed to investigate whether gaze conflicts hinder task performance in collaborative human-robot interactions. Participants engaged in an object selection task with a robot. On each trial, the robot provided spoken instructions of the form "Please pick the [shape] in the [color] zone" where shape and color referred to objects in front of the participant (Figure 5.1). Each of the nine objects was referenced nine times during the interaction, for a total of 81 trials.

On each trial, the robot also provided a gaze cue, which was either congruent with the speech (i.e., looking at the same object), incongruent with the speech (i.e., looking at a different object), or absent (no movement). The robot started each trial

in a neutral position, with gaze directed straight forward and approximately 30 cm below the participant's eyes. To initiate a gaze cue, the robot first attempted to establish joint attention by looking up at the participant's face (mutual gaze), and then engaged in object reference by looking down toward the selected object, then returned to look at the participant's face before returning to the neutral position (Figure 5.2). The robot did not use sensors to confirm that mutual eye gaze was successful; instead, the experiment was pre-scripted and ran autonomously. In no gaze trials, the robot did not move at all and continued looking ahead in neutral position.

In human-human communication, eye gaze moves toward an object prior to a verbal reference and away from the object just as it is named [269]. We carefully aligned the verbal and gaze cues to mimic natural behavior (Figure 5.2).

We measured how much time participants took to select a shape. By comparing response times in the congruent, incongruent, and no gaze conditions, we are able to determine whether gaze has any facilitation or hindrance effect on the speed with which people respond to the robot's instruction.

5.3.1 Apparatus

The experiment apparatus is a black box measuring 120cm by 40cm by 6cm (Figure 5.1). The robot was placed on the table opposite the participant, approximately 80cm away, with the box between them. Three zones are marked by colored paper on top of the box: red, blue, and green. Each zone contains an identical set of white blocks in simple shapes—a cube, a pyramid, and a cylinder—arranged side-by-side in a single row on top of the zone. A momentary pushbutton switch in front of each object is used to select that object, and the precise timings of button presses are recorded on a nearby computer.

We used the same low-cost MyKeepon robot platform described in Chapter 4

(Figure 5.1). This 32cm tall, snowman shaped, interactive robot toy has a rubber yellow skin and four degrees of freedom: rotation around the base, left/right lean, front/back lean, and up/down bob. MyKeepon is a consumer-grade version of a research robot called Keepon Pro, which was designed to be a socially evocative but simple robotic agent [144]. The robot's minimalist design and salient eyes make it a useful platform for HRI studies about eye gaze.

We modified a MyKeepon to make it programmable for this experiment using the same mechanism as in Section 4.3. The MyKeepon internal microprocessors were connected to an Arduino Uno, an open-source electronic prototyping platform. Using the I2C bus on the MyKeepon microprocessor and open-source software [166], the Arduino sends motor commands and retrieves information such as encoder positions from the MyKeepon hardware. This allowed for easy control of the MyKeepon robot platform.

5.3.2 Procedure

Twenty two people participated in this experiment (10 females). Their ages ranged from 18 to 34, with a mean age of 22, and most were Yale undergraduate students. Participants were compensated $8.

Participants were told that they would play an object selection game to help evaluate a new robot platform. They were shown the nine shapes and told that in each round of the game, the robot would provide instructions on which shape to choose. Participants were informed that they should select the shape as quickly and accurately as possible. They were also told to return their finger to a marked start position on the table between trials. This instruction was given to eliminate any "hovering" over the buttons so that response times are consistent across trials.

In each trial, a computer-generated voice provides a verbal cue, which is a sentence that first indicates the type of object and the zone the object is in, for example,

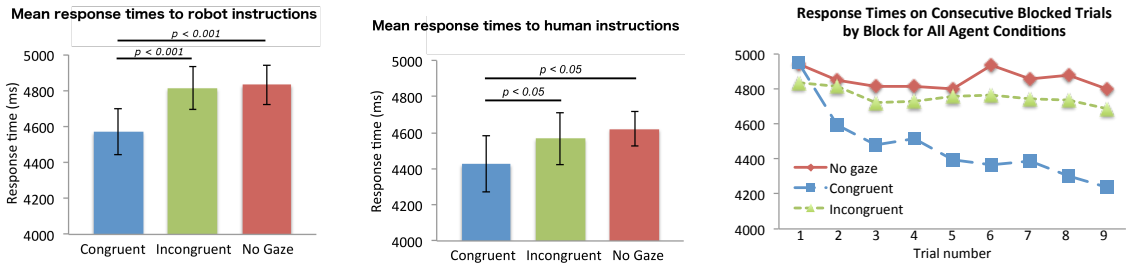
"Please pick the cube in the green zone" (Figure 5.2). The sentence is constructed so that the specific object referred to by the sentence remains ambiguous until the color of the zone is stated near the end of the sentence. Until this point of linguistic disambiguation, there are three potential matches for the sentence (the named shape in the red, blue, and green zones), so participants cannot select an object with more than 33% reliability.

Simultaneous with the verbal cue, the robot also provides a gaze cue by orienting toward one of the shapes. On congruent trials, the robot turns toward the shape named by the verbal cue. On incongruent trials, the robot turns toward a different shape at least three spots away from the correct shape. This restriction ensures that there is no confusion about whether the gaze was directed toward the correct shape. On no gaze trials, the robot remains looking straight ahead.

Participants first practiced two congruent gaze trials under experimenter supervision to familiarize themselves with the task; these practice trials were not recorded, and participants were not told that the robot's gaze would vary in other trials. After the practice, each participant experienced two sections of the experiment with no breaks between them.

In the first section, called the *blocked* section, participants saw each trial type blocked together: first nine no gaze trials, then nine congruent trials, then nine incongruent trials, with no demarcation between the blocks. The purpose of the blocked section is to establish a baseline measure of reaction time (in the no gaze block) and to observe how performance changes as participants become familiar with the robot's gaze.

The second section of *randomized* trials followed the blocked section immediately. During the randomized section, each participant saw a unique random ordering of all 54 combinations of shape, color zone, and gaze type. The purpose of this section is to measure the effects of gaze cues when participants did not know whether the cue



(a) Experiment 1, robot agent.

(b) Experiment 2, human agent.

(c) Blocked trials, both agents.

Figure 5.3: Response times to agent instructions. Figures (a) and (b) show mean response times across all trials for each experiment. Figure (c) shows the blocked trials section separated by trial number. In all figures, congruent gaze facilitates response times, while incongruent and no gaze conditions show no significant difference. Error bars show 1 standard error.

would help or not.

After both sections, participants were given a survey with demographic questions and one free-response question: "Did you notice anything unusual about the robot's behavior?"

5.3.3 Results

Twenty-two participants each completed 27 blocked trials and 54 randomized trials for a total of 1782 data points. Four trials (0.2%) were discarded because no response was recorded within 12 seconds, either because the participant did not press a button or because the button press did not register. We also discarded the no gaze blocked section for one participant (nine trials, or 0.5% of all trials) due to self-reported noncompliance. Participants were highly compliant with the verbal cue, selecting a shape that was different from the robot's spoken instruction on only five trials (0.3%). Results are shown in Table 5.1 and Figure 5.3(a).

A repeated measures ANOVA of response time by gaze type for all trials shows a significant main effect ($F(2, 42) = 43.181, p < 0.001$). Post-hoc tests with a Bonferroni correction reveal that response times to congruent gaze were significantly shorter

than response times to incongruent gaze (by 242ms, $p < 0.001$) and to no gaze (by 262ms, $p < 0.001$). There was no statistical difference between incongruent and no gaze trials.

There are several conclusions to be drawn from these results. First, people were highly accurate and highly consistent in following the robot's speech, complying with speech instructions on 99.7% of trials even though 33% of the trials included a conflicting gaze cue. The high rate of compliance with speech suggests that cases in which participants failed to follow the speech cue involved button press errors, though we did not explicitly ask participants to report errors.

When the robot's gaze indicated the same shape as the verbal cue, participants used gaze to guide their responses, as indicated by the significantly improved response times in the congruent gaze condition. Surprisingly, participants were not hindered by incorrect robot gaze: they responded no slower to incongruent trials—when gaze and speech did not match—than they did to no gaze trials, where there was no gaze cue. In other words, congruent gaze helped people respond to a robot's verbal instructions more quickly, but incongruent gaze did not make them respond more slowly than no gaze at all.

The response facilitation from robot gaze supports previous findings in HRI (such as Boucher et al., 2012; Huang & Mutlu, 2012; Staudte & Crocker, 2011). However, the lack of hindrance from incongruent gaze conflicts with previous findings [232].

To test whether this effect is due to the robot or to the task, we conduct a new experiment with a human in place of the robot. If the same procedure—now with human gaze—yields the same effect, we can conclude that the task, and not the agent, is responsible for the absence of hindrance.

Agent	Gaze type	RT (ms)	SD (ms)	N
Robot	Congruent	4572	256	22
	Incongruent	4814	235	
	None	4834	222	
Human	Congruent	4427	309	9
	Incongruent	4568	289	
	None	4621	189	

Table 5.1: Response times (RTs) for all trials in Experiments 1 and 2. RTs are measured from start of trial, including time to speak sentence. When measured from linguistic disambiguation, RTs are similar to previous work [44].

5.4 Experiment 2

We replicated Experiment 1 with a small number of participants. The apparatus and procedure are identical to Experiment 1, except that the robot is replaced by a human actor (Figure 5.1(b)). For consistency, the verbal cue is still provided by the computer-generated voice from Experiment 1. We took care to make the human gaze as similar as possible to the robot gaze; therefore, the actor practiced looking at the object for the correct duration and shifting her gaze away from the referenced object just before it was named. On the post-task questionnaire, the free-response question was changed to: "Did you notice anything unusual during the experiment?"

Nine participants (2 females) took part in Experiment 2. Their ages ranged from 18 to 20 (mean of 19). They were all Yale undergraduates and they were compensated $8.

5.4.1 Results

Table 5.1 and Figure 5.3(b) show the results of Experiment 2. A repeated measures ANOVA to test the effect of gaze type on response times found a significant main effect ($F(2, 16) = 7.892$, $p = 0.004$). Post-hoc tests with a Bonferroni correction reveal that response times to congruent gaze are shorter than response times to incongruent gaze

(141 ms, $p = 0.018$) and no gaze (194 ms, $p = 0.033$). No significant difference was found between response times in incongruent and no gaze conditions. Participants made an erroneous selection (not following the speech cue) on 20 (2.7%) of the 729 trials.

To compare robot and human gaze, we conducted an ANOVA on response time with gaze type as a within-subjects factor and agent type as a between-subjects factor. The analysis reveals a significant effect of gaze ($F(2, 86) = 6.564, p = 0.002$) but no effect of agent ($F(1, 86) = 0.351, p = ns$) and no interaction ($F(2, 86) = 0.291, p = ns$). Post-hoc pairwise comparisons on the significant result show that congruent gaze led to shorter response times than incongruent gaze (191 ms, $p = 0.017$) and no gaze (228 ms, $p = 0.003$) for all participants regardless of agent condition. No significant difference was found between incongruent and no gaze conditions.

The blocked section of the experiment reveals how participants acclimated to a consistent gaze type. Because there is no significant difference between agent conditions, we can collapse the data across these conditions for this analysis. Figure 5.3(c) shows mean response times for each trial in the blocked section, averaged across participants in both agent conditions. Recall that participants saw nine no gaze trials, then nine congruent trials, and then nine incongruent trials. The no gaze block serves as a baseline for response times without gaze. As shown in Figure 5.3(c), response times remained fairly stable during the no gaze block. Response times improved during the congruent block, indicated by the downward slope of the congruent block line. In contrast, there was no improvement of performance over the nine incongruent blocked trials. Participants performed slightly better on the incongruent block than on the no gaze block that preceded it, though this effect may be due to practice.

Although participants were never explicitly told to follow gaze, they rapidly adapted to using congruent gaze to improve their performance. The rate of improvement does not decrease by the ninth trial, suggesting that more congruent gaze

trials might have led to continuing improvements.

5.5 Discussion

For both robot and human agents, participant response times were faster when the agent's gaze cue was congruent with its verbal cue, compared to incongruent gaze and no gaze conditions. Because the gaze cue is delivered before the point of linguistic disambiguation, the fact that participants responded more quickly on congruent gaze trials indicates that they planned their motion according to the gaze cue before hearing the disambiguation. When the cue was incongruent, however, participants responded no slower than if there were no gaze cue at all. Therefore, while they use gaze to plan their motions, participants quickly recover from erroneous planning when the point of linguistic disambiguation is reached. This facilitation occurs even in the randomized section, when participants could not know ahead of time whether gaze would be congruent, incongruent, or absent. In short, current results suggest that there are scenarios in which adding eye gaze cues to a robot's behavior is a worthwhile investment: at best, it increases comprehension and efficiency, and at worst (when the gaze cue is in error), there is little damage to performance.

Other research has shown that incongruent gaze hinders performance in robot-instruction tasks [232]. However, our study's task involves a lighter cognitive load, which may explain our divergent findings. In both studies, participants identify the referent of the robot's gaze and speech, but in our task, they simply select that referent, whereas in Staudte and Crocker's task, they compare features of that referent to features of other visible objects and then decide if a given statement is true or false. Thus, our experiment's task requires less cognitive processing, which may allow people to quickly overcome the incongruent gaze. This conjecture is supported by findings from a different study that used a similar task to ours [113]. This study also found

no difference between no gaze and incongruent gaze, while confirming the benefit of congruent gaze.

An alternate explanation is that the agent looks at the participant before speaking on incongruent trials, but not on no gaze trials, which may cue participants for the impending selection and negate any hindering effects of incongruent gaze. A revised no gaze condition in which the robot looks at the participant but not to a block would clarify this possibility.

Although mean response times did not significantly differ between robot and human agents, some differences did emerge between these conditions. Participants in the human agent group responded more quickly on average, although the difference was not significant (possibly because of the small group size). Perhaps relatedly, the error rate for participants in the human agent group (2.7%) was higher than the error rate for participants in the robot group (0.3%).

In response to the post-interaction survey question asking whether they noticed "anything unusual" during the experiment, five of the nine participants in the human agent group (56%) made reference to intentional misdirection by the actor, writing things like "She built up my trust and then betrayed me" and "She tried to trick me with her gaze." In comparison, only six of the 22 people in the robot group (27%) included such statements about intentional action from the robot. Even with identical behaviors, there was some difference in agency attributions between robots and humans.

However, human gaze is inherently less precise than robot gaze. Future experiments could record the human actor's face to verify that human gaze timings were comparable to robot timings. To generalize the results, future work should also test different collaborative scenarios to understand the conditions under which facilitation is possible without hindrance. Eye tracking would reveal at which point people decide to follow or ignore a robot's gaze. Future work should also randomize the assignment

of conditions, rather than recruiting independent groups of participants, to rule out the possibility of group effects causing the observed variations.

MyKeepon has limited articulation and a simplified appearance. We chose this robot intentionally—MyKeepon's large eyes and gross body movements make its eye direction highly salient—but it is simpler and smaller than many other robots. Our results, therefore, may be most applicable to this type of robot. Future studies should investigate robots with articulated eyes as well as anthropomorphic robots to find whether physical appearance and eye motion affect gaze cues.

This work provides support for efforts to incorporate nonverbal behavior into robot communication. It suggests that, at least in some scenarios, the benefits of incorporating gaze into referential communication outweighs the risk that those behaviors might be incongruous with each other.

Based on the results from this study, we can begin to envision how to develop our nonverbal behavior models in Chapter 7 and Chapter 8. These models combine speech with gaze and other nonverbal behaviors to recognize and produce object references.

5.6 Summary

Congruent multimodal behavior facilitates human understanding of object references. When a robot simultaneously verbally and visually referenced an object, people identified that object more quickly than when the robot simply verbally referenced the object. In contrast, incongruent multimodal behavior did not hinder people's understanding of object references. People performed no worse when gaze and verbal references were to a different object than when there was no gaze reference at all. This pattern of effect, in which congruent multimodal behavior improves performance but incongruent behavior does not worsen performance, was also seen with a human agent instead of a robot.

Though in this chapter multimodality involved a speech channel and a gaze channel, multimodal communication can include multiple channels of nonverbal behavior, as well. In the next chapter, we investigate the effects of multimodal behavior comprised of two nonverbal channels—gaze and gesture—on people's understanding and compliance with a robot's referential communication.

6

Gaze and Gesture in Robot-to-Human Handovers*

One of the benefits of robots as physically embodied systems is their capacity to manipulate objects to assist human users. In human-human interactions, this kind of collaborative manipulation is accompanied by nonverbal communication that provides additional information about intentions. In this chapter, we examine the role of nonverbal communication in the context of collaboration, the target domain for our nonverbal behavior models. Specifically, we focus on robot-to-human object handovers, a common action during physical collaborations. Robot-to-human handovers are primarily manual tasks, and human attention is therefore drawn to robot hands rather than to robot faces during handovers. In this chapter, we show that a simple manipulation of a robot's handover behavior can significantly increase both awareness of the robot's eye gaze and compliance with that gaze. When eye gaze communication occurs during the robot's release of an object, delaying object release

*This work was originally published as:
Henny Admoni, Anca Dragan, Siddhartha Srinivasa, and Brian Scassellati. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 49–56, 2014.

until the gaze is finished draws attention back to the robot's head, which increases conscious perception of the robot's communication. Furthermore, the handover delay increases peoples' compliance with the robot's communication over a non-delayed handover, even when compliance results in counterintuitive behavior. In other words, our study reveals that people recognize and comply with a robot's referential eye gaze communication, but only when a social gesture (a handover delay) indicates that the robot's gaze is important.

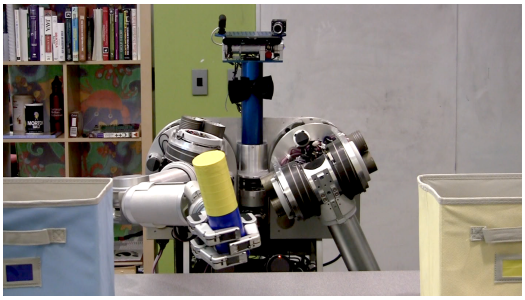
6.1 Introduction

In the future, assistive robots will help people perform manual tasks more easily and efficiently. These robots may retrieve items from high shelves [180], assist in fine motor manipulations [41], or act as extra hands during physically complex tasks [74]. One of the primary challenges for such robots will be the ability to manipulate objects in collaboration with people [1].

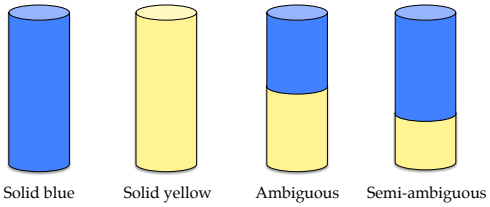
For example, imagine a robot, like the one in Figure 6.1(a), that helps a wheelchair-bound user cook a meal. This robot can move around the kitchen, grabbing the right ingredients and handing them to the user. The robot and user can also prepare parts of the meal simultaneously, passing utensils and ingredients back and forth between them. Finally, the robot can help clean up, taking items from the user and moving them to the sink.

A common task throughout this interaction is the *handover*: the act of transferring an item from one actor to another. For seamless robot-to-human handovers, the robot must generate appropriate social cues that alert the person to the what, when, and where of the handover [56, 235].

But other information, unrelated to the handover itself, may also need to be



(a) A participant's view of the block sorting task.



(b) Blocks were fully colored, ambiguously colored (50% of each color) or semi-ambiguously colored (70% of one color).

Figure 6.1: Participants engage in a collaborative manipulation task with HERB. The robot hands over colored blocks, and participants sort them into colored boxes on the table.

communicated during a handover. For example, the robot might want to indicate where to put an object after giving it. Speech is an obvious mode of communication for conveying such information, but it may not be available or effective in all situations. For instance, speech may be unavailable in a noisy room, when interacting with the hearing impaired, or when a person is already engaged in a listening task, such as holding a conversation while cooking. Even when speech is available, it is not always the most effective means of communicating: a robot that announced every handover before it occurred would hinder the fluency of interactions involving frequent handovers.

Eye gaze is an alternative means of communication that can be used when speech isn't practical. Typical humans can understand the motor intentions of others based on their gaze [193], and robots are able to influence human motor behavior using gaze (e.g., [175]). Based on this, a natural conclusion is to communicate information about where to put the object using eye gaze.

Because eye gaze requires the user to attend to the cue in order to be effective, it becomes critical to select the right time to exhibit the eye gaze cue. This communication should occur during the *transfer phase* of the handover, which starts from the point at which the giver has finished reaching with the object toward the agreed-

upon transfer location, and ends when the receiver has taken hold of the object and retracted it, signaling the end of joint activity [235]. Earlier signals specifying where to put the object may be confused with attempts to establish the what, when, and where features of the handover [235]. Signals sent after the transfer phase may be missed, because the user's attention may have already shifted to the next task location [106]. Thus the transfer phase is the ideal time to indicate temporally relevant but non-handover-related information.

However, handovers are primarily manual tasks that draw attention directly to the robot's hand and away from its eyes. Mutual eye gaze is not a necessary part of handovers [234], and because people fixate their gaze almost exclusively on areas of their environment related to the task [106], attention is often directed somewhere other than the robot's head during a handover. In our first attempts to influence human behavior using eye gaze, we found that drawing attention to the robot's face during the gaze cue was surprisingly difficult. People responded to the unfamiliar experience of robot handovers by focusing intently on the robot's hand, ignoring all nonverbal communication from the robot's head.

To address this, we introduce the idea of a *deliberate delay*—an intentional hiccup in the handover—which prompts users to shift their attention from the robot's hand to its head. In particular, we look at deliberately delaying the transfer phase of the handover by postponing the release of an object from the robot giver to the human receiver until a gaze cue has been delivered. Handovers cause user attention to be focused on the robot's hand. By manipulating the force profile during a handover—deliberately making the robot hold on to an object longer—we can draw attention to other channels like head direction.

In this chapter, we present experimental evidence that adding a deliberate delay to a handover is beneficial for communicating non-handover information nonverbally, even though this delay decreases the smoothness of the handover itself. As we show

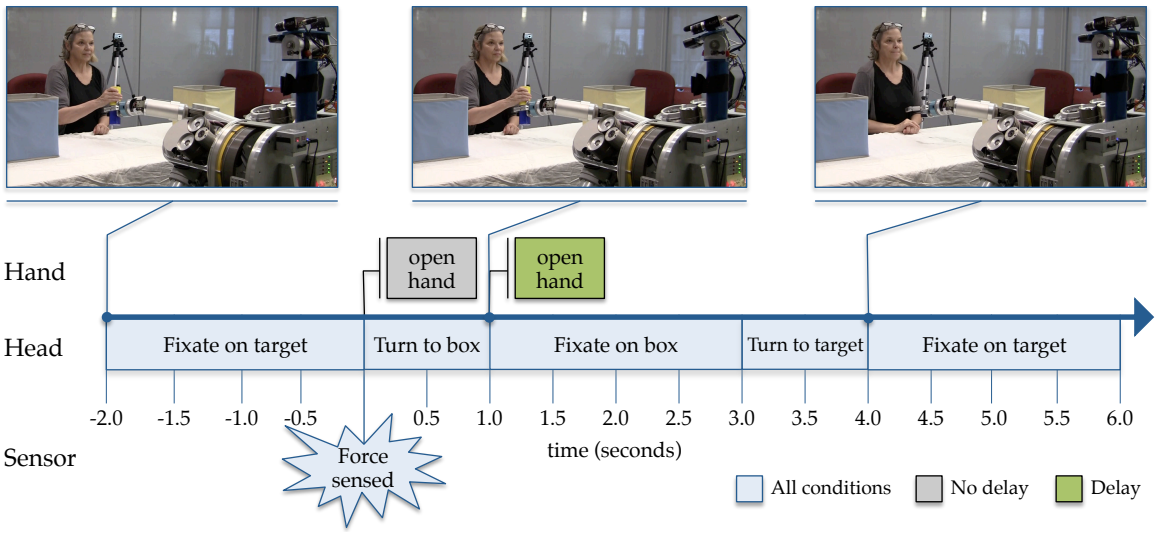


Figure 6.2: A graphical timeline of HERB's head and hand motions turning the transfer phase of the handover. HERB begins by fixating on a target (the participant's face or a mirrored point, depending on gaze condition), then looks at one of the boxes as a suggestion, and finally returns to fixate on the target. The timing of the block release depends on delay condition; it is either simultaneous with the head turn toward a box, or just after the head turn.

below, a deliberate delay not only increases attention to the robot's head, but also increases the rate at which people comply with the robot when its gaze cue leads to counterintuitive actions.

We test the effect of a handover delay on a simplified collaborative task (Figure 6.1). In this task, a robot called HERB hands colored blocks to participants, who sort those blocks into one of two colored boxes according to their personal preference. A 6-axis force-torque sensor in the robot's hand identifies when the participant has grasped the block to begin the transfer phase of the handover.

When it enters the transfer phase, the robot gives sorting suggestions by looking at one of the boxes. We manipulate when this gaze cue is executed relative to the object release during transfer: in the "no delay" case, the gaze cue and release occur simultaneously; in the "delay" case, the release is delayed until the gaze cue is complete (Figure 6.2).

We hypothesize that:

H1 A handover delay will cause people to pay more attention to HERB's gaze communication, and that

H2 Social gaze will lead people to comply more with HERB's counterintuitive suggestion.

Our results validate H1: a deliberate delay during the transfer phase of a handover causes people to pay more attention to the robot's head and to notice the robot's nonverbal gaze suggestion more frequently. More surprisingly, the delay also causes people to comply more often with the robot's sorting suggestion, even when it is contrary to their natural behavior. This compliance effect holds even for people who explicitly notice HERB's suggestion, indicating that the delay itself, and not solely increased attention, is responsible for increased compliance. Interestingly, H2 was not supported: we found no effect of social gaze on compliance.

This work lays the foundation for a new type of robot handover. Instead of working toward behavior seamlessness, researchers can build models of how people will respond to robot behavior, and select robot actions that manipulate these responses in desired ways. As shown in this chapter, this may involve leveraging nonverbal communication channels, such as eye gaze and gesture, and introducing targeted, deliberate imperfections to improve communication and efficiency.

6.2 Related Work

Our work draws from two areas of research in HRI: robot-to-human handovers and robot gaze communication. Though these areas have developed independently, they share considerable overlaps, for instance, using joint attention to signal a handover. Rather than surveying the broad fields individually, we highlight papers in each area that address the overlap between handovers and gaze communication.

Handovers can be divided into three distinct phases [152, 235]: the *approach*,

during which a giver moves toward a receiver; the *signal*, during which giver and receiver communicate their readiness for the handover; and the *transfer*, during which the object is transferred from giver to receiver.

Handovers are a primary part of collaborative robotics, and there is strong interest in automatically generating successful robot-to-human handovers [16, 81, 224, 235]. To produce a successful handover, a robot must first convey its intention to execute that handover, which requires both spatial information (a distinct handover pose) and temporal contrast (a distinct movement profile for handovers) [56, 98]. HRI studies have attempted to determine user preferences for optimal handover behavior, such as maximal arm extension [55], minimum jerk motion profiles [118], and legibility of motion [78]. Metrics for determining human preferences range from surveys and observation [55] to physiological measurements like skin conductance and eye movement [71].

The structure of human-human handovers can also be used to inform human-robot handovers [235]. Investigations of human-human handovers identified that object transfer time (from initial contact by the receiver to final release from the giver) is approximately 500 milliseconds [62].

Mutual gaze is not a predictor of handover initiation; confirming the partner's availability through asynchronous fixations is more important to successful handovers than synchronized mutual eye contact [234]. However, taking a human partner's eye gaze into account when planning a handover increases the success of robot-to-human handovers [101]. More generally, human gaze is task driven, and gaze fixations are rarely directed to locations in the world that are not relevant to the task, even if they are visually salient [106]. Fixations are instead guided by the spatio-temporal requirements of the task, arriving at the relevant location just at the point at which they are needed for task completion [127].

Robot eye gaze, however, is an important communication mode in HRI. Robots

can communicate information through gaze during tasks like storytelling [172] and teaching [23]. Robot gaze cues such as joint attention facilitate performance in cooperative tasks [44] and improve perceptions of a robot's competence and naturalness [117]. Robots can even manipulate peoples' behavior using only gaze cues, prompting people to adopt certain conversational roles [174] or select certain objects from a set [175], even without people consciously registering the gaze cue. Robot gaze is clearly an informative communication channel in human-robot handovers, and we leverage this to provide sorting suggestions in the experiment described here.

6.3 Methods

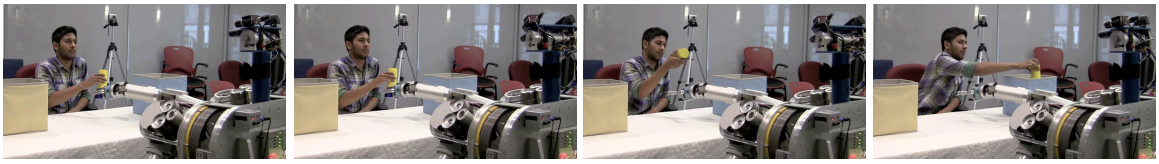
Users engage in a simple collaborative manipulation task with HERB: the robot hands blocks to participants, who are asked to sort these blocks into either a yellow box or a blue box on the table in front of them. Most blocks are unambiguously colored (fully yellow or blue), but some blocks are ambiguously colored (50% yellow and 50% blue) or semi-ambiguously colored (70% of one color and 30% of another), as illustrated in Figure 6.1(b).

The main manipulation in this experiment is the deliberate delay between HERB's sorting suggestion and block release (Figure 6.2). HERB provides a nonverbal sorting suggestion by looking (i.e., orienting its head) toward one of the boxes on the table. On "no delay" trials, HERB simultaneously releases the block to the participant and executes the suggestion behavior by looking at one of the boxes. On "delay" trials, HERB first looks at one of the boxes, and only then releases the block to the participant. The suggestion behavior takes about four seconds to execute: one second for HERB to turn its head toward the box, two seconds to gaze at the box, and one second to return its head to the starting point. Thus, there is a one second difference in when the block is released between no delay and delay conditions.

There are several strategies participants could employ for sorting the blocks. One of the more obvious strategies is to sort by dominant color, putting the primarily yellow blocks in the yellow box and the primarily blue blocks in the blue box. Another strategy is to sort by the first visible color (typically the top color), regardless of how the rest of the block is colored. Other strategies, like sorting randomly or alternating boxes, do not take color into account. There are several possible sorting strategies; our analyses do not rely on participants to follow any particular strategy.

In order to test whether people see and comply with a robot's suggestions, we made HERB's suggestions as counterintuitive as possible. Therefore, when handing over the ambiguous block, HERB always suggested the bottom (i.e., less visible) color when it presented the block, which conflicts with the top color strategy (Figure 6.3). Because the block was exactly half of each color, however, there was no conflict with the dominant color strategy, so the ambiguous case was only mildly counterintuitive. When handing over the semi-ambiguous block, HERB always presented the block with the dominant color on top and always suggested the less dominant color. HERB's suggestion conflicts with both the dominant color strategy *and* the top color strategy, making this a highly counterintuitive suggestion.

We also manipulated whether HERB engages in social or non-social gaze before the suggestion (Figure 6.4). In the joint attention condition, HERB first makes eye contact by looking at the participant's face, then down at the block in its hand, and then back to the participant's face as it reaches with the block to begin the handover. After HERB initiates the suggestion by turning its head to one of the boxes, it again returns to look at the participant's face before retracting its hand. In the mirrored condition, HERB's head moves at the same time and for the same distance as in the joint attention condition, but it moves laterally and remains oriented downward throughout this movement, so the gaze appears non-social. Therefore, we control for total amount of head movement while manipulating whether the gaze is social



(a) Delay

(b) No delay

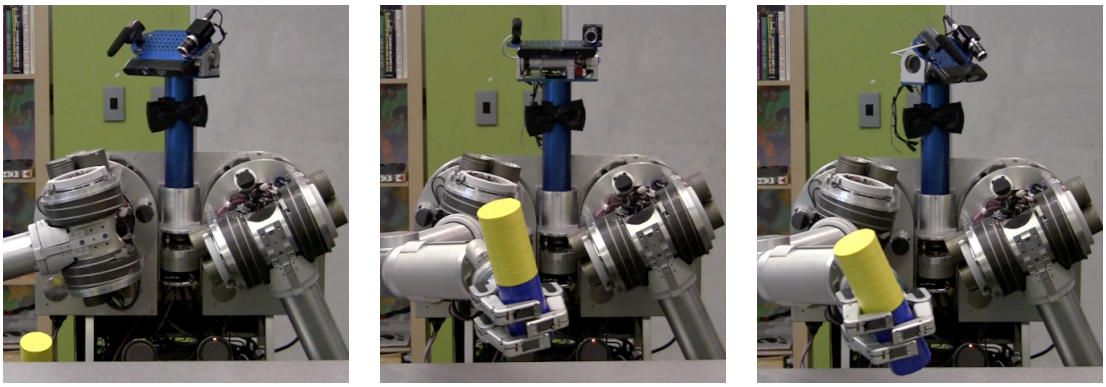
Figure 6.3: A comparison of HERB's head and hand movements during the handover transfer phase in delay and no delay conditions. In the no delay condition, HERB releases the block as it turns its head to a box. In the delay condition, the release occurs only after the head turn (frame 3).

or non-social. This manipulation explores whether social gaze before the suggestion affects how people respond to a counterintuitive suggestion.

6.3.1 Robot platform

HERB (Home Exploring Robot Butler) is a bi-manual robot developed for assistive tasks in home environments at the Personal Robotics lab at Carnegie Mellon University [227]. HERB has two 7-DOF WAM arms, each with a 4-DOF BH8-series Barrett hand with three fingers. In this experiment, only the right arm was used. HERB's hand has a 6-axis force/torque sensor that can detect external forces applied to the joints, for instance when a participant gently pulls on an object in HERB's hand. Motion trajectories for picking up and handing over blocks were pre-planned using CHOMP [198] and played back during the experiment.

HERB also has a pan-tilt head outfitted with a Microsoft Kinect and a camera, though no real-time vision was used in this experiment. The front of the Kinect has two visible round cameras which serve as HERB's "eyes."



(a) Neutral gaze (b) Joint attention (c) Mirrored gaze

Figure 6.4: HERB began every trial in a neutral gaze position, looking at the block in its hand. Herb then displayed either (b) joint attention or (c) mirrored gaze depending on the condition. Mirrored gaze and joint attention involve the same amount of movement at the same time, but the mirrored gaze target is shifted away from the participant's face so as to appear non-social.

6.3.2 Procedure

Participants were randomly assigned to either the delay or no delay condition, and to either the joint attention or mirrored condition. There were 32 participants (18 females), eight in each of the four conditions, with a mean age of 34. Participants were recruited from the Pittsburgh area using an online participant pool website through Carnegie Mellon University. They were compensated $10 for their time.

Participants were told that they would play a sorting game with HERB. They were instructed to take the block from HERB's hand once HERB had extended the block to them. Participants were also told that HERB's head would move and that HERB may provide suggestions about how to sort the blocks, but that the final sorting method was up to them. Participants were not informed about the kinds of blocks they would be seeing, and the blocks were kept hidden under the table until HERB handed them to the participant.

On each trial, HERB picked up a block from below the table and handed it to participants by extending its arm forward while grasping the block. Following previous research [55], HERB's arm became fully extended to clearly communicate

the handover.

HERB's hand contains a force sensor that identified when participants grasped the block during the handover. When the force sensor in the hand registered the participant's grasp on the block, HERB initiated a suggestion behavior by turning its head to one of the two boxes on the table. Depending on the delay condition, HERB either simultaneously released the block (no delay) or waited until its head was fully turned and then released the block (delay, Figure 6.2). Once the block was released, HERB withdrew its hand to begin the next trial.

Each participant engaged in five block handovers. The first two handovers involved solid color blocks, one of each color, both to familiarize participants with the task and to establish their preferred sorting method. The third handover involved the ambiguous block, presented with the first block's color on top. HERB always suggested that this block be sorted according to the color on the bottom, which violated the top-color sorting strategy, but was only a mildly counterintuitive suggestion because it did not violate the dominant color strategy (since there was no dominant color on the ambiguous block). The fourth block was again a solid colored block of the same color as the first block, intended to separate the test trials and to balance the number of blocks in each box as well as possible. For example, if the first block was yellow, the second block was blue, and the third block was ambiguous (presented with yellow on top), then the fourth block would be yellow again; the idea was that participants would sort the first and fourth blocks into the yellow box and the second and third blocks into the blue box, though this was not always the case. The final block was the semi-ambiguous block; this was always presented with the dominant color upward, to increase the saliency of the dominant color, but HERB always suggested sorting by the minority color. For example, if the fifth block was 70% blue, HERB oriented its head toward the yellow box.

By handing the ambiguous and semi-ambiguous blocks over with HERB's sug-

gested color on bottom, we made it as easy as possible for participants to use a sorting strategy that would conflict with HERB's suggestions in these cases. Therefore, we expect to see low compliance with HERB's suggestions in the absence of a manipulation.

First block color, dominant color of the semi-ambiguous block, and the arrangement of the boxes on the table were counterbalanced between participants. The experiment ran fully autonomously using pre-scripted trajectories for block pick-ups, block handovers, and head movements. There was real-time force feedback to measure when participants grasped the block during the transfer phase of the handover. The human-robot interactions lasted approximately 2 minutes and 20 seconds, though the particular amount of time varied by how long the participant took to sort the block.

6.3.3 Data collection

There are four data sources in this experiment. First, task performance was evaluated by whether participants followed HERB's suggestion on ambiguous and semi-ambiguous block trials. This provides a quantitative evaluation of compliance with HERB's counterintuitive suggestions.

Immediately after the interaction, participants completed written questionnaires that asked about their experiences and decisions during the task. These included free-response questions about whether they noticed suggestions from HERB and about their sorting strategy. Questionnaires contained specific questions about the ambiguous and semi-ambiguous blocks (represented with drawings), as well as Likert scale questions about HERB—rating features such as intelligence and friendliness—and about the collaboration—rating statements such as "I felt like HERB and I acted as a team."

After completing the survey, participants also engaged in a semi-structured interview with an experimenter. They were asked to explain their sorting of the ambiguous

and semi-ambiguous blocks. They were also asked whether they noticed any suggestion behavior from HERB. The responses from these interviews are used to support participants' written responses in the surveys.

Finally, the interaction with HERB was video recorded, and these videos were coded for information such as amount of time spent looking at HERB's face and whether or not the participant looked at HERB's head as it executed a suggestion behavior. The videos were annotated by an independent coder naïve to the research hypothesis. We randomly selected 10% of the videos for validation with a second coder; inter-coder agreement was 87% or higher. Because all of the coding measures were objective and inter-coder agreement was above the accepted 80% threshold, we feel confident analyzing the single coder's annotations.

6.4 Results

This experiment yielded quantitative results from the task (such as the rate of participants complying with HERB's suggestion), self-reports in the form of Likert scales and free-responses on the post-task questionnaire and semi-structured interview, and objective observations of the interaction from the recorded videos (such as the amount of time participants spent looking at HERB's head).

Manipulation Check. To verify that HERB's sorting suggestion for semi-ambiguous blocks was counterintuitive, we analyzed the rate at which people chose the "counterintuitive" box when they were unaware of HERB's suggestion. Recognizing HERB's head movements as sorting suggestions significantly correlates with sorting the semi-ambiguous block as suggested (Pearson's $\chi^2(1, N = 32) = 11.567, p = 0.007$). Only one participant out of 14 sorted the semi-ambiguous block as HERB suggested without recognizing HERB's suggestion, verifying participants' bias against HERB's sorting suggestion and supporting the semi-ambiguous block as a valid manipulation

to test for compliance.

Compliance: Semi-Ambiguous Block. The central research question is whether users comply with HERB's suggestion in the semi-ambiguous case. To test the effects of delay and gaze on correctness, we ran a factorial nominal logistic regression, which found that delay has a significant effect on compliance ($\chi^2(1, N = 32) = 6.77, p = 0.0092$). Without delay, only 19% of users sorted the semi-ambiguous block according to HERB's suggestion; delaying the release leads to 63% of users matching HERB's suggestion (Figure 6.5(a)).

When we analyze only participants who reported recognizing HERB's head movements as suggestions, the rate of compliance increases to 83% for participants in the delay condition and 33% for participants in the no delay condition (Figure 6.5(b)), with delay playing a significant role in this outcome: a nominal logistic regression for compliance with gaze and delay as factors, on only users who recognized HERB's head motions as suggestions, reveals a significant effect of delay ($\chi^2(1, N = 18) = 4.46, p = 0.0346$).

Compliance: Ambiguous Block. The ambiguous block represents a relatively low-conflict suggestion. Even though HERB always suggests sorting by bottom color, which violates the top-color strategy, most participants (59%) followed HERB's suggestion for sorting the ambiguous block.

An effect likelihood ratio test reveals a borderline significant effect for delay ($\chi^2(1, N = 32) = 3.632, p = 0.0567$), with 75% of participants in the delay condition following the ambiguous block suggestion, but only 56% of participants in the no delay condition following the suggestion (Figure 6.5(c)). There was no effect of gaze or an interaction.

Sorting ambiguous and semi-ambiguous blocks by HERB's suggestions are highly correlated (Pearson's $\chi^2(1, N = 32) = 9.85, p = 0.0017$). Ninety-two percent of users who sorted the semi-ambiguous block according to HERB's suggestion also previously

sorted the ambiguous block according to HERB's suggestion.

Gaze. Gaze type (joint attention versus mirrored) did not significantly affect compliance, and there was no significant interaction effect. Similarly, the analysis of compliance in only participants who reported recognizing HERB's suggestions found no significant effect of gaze. Joint attention does correspond to a higher probability of following HERB's suggestion (44% with joint attention, 38% with mirrored), but the difference is not statistically significant.

Self-Reports. Participants' free responses on the questionnaire and interview revealed that 75% of participants in the delay condition and 38% of participants in the no delay condition noticed and interpreted HERB's head movements as sorting suggestions (Figure 6.5(e)). A nominal logistic regression with gaze and delay as factors shows that delay significantly affects whether participants thought HERB had a sorting suggestion ($\chi^2(1, N = 32) = 4.69, p = 0.0302$, Wald test $\chi^2(1, 32) = 4.32, p = 0.0377$). No significant effect was found for gaze, and no interaction was found.

None of the participants in the no delay condition thought they complied with HERB's suggestion on the semi-ambiguous block trial; they either did not state that HERB gave them a suggestion, or they explicitly stated that they did not follow HERB's suggestion (Figure 6.5(f)). In the delay condition, 50% of users explicitly stated that they chose their sorting strategy based on HERB's suggestion for the semi-ambiguous block. The effect is significant according to the effect likelihood ratio test ($\chi^2(1, N = 32) = 13.8, p = 0.0002$).

Video Coding. Videos were coded for how long participants looked at HERB's head, which reveal how much visual attention participants devoted to HERB's head during the task. Videos were also coded for events in which participants looked at HERB's head while HERB looked at one of the boxes, which indicate whether attention was directed to HERB's head at the right time to notice HERB's gaze cues.

A two-factor analysis of variance investigating the effect of delay and gaze on the total amount of time the participant looked at HERB's head showed a significant effect of delay ($F(1, 31) = 12.9828, p = 0.0012$). The handover delay more than doubled the mean looking time, from 24.7 seconds to 50.6 seconds (Figure 6.5(d)). There was no significant effect of gaze or an interaction.

To understand whether this additional time spent looking at HERB's head was useful, we ran a nominal logistic regression analyzing the effect of delay and gaze on whether the participant noticed HERB's suggestion in each trial, as measured by whether the participant looked at HERB's head while it was oriented toward one of the boxes (Figure 6.5(g)). The test found a significant effect of delay on noticing HERB's suggestion in the third and fourth trials, and a borderline significant effect in the fifth trial, but no significant effect of delay in the first or second trials ($\chi^2(1, N = 32) = 9.113, p = 0.003$ for trial 3, $\chi^2(1, N = 32) = 6.974, p = 0.008$ for trial 4, and $\chi^2(1, N = 32) = 4.993, p = 0.0254$ for trial 5, lowering the α to 0.01 for this analysis based on a Bonferroni correction, the most conservative control for multiple comparisons).

Given that HERB first presents the delay in the third trial, these results show that a deliberate delay led people to attend more to HERB's suggestions, even in subsequent trials when no delay was present (the fourth trial). The analysis did not find any effect of gaze on any trials.

6.5 Discussion

Our results yield two main findings about the effects of deliberate handover delays.

Result 1 *Handover delays increased the amount of attention participants paid to the robot's head, which increased participants' awareness of the robot's nonverbal gaze cues.*

As predicted by hypothesis H1, deliberate delays increased the amount of time participants spent looking at the robot's head in general. This increase was not spurious: deliberate delays also increased time spent looking at the robot's head specifically when the robot made a suggestion. Therefore, a handover delay drew attention to the robot's head even though the head was not involved in the handover. These results are supported by both self-reports and video observations.

Result 2 *In addition to increasing recognition of the robot's suggestions, the handover delay also increased compliance with those suggestions.*

In our analysis of just the participants who reported recognizing the robot's head movements as suggestions, there was still a significant effect of delay on compliance. In other words, even once participants are aware that the robot is making a suggestion, they are still more likely to comply with that counterintuitive suggestion if the handover has a delay.

There are several interpretations for this second finding. The non-agentic explanation is that the delay drew peoples' attention to the robot's head which, because it was moving, increased the saliency of the suggestion behavior to the point where people followed it. This explanation does not require participants to attribute any kind of meaning to the handover delay. It is not wholly satisfactory, however, because when we exclude people who do not explicitly report seeing the robot's suggestion, there is *still* an effect of delay on compliance. Because these participants already notice the robot's head movements and explicitly interpret them as suggestions, the saliency of the robot's head movement seems unlikely to have a further effect.

A more agentic explanation is that the robot's handover delay was interpreted as *purposeful*, and that people were more likely to comply with the robot's suggestion when they believed that it came from an intentional agent. In order to test this explanation, we would need to measure peoples' attributions of agency to the robot before and after experiencing the delayed trials, an interesting point for future work.

In the current experiment, when the robot expressed a delay, there was always another meaningful channel of communication (eye gaze) to draw information from. However, it would be interesting to explore the effect of a delay when there is no other salient feature on which to focus attention. Perhaps in that case the delay would be interpreted as less agentic; the robot might even be seen as broken.

Mechanics of Deliberate Delays. We used a handover delay of one second in this study because that was the amount of time it took for HERB to turn its head toward a box. By the strength of the results, this duration was effective for drawing attention back to the robot's head.

A minority of participants in the delay case did not shift their attention in response to the delay; instead, these people seemed to focus more intently on pulling the block from the robot's hand. In these few cases, perhaps the delay was so unexpected that it served to draw attention to the hand, rather than release attention from it. More work is necessary to find the balance point where the delay is long enough to notice but not so long as to be problematic. This spot may also vary among people and depend on factors such as comfort with the robot.

Tasks for Investigating Compliance. The analysis supports our use of a semi-ambiguous block to investigate compliance. We found a strong correlation between noticing the robot's suggestion and sorting the block according to that suggestion. Of the 14 participants who did not report noticing the robot's suggestions, only one of them sorted the semi-ambiguous block in the same box that the robot suggested, emphasizing the counterintuitive nature of that suggestion. More users overall matched the robot's suggestion in the ambiguous case (59%) than in the semi-ambiguous case (41%). This is expected, as the robot's suggestion for the semi-ambiguous block conflicted with both the dominant color strategy and the top color strategy, whereas the ambiguous suggestion only conflicts with the top color strategy.

A block-sorting task is a useful proxy for other collaborative manipulation tasks

because it involves many of the same behaviors in a simplified format. Our task included handovers, object manipulation, classification decisions, joint attention, referential gaze, and mutual gaze. It subtly addressed the issue of compliance and touched upon animacy and intentionality of robot agents. The task was performed in a constrained environment with high repeatability and few distractions, but it remained easy to understand and natural to complete.

Gaze. Our gaze manipulations had little effect on attention and compliance in this handover task. Though people could understand HERB's gaze to the box as a suggestion, there was no difference between joint attention and mirrored gaze conditions in terms of how much attention was directed at HERB's head or the rate at which participants complied with HERB's counterintuitive suggestions. Thus, hypothesis H2 was not supported.

While studies have shown that robot gaze is a strong social cue, many of these studies used tasks in which gaze was a primary component, such as conversation. In these situations, a person attends to the robot's gaze as part of the task, and therefore gaze cues may be more salient or useful.

Furthermore, HERB's "face", as seen in Figure 6.1(a), is relatively abstract: the entire head consists of a flat platform with a pair of cameras for perception and a microphone. On the spectrum of biological realism in Section 2.2.1, HERB falls quite far to the left, with robots that have low levels of realistic gaze capabilities. It is possible that HERB's non-anthropomorphic head affected how well people actually perceived joint attention, and that the joint attention condition would have yielded different results on a robot with more defined eyes.

In the delay condition, joint attention increased the amount of time spent looking at HERB's head to 61 seconds from 41 seconds for mirrored gaze. Joint attention also doubled the probability of a participant complying with HERB's suggestion in the no delay condition (from 12.5% to 25%). However, neither of these effects reached

significance. More research is needed to understand how social gaze affects people in tasks where gaze is not a central component.

Future Work. Speech can be more precise and noticeable than gaze in many situations. Because this study focused on nonverbal communication (handover fluency and eye gaze), adding speech to the system would have been a confound with our current manipulations (handover delay and social gaze). However, future work on a robust human-robot handover system should incorporate spoken cues.

The decision to present the gaze cue during the handover transfer phase was based on a pilot and previous experience with HERB. Future studies can explore the effects of presenting the gaze cue at different points in the interaction.

Implications. The results reported in this chapter provide insight into the design of effective robot-to-human handovers. When information needs to be conveyed during handovers, we suggest that seamlessness should be secondary to communication. For instance, by manipulating the force profile of the handover so that the robot deliberately delays releasing an object, robot designers can draw attention to important features like eye gaze and other nonverbal communication. This idea is in line with previous research that has shown that deliberately manipulating other aspects of a handover, like the spatio-temporal motion, can help convey information about the task [55]. The current work is novel because it uses a feature of the handover (the force profile) to convey information toward an unrelated mode of communication (eye gaze) about a subsequent task (block sorting).

6.6 Summary

This study investigated the nonverbal communication behind robot-to-human handovers. It showed that using one type of nonverbal communication—a gesture, in the form of a handover delay—could influence attention to, and compliance with, a

second type of nonverbal communication—a directional eye gaze.

In Chapter 3, we showed that directional eye gaze from a robot is not cognitively processed the same as directional gaze from a human, at least when measured on a micro scale. Interestingly, the current study shows a macro scale difference between how human and robot gaze is processed. Psychology research suggests that the joint attention condition used in this study should generate more engagement and solicit more compliance from a participant than the similar but non-social mirrored gaze condition. However, we failed to find any effect of gaze type on human behavior—people did not comply with the robot any more when it established joint attention than when it simply moved its head to a non-social gaze position.

On the other hand, the study in this chapter showed that a *combination* of nonverbal behaviors *can* influence human responses. In this case, people complied with the robot's sorting suggestions as conveyed by gaze, but only when they saw a deliberate gesture in the form of a handover delay. That is, the human partner must recognize that the robot's gaze behavior is intentional before they can interpret it as communicative.

This chapter focused on a specific part of human-robot physical collaboration, namely the handover. In the next chapter, we explore how people use a broad array of nonverbal communication during interpersonal collaborations, to try to identify how and when they use gaze and gesture in naturalistic interactions.

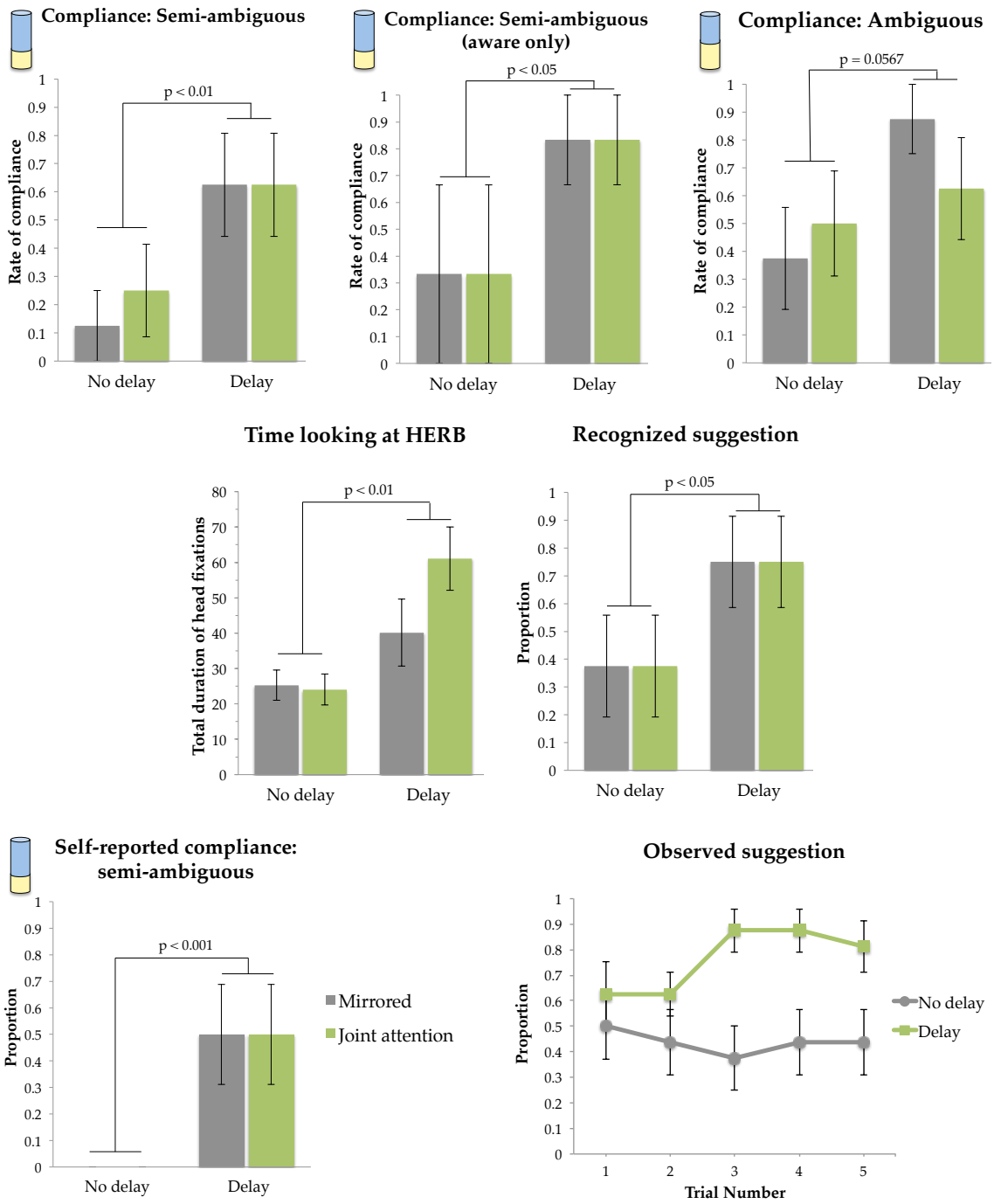


Figure 6.5: Results of the experiment from task responses, self reports, and video observations. Error bars indicate ±2 SE.

7

Data-Driven Model of Human Nonverbal Behavior*

Up to this point, we focused on how people respond to specific elements of robot nonverbal behavior (such as attentional eye gaze and handover gestures) in well-controlled laboratory studies. In this chapter, we begin modeling these nonverbal behaviors. We start by focusing on human-human interactions, which provide a rich source of examples for communicative nonverbal behavior. This chapter focuses specifically on a tutoring scenario, where a teacher conveys a set of information to a student. We chose a spatially-oriented tutoring scenario (teaching a map-based board game) to evoke a range of nonverbal behavior, including directional eye gaze and gestures. We recorded student-teacher pairs and built a model of nonverbal behavior from their interactions. This model is bidirectional: it can both predict *the context of a communicative nonverbal behavior and successfully* generate *nonverbal behaviors to match a particular context. This bidirectionality enables a robot*

*This work was originally published as:
Henny Admoni and Brian Scassellati. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proceedings of the 6th ACM International Conference on Multimodal Interaction (ICMI)*, Istanbul, Turkey, 2014.

to understand and use nonverbal behaviors from the same underlying data representation.

7.1 Introduction

Socially assistive robotics (SAR) focuses on building robots that help people through interactions that are inherently social [83]. Application areas for SAR include tutoring [130, 142], autism therapy [216], and elder care [258]. Social robots augment traditional human-human interactions in these areas by providing additional interactions that are impractical, time-consuming, or impossible to achieve with a person.

For example, a social robot can act as a peer tutor, helping students practice skills or solidify knowledge through one-on-one interactions outside of the classroom. By presenting itself as a peer, the robot can encourage students to practice previously-learned knowledge by re-teaching it to the robot. In this way, the robot provides educational support beyond what a classroom teacher has time for, and with potentially more consistent quality than a human peer.

For social robots to be effective communicators, they must understand the *context* of their human partner's communication, that is, the communicative goal or perspective. In the tutoring robot example, for instance, the robot must be able to recognize whether its partner is referring to a location in the environment, asking a question, or explaining some knowledge. Similarly, social robots must be able to convey the context of their own communication effectively.

The cues to understanding such context can come from speech, but often come from nonverbal behaviors like eye gaze [28] and gesture [164]. Gestures, for instance, reflect ideas that are not necessarily conveyed in speech [99], and teachers frequently use gestures to ground their spoken utterances to the objects of instruction [17, 204]. Eye gaze is critical for joint attention—simultaneous attention toward a particular

Figure 7.1: Screenshots from human-human teaching interaction videos. The student (top) displays gaze to the referent, while the teacher (bottom) displays gaze to the partner and a deictic gesture to the map.

object or location—which is fundamental for learning [247]. Therefore, the effectiveness of the tutoring robot, or any socially assistive robot, depends on its ability to recognize and utilize the nonverbal context clues that people use naturally.

In this work, we take a *data-driven* approach, using empirical data from human-human interactions to build a model of nonverbal robot behaviors. By training on previously-observed human behavior, we take advantage of the frequency and ease with which people use nonverbal behaviors to design more communicative robot behaviors.

Other work uses a similar data-driven approach for nonverbal behavior modeling. Researchers have generated robot behavior, such as gaze aversions [24] and narrative gestures [116], by analyzing videos of people conversing or telling stories. For virtual agents, empirical observation has driven gesture formation for iconic gestures [40] and

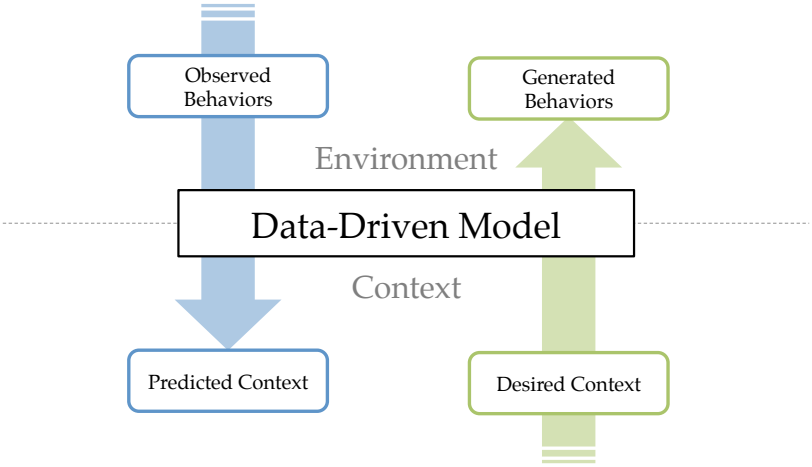


Figure 7.2: The model performs both context prediction (left) and behavior generation (right).

narrative performance [178, 233].

However, much of this previous work focuses exclusively on the speaker's behaviors. In contrast, our work considers the behaviors of *both* interaction partners simultaneously. Tutoring is an activity with bi-directional communication—the teacher makes a statement, the student asks a question, the teacher replies, the student confirms—and peoples' nonverbal behavior is influenced by the behavior of their partner. For instance, joint reference is a common social behavior that involves one person deictically referring to an object or location, then another person looking at that referent in response. With a view of both partners' behaviors, our model captures this kind of bi-directional behavior.

Our work also hinges on the idea that a model for nonverbal behavior should be simultaneously *predictive* and *generative*. In other words, the model should be able to both predict (or classify) the context of a newly observed set of nonverbal behaviors, and generate a set of nonverbal behaviors given a context of communication (Figure 7.2), without needing to collect and train on different sets of data. Some other work has this capability (such as [40]), but we elevate this to a central requirement for our system.

In this chapter, we introduce the context and features that comprise our model and describe our preliminary data collection of real-world human-human teaching interactions. We then describe our model in terms of these features, detailing how it can both predict new contexts and generate new behaviors. We evaluate the model and show that it is effective at both of these tasks. We conclude with ideas for extensions of this model.

7.2 Collecting Human-Human Interaction Data

To create a model of nonverbal behavior, we first collected examples of nonverbal behavior during tutoring (Figure 7.1). We recruited two pairs of participants (mean age 22), randomly assigning one as teacher and the other as student, and recorded their interaction as the teacher taught the student how to play a board game called TransAmerica.

In TransAmerica, players must place game pieces representing railroad tracks along a grid overlaid on a map of the United States. Players score points for successfully building a track network that connects the cities specified in their randomly-selected hand of cards. We chose this game specifically because teaching the game involves spatial references, which encourage deictic gestures and demonstrations in addition to statements of facts and rules.

Neither student nor teacher had played the board game previously. Before the recorded interaction, the teacher was given a lesson on the game from an experimenter for approximately five minutes. The teacher was also provided with a rule sheet that described all of the rules of the game.

We audio- and video- recorded both teacher and student during this interaction, which lasted approximately five minutes per dyad. We then manually coded these recordings for five *predictors*: the *teacher's gaze, gestures*, and *deictic references*, as

well as the *student's gaze* and *gestures* (Table 7.1). The student's deictic reference was infrequent, so we did not code for that predictor. We also coded the *context* of each utterance.

Values for gaze follow previous work [116], and represent possible gaze locations: to the *partner*, to the *referent* of current speech (regardless of who is speaking), to one's *own gesture*, or to some *other* location in the environment.

Values for gesture include those from established categorizations as well as additional values specific to physically-based teaching tasks. *Iconic*, *metaphoric*, *deictic*, and *beat* gestures are defined as in the literature [164]. Iconic gestures are closely related to the topic of speech, and often represent physical concepts, such as meshing the fingers together when referring to a "track network." Metaphoric gestures indicate abstract concepts, for instance waving the hand between two players to represent "taking turns." Deictic gestures involve pointing, which can be accomplished with the finger or the whole hand. Beat gestures are linked to transitions in speech, but are often unrelated to the content of speech. *Demonstrations* involve physical movements that mimic the topic of speech. *Functional* movements are not intended for communication, but are used to accomplish game-related tasks such as dealing cards. Actions outside of these categories, such as brushing hair behind an ear, were categorized as *other*.

The deixis category encodes gesture types—*pointing* to a single target, *sweeping* over a range of targets, and *holding*—as well as gesture locations—the game *map*, *cards*, *playing pieces*, and *box*. Though every deixis value must have an associated gesture, not every gesture must have a deixis value. Deixis values can occur with any gesture, especially demonstrations and functional gestures.

The nine contexts each represent a particular kind of communication, and contexts are determined based on both speech and nonverbal behaviors. The *rules* context indicates communication about the rules of the game. *Fact* contexts involve commu-

Name	Values
Gaze (A)	partner, referent, own gesture, other
Gesture (E)	iconic, metaphoric, deictic, beat, demonstration, functional, other
Deixis (D)	point {map, own cards, partner cards, box}, sweep {map, box}, hold {cards, game piece, box}
Context (C)	backchannel, deixis, expository, fact, filler, question, reply, rules

Table 7.1: Model parameters and their values.

nication about facts that don't include game rules, such as "The name of the game is TransAmerica." An *expository* context indicates communication that elaborates on previous statements without providing new rules or facts. *Question* and *reply* contexts involve asking questions or providing direct answers, respectively. *Deixis* indicates communication that explicitly refers to physical locations or nearby objects. *Confirmation* contexts involve confirmation-seeking questions or statements such as "do you understand?" *Backchannel* contexts are utterances that indicate a listener's attention. *Filler* are non-meaningful communications that stand in for silence, often at the beginning of a new phrase.

Contexts are mutually exclusive, though two sets of identical nonverbal behaviors may be classified as different contexts, for instance based on different speech during those behaviors. For example, the deixis context always involved concrete object references, while the other contexts involved more abstract descriptions of the game.

We developed the list of contexts before examining the human-human interaction data, so that we would not be influenced by individual preferences for certain contexts. Interestingly, we did not note a single instance of confirmation context in the interactions we annotated, despite their expected appearance in a teaching task. It is possible that a more experienced teacher might employ confirmation seeking behaviors, even though our current participants did not.

7.3 Nonverbal Behavior Model

A model of nonverbal behavior should be able to classify the context given new observations of nonverbal behavior, as well as generate appropriate behaviors to suit a desired context (Figure 7.2).

We discretized the human-human interaction recordings into one-second segments. Each segment provides one observation $o \in O$, which is described by a tuple of predictors $o = \{a_T, e_T, d_T, a_S, e_S\}$ where $a_T, a_S \in A$ are the type of eye gaze exhibited by the teacher and student, respectively, $e_T, e_S \in E$ are the types of gestures exhibited by the teacher and student, respectively, and $d_T \in D$ is the deictic referent of the teacher's gesture in that segment. We chose one-second segments after observing the interactions, though the level of data granularity is flexible and may be adjusted for different applications.

Sometimes it is useful to take history into account, as well. An observation with history,

$$o_h = \{a_{T_t}, e_{T_t}, d_{T_t}, a_{S_t}, e_{S_t}, a_{T_{t-1}}, e_{T_{t-1}}, d_{T_{t-1}}, a_{S_{t-1}}, e_{S_{t-1}}\}$$

is defined by predictor values at current time t and predictor values from the previous time step $t-1$, if available. The set of observations including history is O_h.

Using this formulation, we can represent each observation as a point in high-dimensional space.

7.3.1 Predicting Context

Given a set of observations of nonverbal behavior, our system can predict the context of the communication. To do so, observations from the human-human interactions were used to train a prediction algorithm using k-nearest neighbors. In this algorithm, predictors are attributes and context is the class label. We can denote this as $label\,(o_h) = c$ for observation o_h and context c. Note that *label* is not a function,

since identical observations can have different contexts.

To classify the context of a new observation, the algorithm performs operation

$$nclosest : (O_h, o_{new}, k) \rightarrow K \tag{7.1}$$

which takes a set of observations O_h, a new observation o_{new}, and a positive integer k and returns a set $K = \{o_{h_1}, \ldots, o_{h_k}\}$ containing the k closest observations to o_{new}.

Because the predictor values are categorical, rather than continuous, our KNN algorithm uses the Hamming distance to identify nearest neighbors. For each existing observation $o_{old} = \{x_1, \ldots, x_n\}$, the algorithm calculates the distance D between o_{old} and the new observation $o_{new} = \{y_1, \ldots, y_n\}$,

$$D = \sum_{i=1}^{n} h(x_i, y_i), \quad h(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases} \tag{7.2}$$

Once it has evaluated the k nearest neighbors, the model assigns o_{new} a context based on a majority vote of the contexts of the observations in K. Ties are resolved by selecting randomly. Since there may be several different behaviors applicable in the same context, we extend the context assignment to o_{new} such that the probability of context assignment is proportional to the number of observations with that context in K. In other words, the probability of assigning o_{new} a context c is $p(c) = \frac{count(label(o_h)=c)}{k}$ for each $o_h \in K$, where $count(x)$ is a function that returns the number of instances of x in the data.

We empirically determined that $k = 2$ was the most accurate value for our data, though k may vary by application. Our model examines the two most similar examples of previous behavior to judge a new behavior's context.

7.3.2 Generating Behavior

Given a context, the model can also select appropriate nonverbal behaviors. It does so by finding the largest cluster of examples for the context, then selecting the nonverbal behaviors that are most common in that cluster.

Mathematically, given a desired context $c_{des} \in C$, the model searches over all observations $o \in O$ for

$$\{\{a_T, e_T, d_T, a_S, e_S\} \mid \begin{array}{l} count\left(label(o) = c_{des}\right) > \\ count\left(label(o) = c_i\right), c_{des} \neq c_i\} \end{array} \tag{7.3}$$

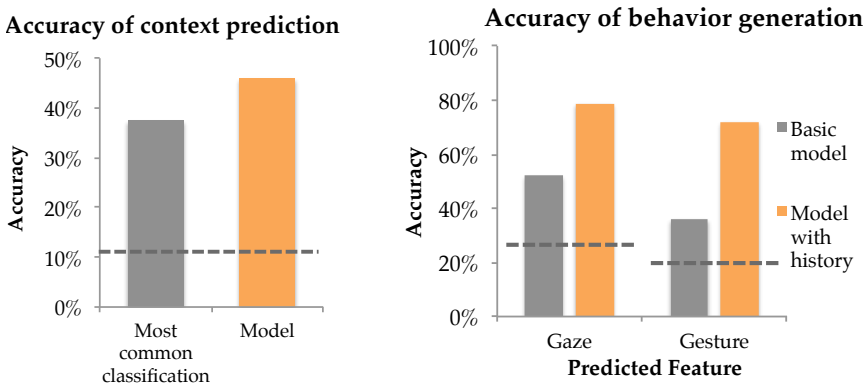
Since this can yield multiple qualifying sets of behavior, the model can weight its behavior choice based on the frequency of observations containing that behavior for the desired context. This allows behavior variability in proportion to observed examples.

In effect, the model is replicating the most common behaviors it has observed for a given context. This follows the idea that people learn to communicate by mimicking observed behavior in given situations.

This behavior generation algorithm is amnesic because it does not account for history. To account for behavior from the previous time step, we use the history-aware representation of an observation, o_h. The process for generating new context (equation 7.3) remains the same, except that every step now uses o_h instead of o.

7.4 Model Evaluation

Given a new observation containing the five predictors, how accurate is the model at identifying the correct context? We performed a 10-fold cross validation: combining all observations from both dyads, this validation segmented the data into 10 groups, trained the model on nine of those groups, and calculated the accuracy of context predictions using data from the remaining, untrained group (Figure 7.3(a)). On



(a) Accuracy of predicting context given behavior observations.

(b) Accuracy of generating new behaviors given context.

Figure 7.3: Results of model evaluation. The dashed line indicates chance.

	backchannel	deixis	expository	fact	filler	question	reply	rules
backchannel	10	13	23	1	11	2	2	21
deixis	4	14	16	0	6	4	0	15
expository	6	13	59	3	29	8	3	28
fact	1	0	4	0	1	2	0	1
filler	1	13	36	2	32	1	3	22
question	5	6	7	2	8	26	0	5
reply	1	2	4	0	1	1	2	4
rules	7	19	47	3	25	6	3	72

Table 7.2: A confusion matrix for context prediction with the cross-validated model. Rows represent ground truth and columns represent predicted classifications.

average, cross-validation accuracy was 45.9%. This value is significantly better than chance, which is 11.1% for nine classifications. It is also better than simply predicting the most common classification—rules—which only leads to an accuracy of 37.5% using the cross-validated model. Table 7.2 shows a confusion matrix for the cross-validated model.

Given a context, how well does the model generate gaze and gesture behaviors? To test this, we compared the recorded human behavior for each observation in our data set against the most likely behavior generated by the algorithm for that obser-

vation's context (Figure 7.3(b)). When using the amnesiac generation method (that is, behavior generation that ignores any history), our system matches actual human gaze behavior 52.0% of the time, and human gesture behavior 36.0% of the time. This is an improvement over randomly selecting behavior values, which would yield 25.0% accuracy for gaze and 14.3% accuracy for gesture. Taking a single time-step of history into account significantly improves performance. The historically mindful generation method yields 78.8% accuracy for gaze behaviors and 72.0% accuracy for gestures.

7.5 Discussion

In this chapter, we describe a model for communicative context. The model is trained on data from human-human tutoring interactions. It can successfully predict a speaker's communicative context from only their nonverbal behavior, and it can generate appropriate nonverbal behaviors to convey a particular context.

The model currently uses a subset of the nonverbal behaviors that people use to communicate. We carefully selected the gaze and gesture categories from existing work on nonverbal communication, and added a few categories (demonstrations and functional gestures) that were specific to our scenario. Extending the model to include other categories, or other behavior features such as head pose or body posture, might yield even greater accuracy and expressiveness.

As a data-driven model, the effectiveness of the system depends on the quality of the data provided. This chapter uses a relatively small data corpus (two dyadic interactions). While many nonverbal behaviors are consistent across people, it is not clear whether this model can account for the large variability in human nonverbal behavior expression, given our small sample size. Even with this small corpus, however, the model successfully predicted and generated nonverbal behavior in a cross validation

test.

The small sample size is largely due to the fact that data annotation presents a major challenge for data-driven modeling. Collecting human interaction examples and manually annotating them takes time. Automatic annotations are not yet robust enough to correctly identify all of the features used by the model, particularly the context. In future work, automatic gaze and gesture detectors may ease some of the burden of manual annotation. However, in Chapter 8 we take a different approach that does not require this manual coding process.

Additionally, the model developed here is specific to this kind of spatial tutoring scenario. It is not clear, for example, that the nonverbal behavior combinations displayed in this situation would map to the same contexts in a collaborative building task where people's hands are occupied. This scenario dependence is a general limitation of data-driven models, and it is also addressed by the heuristic model in Chapter 8.

7.6 Summary

We built a data-driven model of nonverbal behavior for tutoring. Data was collected from human teacher-student pairs interacting around a spatially-oriented board game, which elicited a range of gaze and gesture behaviors. We trained a simple nonverbal behavior model using five predictors—the teacher's gaze, the teacher's gestures, the teacher's deictic references, the student's gaze, and the student's gestures—along with a context label for each communicative utterance. This model recognized the context of a communication from just the set of nonverbal behaviors present during that communication. It also successfully predicted what nonverbal behavior should accompany a desired context.

This work helps develop social robots that can both understand and use nonverbal

behavior in their interactions with people. A tutoring robot can apply the empirical model described here to interpret a human partner's communication. It can also dynamically generate an appropriate nonverbal behavior to support or augment its next utterance, based on the context of that utterance.

However, such data-driven models as the one described in this chapter rely on collection and meticulous annotation of training data, which can be time consuming and scenario-dependent. In the next chapter, we introduce a scenario- and robot-agnostic model for generating nonverbal behavior that relies on features of the scene, rather than previous human examples, to select a robot's nonverbal behavior.

8

A Generative Model of Robot Nonverbal Behavior*

As we saw in the previous chapter, nonverbal communication can help predict the context of communication in a collaborative interaction. While Chapter 7 explored a broad array of communicative nonverbal behaviors, the current chapter focuses again on a specific type of communication covered in Chapters 5 and 6: referential communication. These earlier chapters investigated human responses to a robot's nonverbal referential communication. In the current chapter, we investigate how to dynamically generate such communication. We develop a heuristic model for generating a robot's nonverbal behavior that is scene- and robot-agnostic, which addresses some of the limitations of the data-driven model in Chapter 7. The current model uses a robot's verbal and nonverbal behaviors to successfully communicate object references to a human partner. This model, which is informed by computer vision, human-robot interaction, and cognitive psychology, simulates

*This work was originally published as:
Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

how low-level and high-level features of the scene might draw a user's attention. It then selects the most appropriate robot behavior that maximizes the likelihood that a user will understand the correct object reference while minimizing the cost of the behavior. This chapter presents a general computational framework for this model, then describes a specific implementation in a human-robot collaboration. Finally, we analyze the model's performance in two human evaluations—one video-based (75 participants) and one in person (20 participants)—and demonstrate that the system predicts the correct behaviors to perform successful object references.

8.1 Introduction

A key challenge for social robots is natural communication between robots and humans. Typical human communication occurs both verbally, through speech, and nonverbally, through modalities like eye gaze [27] and gestures [164]. Nonverbal behaviors serve to augment verbal communication by reinforcing or extending spoken communication [99]. Collaborative robot systems benefit from this multimodal communication to improve the fluency and efficiency of human-robot teams.

For example, imagine a robot that provides guidance to a user during a collaborative task, such as the furniture assembly task pictured in Figure 8.1. The robot has knowledge about the task steps, and may assist the user by identifying which part is required next. In this situation, a verbal description of the part, such as "the red hammer," can often be sufficient. However, there may be situations, such as the one pictured, in which the robot must disambiguate between similar objects. In this case, nonverbal behaviors like gaze and gesture become important. Users viewing the scene in Figure 8.1 may not understand which object is being referenced by the phrase

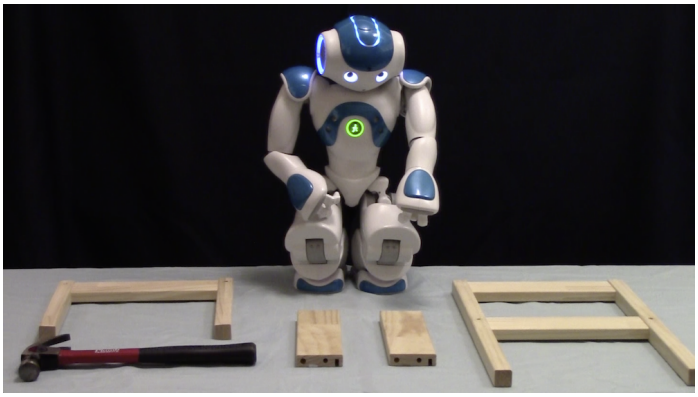


Figure 8.1: A behavior model must select enough nonverbal behaviors to disambiguate object references, while minimizing the cost of these behaviors.

"the small wood block," but the robot's head orientation and pointing, performed in conjunction with that phrase, make the reference easy to understand.

In this chapter, we focus on the task of *communicating object references* about objects of interest in the environment. As people perform tasks alone, their visual attention is directed almost exclusively at the objects required for that task, and almost never at other parts of the scene [106]. By directing a user's visual attention to relevant objects through effective object references, a robot can help focus the user's energy on the necessary components of their task. Research has shown that when a robot uses deictic nonverbal behaviors (e.g., gazes and gestures) to augment its referential speech, users respond to those references more quickly than when those nonverbal behaviors are absent [3, 44, 47, 117, 231].

A naive robot controller might augment its referential speech by selecting every relevant nonverbal behavior available to it, with the thought that this would maximize communication. However, frequent pointing and gaze behaviors use effectors that the robot might need for other parts of its task, like its arms for object manipulation and its head for vision. For robots operating on batteries, energy cost may be a barrier to frequent nonverbal behavior. Finally, people collaborating in teams often use nonverbal behavior for subtle, implicit communication [219], so overuse of

pointing and gaze may become visually distracting to the human partner. Therefore, a robot behavior system should select the fewest possible nonverbal behaviors that will maximize communication while maintaining interaction seamlessness.

To correctly select nonverbal behaviors for communication, the robot must have a model for how these nonverbal behaviors influence the interaction. Other research has modeled nonverbal behavior to communicate information in a variety of domains. Researchers have created models for generating expressive nonverbal behaviors for narration [114], conversational turn-taking [24], object manipulation [78].

Our research is novel because it considers nonverbal behaviors as elements in the environment that influence a user's attention, along with other parts of the greater visual scene comprised of both bottom-up and top-down features. This chapter presents a model for generating referential robot behaviors that considers the expected attention draw of the scene when predicting which behavior will be most effective. The model first simulates how different elements of a visual scene might capture a user's attention, then selects behaviors that most efficiently direct that attention to a target object.

Psychologists have established that people's attention is influenced by both bottom-up saliency cues and top-down context cues [126]. Bottom-up cues are highly salient visual features that are distinguished very quickly by the visual system, such as color and orientation. The effect of these bottom-up cues is modulated by top-down processes that depend on the context of the task [106]. Nonverbal behaviors such as pointing and gaze are top-down communicative factors that influence where people will attend.

Our model uses both bottom-up and top-down cues when calculating how the robot's nonverbal behaviors will influence a viewer's understanding of object references. We introduce a metric called the *referential likelihood score*, a score for each object in the scene that, given a verbal or nonverbal referential behavior, indicates

how much the model expects the user to see that object as the target of the reference. The referential likelihood score is calculated using features of the scene, the context of the interaction, and the robot's behaviors.

To select a communicative behavior, this model simulates referential likelihood scores for all objects under each of its possible verbal and nonverbal behaviors. It predicts the effect of each behavior on the user's attention, and then selects the behavior that maximizes attention toward the desired object while minimizing unnecessary actions.

This chapter makes several contributions:

1. A mathematically defined model for generating object reference behaviors (Section 8.3.1),

2. An implementation of the model on a Nao robot in a collaborative building task (Sections 8.3.2 and 8.3.3)

3. Two evaluations of the implementation—one video-based (Section 8.4.3), and one in person (Section 8.4.4)—that show how the model successfully minimizes extraneous actions while maximizing user comprehension of object references

8.2 Related Work

Modeling visual attention is not new to robotics [91], but most visual attention systems compute where a robot should look from the robot's perspective, rather than modeling the *user's* attention to determine the best robot behavior. For example, a gaze behavior controller for the social robot Kismet depends on both bottom-up saliency (detecting color, motion, and faces) and top-down motivations of the robot [49]. A bottom-up attention system controls an iCub robot's gaze based on robot-centric visual and auditory saliency maps [205]. A Keepon robot can develop joint

attention behaviors by learning the sensorimotor coordination that occurs when it looks at salient objects in its view [177].

The kinds of visual cues discussed in this chapter—saliency, pointing, and gaze—have been previously incorporated into attention models to provide information about where a social robot should attend. For example, to improve the speed of visual search, one system uses human pointing as an additional cue along with low-level saliency maps [217]. Another system incorporates haptic-ostensive references—which occur when an object is being referenced as it is manipulated—into the dialogue system for a cooperative human-robot assembly task [86]. Gesture recognition performance in a human-robot interaction improves markedly when head orientation is taken into account [181].

Our model is inspired by psychology's understanding of how people direct visual attention. Visual attention involves both bottom-up mechanisms to isolate objects of interest from their background, and top-down mechanism to select task-relevant objects [73]. According to the feature-integration theory of visual attention, visual processing proceeds in parallel when objects can be detected based on a single feature (such as color or shape), but serially when more than one feature is needed to distinguish between objects [249]. This idea provided the basis for a foundational computational model of visual attention designed by Itti and colleagues [125].

Gaze and pointing both have communicative value during interactions, specifically in drawing attention to a particular visual region [32]. However, directing attention through gaze or pointing refers to approximate spatial zones rather than precise linear vectors [54]. Therefore, these nonverbal behaviors can be seen as directing a cone of attention out toward the scene, rather than a single line.

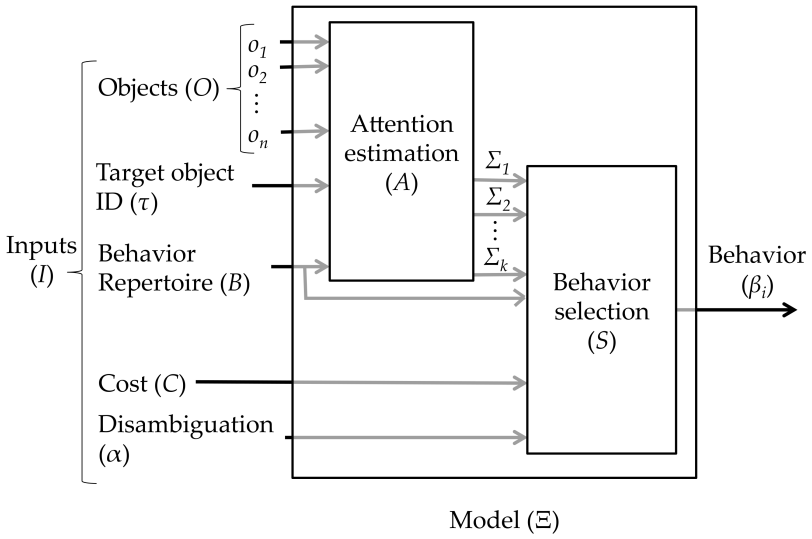


Figure 8.2: A schematic of the model, which takes features of the scene, a target object, and a robot's capabilities as inputs and outputs the best behavior for referencing the target object.

8.3 Behavior Model

This section presents a mathematical specification of the behavior model for object references (Figure 8.2).

8.3.1 Model Overview

The model takes a set of inputs I that describe the scene from the user's perspective, as well as the robot's target object and its capabilities in terms of behaviors. The model outputs a single behavior β that maximizes the likelihood of a user understanding that object reference while minimizing extraneous actions.

Mathematically,

$$\Xi : I \rightarrow \beta \tag{8.1}$$

where input I is a tuple

$$I = (O, \tau, B, C, \alpha) \tag{8.2}$$

The set of objects $O = \{o_1, \ldots, o_n\}$, denotes all the possible objects in the environment that might be referenced. One of these objects, o_τ, is the target object. In our example from Section 8.1 (and Figure 8.1), the target object is the small wood block on the right, and the set of objects O is comprised of that block, three other wood pieces, and a hammer.

The set $B = \{\beta_1, \ldots, \beta_k\}$ is the robot's repertoire, the behaviors it can perform. For example, the robot in Figure 8.1 can point to a location, gaze to a location by turning its head, or speak an utterance. A behavior may even be a combination of actions, such as a simultaneous head turn and point; B contains all possible combinations of behaviors as well. The repertoire of behaviors is dependent on the robot's capabilities, but includes both verbal and nonverbal behaviors.

C is a ranking of behaviors according to their relative costs, as described below. The value α indicates the disambiguation level for a reference. This specifies how clearly (or ambiguously) the referential behavior should indicate the target object, compared to other objects in the scene. Because it is a ratio measure, α has no units and exists on an arbitrary scale that may be situation dependent.

In actuality, the model is composed of two functions, A and S, in sequence

$$\Xi : I \xrightarrow{A} (\Sigma, C, \alpha) \xrightarrow{S} \beta \tag{8.3}$$

To select a successful referential behavior, we must compare how each behavior affects a user's interpretation of the scene. To do so, we calculate a *referential likelihood score* for each object under each behavior. We denote an object o_i's likelihood score under behavior β_j as $\sigma_{i,j}$. The set of likelihood scores for all objects under a behavior β_j is Σ_j.

The attention estimation function A calculates these referential likelihood scores

for each object. This function,

$$A : (O, \tau, B) \rightarrow \Sigma \tag{8.4}$$

takes the set of all objects and the index of the target object, along with the robot's behavioral repertoire. For each behavior $\beta_j \in B$, A calculates a set $\Sigma_j = \{\sigma_{1,j}, \ldots, \sigma_{n,j}\}$ of referential likelihood scores for each object under that behavior; mathematically, $A(O, \tau, \beta_j) = \Sigma_j$.

For example, if behavior 1 is "pointing," the values in Σ_1 would indicate the model's evaluation of likelihood scores for each object in a scene where the robot is pointing to the target object. $\Sigma = \{\Sigma_1, \ldots, \Sigma_{|B|}\}$ denotes the set of referential likelihood scores for each object under each behavior.

Depending on how function A is implemented, likelihood scores can depend on features of the objects, such as visual saliency; features of the scene, such as density of objects; and features of the robot's currently active behavior, such as pointing, gazing, or verbal reference. Our implementation of attention estimation function A is detailed in Section 8.3.2.

Once the set of likelihood scores under each behavior is calculated, it is given to a behavior selection function S along with a cost function and the desired disambiguation level. This function

$$S : (\Sigma, C, \alpha) \rightarrow \beta \tag{8.5}$$

selects a behavior to perform that minimizes cost while maximizing the likelihood of reference toward the target object. Parameter α provides a measure of the minimum level likelihood for the object reference.

The cost function C takes in behaviors from the robot's repertoire and returns their relative costs. It depends on the robot and the task at hand. For example, the cost of moving an arm to point might be greater than the cost of turning a head

to gaze. Costs can be measured in terms of energy expenditure, cognitive load for the user, or any other metric that should be minimized for easy, efficient human-robot collaboration. The ordering generated by C is a relative ranking of the robot's behaviors indicating their relative expense. Our implementation of behavior selection function B is detailed in Section 8.3.3.

Detailed specification of each function is intentionally flexible, as it depends on task, objects, and robot capabilities. In the next sections, we show an implementation of this model for a collaborative building task with a Nao robot.

8.3.2 Attention Estimation (A)

Our approach accounts for both bottom-up and top-down information when determining an object's likelihood score. It considers the *saliency* of a scene in terms of low-level features like color, intensity, and orientation. It also recognizes top-down *verbal context* by comparing descriptive words about the object of interest from a natural language utterance to a set of known object features. Finally, it considers *gaze* and *pointing* gestures to provide additional clarity about the object reference.

We combine these four features into a weighted linear sum

$$\sigma_{i_j} = \omega_s S_i + \omega_v V_i + \omega_g G_{i,j} + \omega_p P_{i,j} \tag{8.6}$$

The likelihood score $\sigma_{i,j}$ of an object $o_i \in O$ under behavior $\beta_j \in B$ depends on the object's low-level visual saliency in the scene S_i, the high-level verbal context V_i, the current gaze $G_{i,j}$ and the current pointing $P_{i,j}$. Each feature has a weight ω that indicates its relative importance to the likelihood evaluation.

In this work, we chose to implement the feature calculation as a linear weighted sum for simplicity. However, the feature calculation need not be a linear sum. It could be a more complex function to represent complex dynamics of different features.

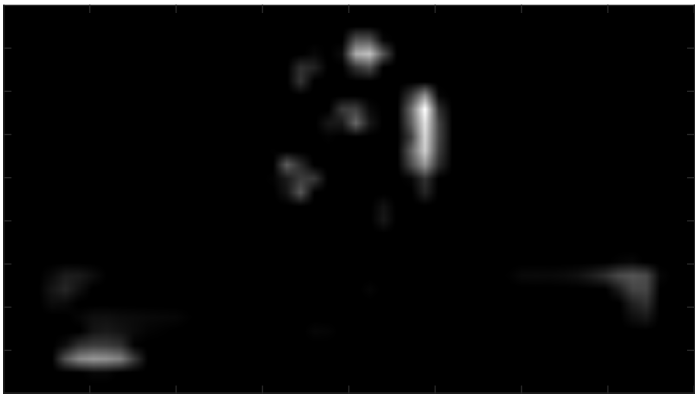


Figure 8.3: Visual saliency is a bottom-up cue that influences attention. This is a saliency map for the scene in Figure 8.1.

Visual Saliency

The visual saliency score S is calculated for a 2-dimensional snapshot of the scene taken from the user's perspective by a camera mounted above and behind the user. It is critical for this work that the scene can be observed from near the user's point of view, which applies some limitation to the sensor setup.

Using features such as color, orientation, and intensity, we generate a saliency map of the scene from the user's point of view (Figure 8.3). An object's saliency score S_i is incremented every time the saliency value for a pixel in a saliency map corresponding to o_i is above the average saliency for all pixels in that map.

Verbal Context

The verbal context V is a calculation of the proportion of descriptor words in an utterance that match descriptors of the object. In our implementation of this system, behaviors β can include utterances u of descriptive words such as "red" and "hammer." This score ignores linking words like "the" and "and." The specification for objects o includes a corresponding list of descriptors for that object.

$$V_{i,j} = \sum_{w \in u} \frac{\text{member}(w, D)}{|D|} \tag{8.7}$$

for utterance u, words w, and object descriptors D, where

$$\text{member}(w, D) = \begin{cases} 1 & \text{if } w \in D \\ 0 & \text{otherwise} \end{cases} \tag{8.8}$$

The range of values for $V_{i,j}$ is $[0, 1]$. If all of the descriptive words in an utterance describe an object, $V_{i,j} = 1$.

We demonstrate in this chapter that our system works even with this very simplistic language model. However, a more complex language system could be easily implemented as part of this nonverbal behavior model; it would simply have to generate a verbal reference score $V_{i,j}$.

Head Orientation and Pointing

To identify whether an object is within the cone of attention indicated by head orientation and pointing, we use a raytracing model. Each ray begins at the origin h of the behavior (i.e., head or hand). The yaw θ and pitch ϕ of the rays are measured as angles off the center of attention and determine the attenuation of the ray's communicative strength relative to the utilized modality. As an example, because head orientation yields a larger but less focused cone of attention than pointing, the attenuation function (Equation 8.10) provides a slower decay of signal strength as it diverges from the focus. This is mediated by the total communicative power being distributed over a wider volume of space, effectively reducing the baseline strength of any single head-originating ray. These attenuation functions effectively determine the strength and spread parameters that define the communicative meaning of head orientation and pointing (Figure 8.4).

To calculate gaze score $G_{i,j}$ for a given object $o_i \in O$ and behavior $\beta_j \in B$, we

find

$$G_{i,j} = \int_{\theta} \int_{\phi} a^G(\theta, \phi) \cdot I(h, \theta, \phi, o_i)^{-1} \cdot r_G \ \mathrm{d}\phi \, \mathrm{d}\theta \tag{8.9}$$

where θ and ϕ comprise the pitch and yaw of the robot's field of view, a_G is the focus attenuation function, I is the distance between ray origin (h) and intersection with o_i (equal to ∞ if no intersection), and r_G is the base score attributed to a perfectly focused gaze ray.

To represent the attenuation of gaze rays as their angular deviation increases from the center ray, we devise the gaze attenuation function

$$a^G(\theta, \phi) = (1 + \theta^2 + \phi^2)^{-1} \tag{8.10}$$

This function indicates that gaze signal strength diminishes congruent to the inverse squared distance of the angular deviation.

Computing the pointing score $P_{i,j}$ for object o_i is identical to Equation 8.9, but with a different base ray score r_P and pointing attenuation function

$$a^P(\theta, \phi) = (1 + e^{\sqrt{\theta^2 + \phi^2}})^{-1} \cdot 2 \tag{8.11}$$

This function indicates that pointing signal strength is more tightly concentrated, decaying exponentially with deviation from focus.

8.3.3 Behavior Selection (S)

Once each object has a likelihood score under each behavior, the model can select the behavior that maximizes the likelihood of the target object compared to other objects, while minimizing the behavior cost. Recall that we denote the likelihood score of an object i under behavior j as σ_{i_j}.

Likelihood scores represent a *relative* likelihood of reference, so what matters is

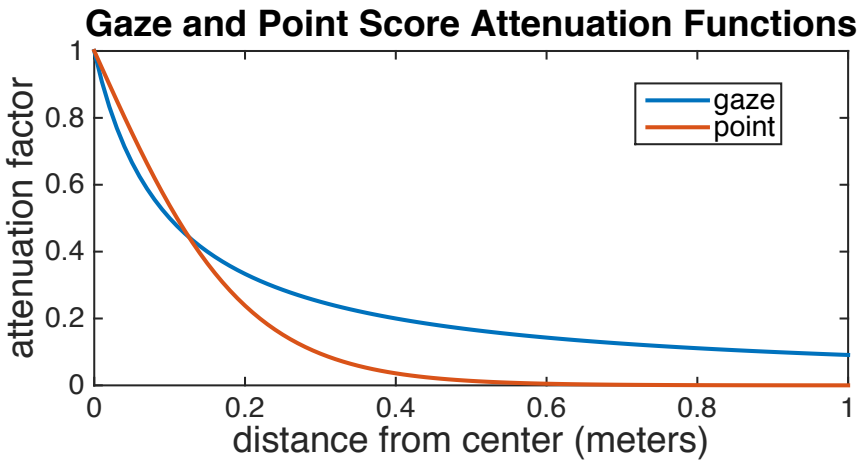


Figure 8.4: Attenuation functions for gaze and pointing cones.

the value of the target object's likelihood score, σ_τ, relative to every other object's likelihood score. Therefore, we calculate the behavior selection score b for a specific behavior β_j as the number of standard deviations the likelihood score for the target object, $\sigma_{\tau,j}$, is above the mean likelihood score $\bar{\sigma_j}$.

$$b_j = \frac{\sigma_{\tau,j} - \bar{\sigma_j}}{\sqrt{\frac{\sum_{i=0}^{n}(\sigma_{i,j} - \bar{\sigma_j})^2}{n-1}}} \tag{8.12}$$

This value must be greater than the empirically-determined disambiguation threshold ($b > \alpha$). We can formalize this as an optimization that selects a behavior of minimum cost while maintaining a behavior score above threshold. In other words, if c_{β_j} represents behavior β_j's index in ranking C, select

$$\min_j(c_{\beta_j}) \text{ such that } b_j > \alpha \tag{8.13}$$

8.4 Evaluation

To evaluate our model, we compare the model's suggested behavior with how people actually performed at identifying the target object on a variety of novel visual

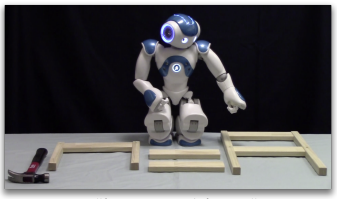
Figure 8.5: The in person evaluation demonstrates the model operating in a real-world human-robot interaction.

scenes. We find that the model correctly suggests the nonverbal behaviors required to maintain high user accuracy in identifying the target object in these scenes.

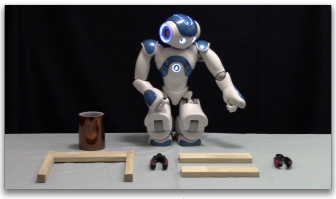
To perform this evaluation, we implemented the model from Section 8.3 on a humanoid robot in a collaborative building task and evaluated it in two experiments. The first experiment, a video-based analysis, allowed us to collect data from a large number of participants using a variety of visual scenes and provides initial confirmation of the model's function. The second experiment, an in person human-robot interaction, demonstrates that the model works well in real-world environments (Figure 8.5).

8.4.1 Study Design

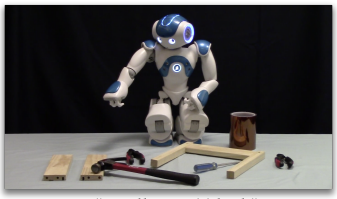
In both evaluations, a Nao robot references objects in the scene, such as parts of an Ikea chair or a set of Lego-like blocks. The robot's behavioral repertoire (B) contains three actions: speaking, gazing, and pointing. Nao's built-in text-to-speech generator was used for verbal references. The robot gazes at an object by orienting the center of its face toward that object. Because Nao does not have independently maneuverable eyes, head orientation serves as a substitute for true gaze, which involves eye movement as well as head movement. The robot points to an object by extending



(a) Low ambiguity: Obvious verbal reference does not need disambiguation.



(b) Medium ambiguity: Ambiguous verbal reference disambiguated with gaze or pointing.



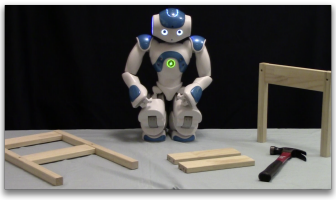
(c) High ambiguity: Ambiguous verbal and gaze reference, disambiguated with pointing.

Figure 8.6: Three scenarios show the range of scene difficulties. The model selects nonverbal behaviors only when they are required to disambiguate a verbal reference.

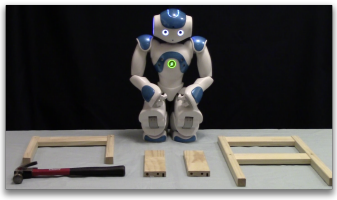
its arm to create a ray from shoulder joint through wrist joint to the center of the target object. The robot uses whichever arm is closer to the object to point. The cost ranking of these behaviors (C) was determined according to energy expenditure. In order from least to greatest cost, the behaviors are speaking, head turning, and pointing.

Higher cost behaviors may be needed when an interaction scenario is ambiguous. We can categorize the ambiguity of interaction scenarios along two dimensions: verbal ambiguity and visual ambiguity. Verbal ambiguity occurs when objects share descriptor words, so that verbal references do not uniquely identify one object. For example, in Figure 8.6(a), the verbal reference "large wood frame" is verbally unambiguous, because only one such object is visible, but the verbal reference "small wood block" would be verbally ambiguous because this phrase could refer to two such objects. Visual ambiguity occurs when there are many similar objects near the target, making gaze and pointing references unclear. For example, Figure 8.6(c) shows high visual ambiguity for the small wood blocks, which are side-by-side.

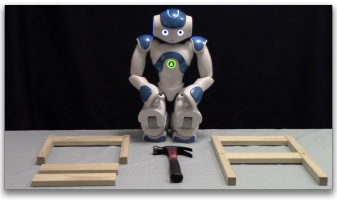
For our evaluation, we select three distinct ambiguity levels along this two-dimensional scale. In low ambiguity scenes, the target object is both verbally and visually unambiguous (Figure 8.6(a)). Medium ambiguity scenes have have high verbal ambiguity but low visual ambiguity (Figure 8.6(b)). High ambiguity scenes are both visually



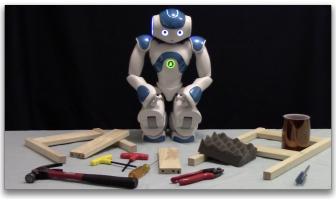
(a) Scene 1



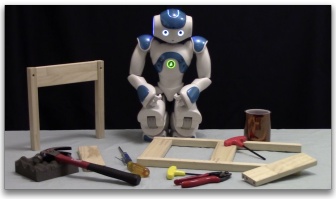
(b) Scene 2



(c) Scene 3



(d) Scene 4



(e) Scene 5

Figure 8.7: These five scenes were used to train the weights ω in Equation 8.6. The robot referenced objects in these scenes by speaking, orienting its head, and pointing.

and verbally ambiguous (Figure 8.6(c)). We did not select scenes with low verbal ambiguity and high visual ambiguity because these are unlikely to require nonverbal behaviors for disambiguation.

We determined the correct object reference behaviors empirically by evaluating human viewers' accuracies in identifying the target object in these three difficulty levels. We establish a threshold of 70% accuracy for a "successful" reference. Though the two evaluations use different scenes, the process is the same: assess human accuracy at identifying the target object under the different behaviors in the robot's repertoire, then compare which behavior was successful for human understanding against the behavior predicted by the model.

8.4.2 Empirically Determined Parameters

Before beginning the evaluations, some model parameters must be set, such as the weights ω_i in the linear sum (Equation 8.6), and the disambiguation threshold α. We empirically determined values for these parameters by performing an initial user study.

We seek to identify the relative importance of the four features identified in Section 8.3.2 on accuracy and naturalness: low-level visual saliency, high-level verbal context, gaze behaviors, and pointing behaviors. We do so by using linear regression to model the effects of each feature on human performance when identifying a target object in a number of example scenes.

We recorded video clips of a Nao robot referencing objects on the table in front of it (Figure 8.7). A reference involved one of seven modes representing combinations of verbal, gaze and gesture behaviors: verbal only; head orientation only; pointing only; verbal and head orientation; verbal and pointing; head orientation and pointing; and all of verbal, head orientation, and pointing. The objects include tools like hammers and wrenches, as well as pieces of wood and basic shapes like cylinders. We recorded each of the seven references to 16 target objects (across five different scenes) for a total of 112 videos.

For each video clip, which lasted between 10 and 15 seconds, we asked users to identify which object the robot was indicating (forced multiple choice with a drop-down menu) and to rate its naturalness on a scale from -3 to 3, where lower numbers meant "very unnatural" and higher numbers meant "very natural."

Participants were recruited from Amazon Mechanical Turk. One hundred users completed the study, with a mean accuracy of 85%. The data from three users was eliminated for failing the preliminary evaluation question and having an accuracy below the 95% confidence interval. Each user evaluated 21 videos, so our analysis is based on a total of 2,037 examples.

To find the relative importance of each feature, we conducted a stepwise linear regression to calculate regression coefficients for each predictor variable. We used five predictor variables: the three robot actions (speaking, head turning, and pointing), scene complexity, and target object salience. The three robot action predictor variables are categorical, simply indicating their presence or absence in the video. Simi-

Measure (y)	R^2 (adjusted)	Predictor	β	Significance
Accuracy	0.181	Pointing	0.10	$p < 0.001$
		Speech	0.29	$p < 0.001$
		Complexity	0.10	$p < 0.001$
Naturalness	0.216	Head orientation	0.91	$p < 0.001$
		Pointing	1.64	$p < 0.001$
		Speech	0.57	$p < 0.001$
		Complexity	0.29	$p < 0.001$

Table 8.1: Linear regression models for features of the scene used to estimate where a user's attention will be drawn.

larly, the scene complexity is categorical, rated either "high" or "low." The salience predictor is an integer value corresponding to the saliency of the target object in that scene, calculated using the SaliencyToolbox, a Matlab software implementation [260] of Itti's saliency model [125].

We constructed two linear models: one for accuracy, using a boolean response variable indicating whether the user correctly identified the target object, and one for naturalness, for which the response variable was the naturalness score between -3 and 3.

The results of the linear regression are shown in Table 8.1. Because we want to maximize both accuracy and naturalness, we used the coefficient values from the more comprehensive model (i.e., the one with the higher adjusted R^2). Therefore, we set each ω in Equation 8.6 to its corresponding β value in the naturalness linear model. Because saliency was not found to be a significant part of the model, we set $\omega_s = 0$, so saliency is not considered in the video-based evaluation below (though it was included in the in person evaluation, see Section 8.4.4). We also determined that $\alpha = 0.75$ provides sufficient object reference disambiguation.

Scene Ambiguity	Metric	Correctness Threshold	Behavior			
			Verbal	Verbal +head	Verbal +point	Verbal +head +point
Low	Model prediction	0.75	1.096	1.537	1.607	1.626
	Human accuracy	70%	100%	100%	100%	100%
Medium	Model prediction	0.75	0.0	0.886	0.926	1.573
	Human accuracy	70%	11%	84%	80%	81%
High	Model prediction	0.75	0.236	-0.209	1.510	1.484
	Human accuracy	70%	48%	36%	73%	78%

Table 8.2: Results for video-based evaluation. For each scene (Figure 8.6), the model prediction row contains behavior selection scores b and the human performance row contains human accuracy rates. Results are colored green if they are above the correctness threshold set for this evaluation (Section 8.4.3).

8.4.3 Evaluation 1: Video-Based

The video-based evaluation used novel scenes containing the same objects as the parameter training videos (Section 8.4.2) but in different configurations (Figure 8.6). Evaluating novel scenes demonstrates that the empirically-determined parameter weights generalize to new scene configurations.

Seventy-five participants (recruited on Mechanical Turk) were shown videos of the robot in various scenes, including the three test scenes in Figure 8.6. In each video, the robot exhibited one of four behaviors: verbal reference only (no movement); verbal and gaze (head movement simultaneous with speech); verbal and gesture (pointing simultaneous with speech); or verbal, gaze, and gesture (head movement and pointing simultaneous with speech).

Table 8.2 shows the model's predicted behaviors as well as participant accuracy rates for each robot behavior on the three ambiguity level scenarios. Red values indicate certainty scores (for model prediction) or accuracy rates (for human accuracy) that fall below the predetermined threshold ($\alpha = 0.75$ and 70%, respectively). Green values indicate scores or accuracy rates above threshold.

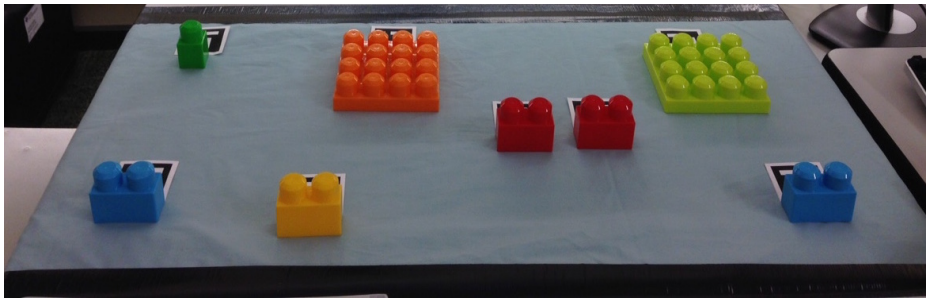


Figure 8.8: Participant view of blocks in the in person evaluation.

As expected, users were extremely accurate using only verbal references for scenario 1, which has low ambiguity. Verbal references were insufficient for distinguishing between the two red pliers in scenario 2, which has medium ambiguity, though users were accurate when the robot displayed any of the nonverbal behaviors. Finally, users had low accuracy on scenario 3 (the high ambiguity scene) with just head orientation as disambiguation, but higher accuracy with pointing behaviors as disambiguation.

Comparing our model's predictions for each scenario with empirical human accuracy, we can see that our model successfully selected the least-cost behavior that still yielded high response accuracies from human users on the example scenarios.

8.4.4 Evaluation 2: In Person

Instead of Ikea furniture and lab tools, the in person evaluation uses brightly colored Lego-type blocks with attached fiducial markers to simplify real-world object recognition (Figure 8.8). The bright colors of the blocks allow the system to automatically identify object positions through color segmentation, and fiducial markers attached to the objects enable the system to uniquely identify the blocks, even those that are visually identical. The robot in this study performed real-time object localization using a Kinect v2, and its pointing and gaze behaviors were not pre-scripted. Fast saliency calculation was implemented using a Python-based computer vision algorithm [255].

In this experiment, the three ambiguity levels described in Section 8.4.1 are com-

Scene Ambiguity	Metric	Correctness Threshold	Behavior			
			Verbal	Verbal +head	Verbal +point	Verbal +head +point
Low	Model prediction	2.0	2.252	2.405	2.437	2.454
	Human accuracy	70%	99%	99%	99%	100%
Medium	Model prediction	2.0	1.239	2.127	2.260	2.391
	Human accuracy	70%	63%	83%	85%	85%
High	Model prediction	2.0	1.366	1.848	2.148	2.196
	Human accuracy	70%	37%	68%	75%	65%

Table 8.3: Results for in person evaluation. For each ambiguity level, the model prediction row contains behavior selection scores b and the human performance row contains human accuracy rates. Results are colored green if they are above the correctness threshold set for this evaluation (Section 8.4.4).

bined into a single scene. As can be seen in Figure 8.8, four blocks (orange, lime, green, and yellow) have low ambiguity, two blocks (the pair of blue) have medium ambiguity, and two blocks (the pair of red) have high ambiguity.

A few changes were made to the model's parameter values based on pilot testing. The disambiguation value (α) was set to 2.0. The weight on verbal scores (ω_v) was raised to 1.07 from 0.57, and the weight of saliency scores (ω_s) was set to 0.25 to account for the importance of visual saliency in real-world scenes. All of the other parameter values were carried over from the empirical evaluation (Section 8.4.2), despite the shift from screen-based to real-world interaction.

Twenty people participated in this study. The robot performed each of the four possible object reference behaviors toward each of the eight objects on the table, for a total of 32 object references per person. The order of object references was randomized for each participant.

Table 8.3 lists the model's predictions and peoples' accuracies in identifying the target object for each of the ambiguity levels and referential behaviors. As expected, participants' performance followed a similar trend to the video-based performance, with high accuracies across the four behaviors for low ambiguity scenarios, high ac-

curacy with any nonverbal behavior for the medium ambiguity scenarios, and high accuracy only with the more precise nonverbal behaviors for the high ambiguity scenarios. Red values indicate scores or accuracy rates that fall below the correctness threshold ($\alpha = 2.0$ and 70% for model predictions and human accuracy, respectively), while green values indicate scores and accuracies that fall above threshold.

One anomalous finding is that people seemed to struggle with ambiguity when the robot performed both head turn and pointing in the high ambiguity condition. The accuracy for the head turn behavior was similar to the accuracy for head turn and pointing behavior, and lower than the target accuracy of 70%. This may be caused by the ambiguity introduced by the head turn behavior, which actually serves to weaken the pointing cue. More research is needed to identify whether pointing alone is a stronger cue than pointing with a head turn in some conditions, particularly in high ambiguity scenes.

8.5 Discussion

This chapter presents a model for generating robot behaviors that help guide a user's attention toward a target object in a minimally distracting but maximally communicative way. It takes the user's perspective to simulate where their attention will be directed in response to the robot's behaviors, then selects the most effective, least expensive behavior.

In this chapter, we implemented the model for a particular assembly task with a Nao robot, but the general model described in section 8.3 is flexible. Different robot capabilities, sensing requirements, and cost evaluations can be implemented to best suit the particular task at hand.

The model assumes that the robot and human users have relatively similar, unobstructed, complete views of the entire workspace. In particular, we assume that

there are no substantial differences in what either agent can perceive of the scene. Accounting for occlusion and partial information, particularly when that information is not shared, would allow the model to be extended to more complex interactions.

Similarly, the model currently requires a view of the environment from the user's perspective in order to calculate the saliency score on the visual scene. This requirement for perspective taking presents a limitation on the sensor setup, because a camera has to be available from approximately the participant's viewpoint. A preliminary solution to this problem is to mount a fixed camera above and behind the user's expected position. However, this requires knowing in advance where the user will be and restricting their movements around the environment. A more elegant solution is to capture a view of the scene from a head-mounted camera, for instance a camera embedded in the user's eye glasses. This allows the user to move about the environment while collecting an accurate user view.

Though the model has been successfully implemented and evaluated in the real world, it does not yet operate in real time due to the time requirements of saliency and raytracing algorithms. Currently, the robot pre-calculates saliency, gaze, and gesture scores before each interaction, which takes up to 10 seconds. Improvements to these algorithms will make real-time interaction feasible.

This model does not account for temporal dynamics of communication. For example, a user's attention may shift as a sentence unfolds and more information is provided to narrow the range of possible references. Similarly, saliency is a relevant feature in tasks that require quick responses, and therefore rely on instantaneous scene evaluations, but the saliency of an object fades with habituation to that scene. A more complex model would account for such temporal dynamics as part of the attention estimation function.

These limitations provide a broad scope for future implementation of the model, which shows promise for making human-robot interaction more comfortable, natural,

and efficient.

8.6 Summary

In this chapter, we introduced a flexible model for generating nonverbal behaviors to communicate spatial references. This model, inspired by psychological understanding of top-down and bottom-up influences on cognition, calculates a nonverbal behavior for a robot to perform that most clearly indicates the target object while minimizing the cost to perform that behavior. The model successfully selects the lowest-cost nonverbal behavior that conveys the spatial reference effectively, as measured in two evaluations, one online and one in person.

The model described in this chapter is flexible. It can be used for robots with different nonverbal capabilities; for example, for a robot with independently movable eyes, the model could include eye shifts as a behavior along with head turns and pointing. Furthermore, the model is not dependent on a particular scene or type of interaction; unlike data-driven models (such as the one in Chapter 7), our model uses features of the scene which are detected dynamically by cameras. Therefore, it can be used to generate a robot's spatial references in a variety of interactions, from tutoring to collaborative manufacturing.

The evaluations of the model in this chapter were fairly contrived: they measured people's accuracy in response to spatial references in a series of discrete, disconnected trials. While this is important for initially validating that the model performs as expected, it does not indicate how the model will perform in a more naturalistic interaction. In the next chapter, we use the nonverbal behavior model described here to generate a robot's nonverbal behavior during a collaborative human-robot construction task, to understand how the scene-dependent generation of nonverbal behavior impacts human performance.

9

Nonverbal Communication in Human-Robot Collaboration*

The previous chapter detailed a model for generating robot referential behavior. In this chapter, we apply this model in a human-robot collaboration. We investigate whether the usefulness of nonverbal behaviors generated by the model changes based on task difficulty. A robot provides instructions for people to construct structures out of Lego blocks. We manipulate the difficulty of a spatial memorization task in two ways: by adding steps to memorize, and by introducing an interruption. We analyze how a robot's deictic nonverbal behavior (looking and pointing), which accompanies the spatial instructions to be memorized, affects people's recall and task completion times under differing difficulty levels. Results indicate that for easy tasks, people generally perform at a high level already, and instructions provided with nonverbal behaviors don't improve performance compared to instructions delivered by speech alone. However, for difficult tasks, seeing the nonverbal behaviors

*This work was originally published as:
Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. Robot nonverbal behavior improves task performance in difficult collaborations. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

led to higher accuracy and shorter completion times, indicating that as the task became more difficult, the referential nonverbal behaviors mitigated the negative effects of task difficulty. In short, nonverbal behavior may be even more valuable for difficult collaborations than for easy ones.

9.1 Introduction

People use nonverbal behaviors (NVBs) to augment spoken references, clarify ambiguous language, and convey attention, among many other functions [28, 99, 164]. Joint activity, which involves coordinating action among two partners, requires NVBs that direct attention to particular objects or regions of space [66]. These actions can take the form of pointing (i.e., *deictic*) gestures, which can be enacted with the hand, the head, or other body parts [66, 164]. In this chapter, we focus on two specific deictic NVBs: pointing with the hand and looking with the head.

Robots can take advantage of deictic NVBs to improve human-robot collaborations. For example, imagine a robot assistant on a factory floor that is training a new employee in how to construct an assembly out of component parts. The robot can look and point to the parts as it refers to them in order to clarify the references. This is especially important when there are multiple parts that can be described the same way, but need to be placed in a particular order, for example, a left and right version of the same bracket piece. Instead of saying "the left bracket piece," the robot can say "that bracket piece" and use pointing to disambiguate the reference.

Human-robot interaction research has shown that people can benefit from this kind of deictic NVB from robots. Pointing and gaze from robots during object references allows people to more quickly locate objects and to disambiguate object references, increasing the efficiency of the collaborations [3, 44, 115]. People also have more positive evaluations of a robot when it uses gestures along with speech [210].

In this chapter, we ask whether NVBs work more effectively in some tasks than in others. In particular, we explore whether the difficulty of a task affects how well a robot's deictic NVBs serve to communicate spatial references. To answer this research question, people are asked to complete a memorization task based on instructions provided by a humanoid robot. We manipulate task difficulty in two ways: by increasing the number of steps people need to memorize, and by introducing an interruption that distracts people momentarily from their task.

We hypothesize that:

H1 Using nonverbal behaviors while providing spatial task instructions will improve recall accuracy and reduce task completion times,

H2 When the task difficulty increases, the effect of nonverbal behaviors will increase, and

H3 A robot that displays nonverbal behaviors will be rated more positively than a robot that only uses speech for communication.

Generating NVBs for robots is not trivial. A naïve NVB controller for a robot might always select all possible nonverbal behaviors, looking and pointing at every possible reference. But there is a benefit to being selective about generating NVBs. Frequent nonverbal behavior is undesirable when it engages the effectors that the robot might otherwise need, such as hands for object manipulation and head for vision. Additionally, in human collaborations, people use nonverbal behaviors as subtle, implicit mechanisms of communication [219], so excessive NVBs may be visually or cognitively distracting to a viewer. For robots powered by batteries, the energy expenditure from moving effectors to perform NVBs might also become a concern.

For these reasons, we have designed a behavior model that is selective about when to generate NVBs. The model considers elements of the scene and the task to select

the most communicative and least expensive NVBs for the particular reference and environment at hand. This model is described in Chapter 8.

In this chapter, we use the model to generate deictic NVBs for a human-robot collaboration in order to investigate the effectiveness of NVBs under different task difficulties. Spatial collaboration, like the task employed in this study, involves manipulating and moving objects in the environment. Because the position of these objects is not restricted, the model cannot simply pre-script the NVBs for each object. Instead, our real-time robot behavior model continually calculates the best NVB for each object reference as the objects in the environment are manipulated.

Section 9.3 provides details about the implementation of the model and the experiment. Section 9.4 describes the results of the study, and Section 9.5 discusses these results as well as suggestions for future research.

9.2 Related Work

Deixis is a critical part of cooperative action between people [66]. In particular, people use deictic gestures like pointing to focus attention on a target spatial region [32]. As pointing becomes more precise (because the pointing targets are closer), people rely more on pointing and less on language for references [32]. Deictic gestures are especially useful in communicating how to assemble objects [160], which is the task we have selected for the present study.

Human-robot interaction (HRI) research has shown the benefit of deictic gesture to human-robot collaboration. Implicit nonverbal communication—including deixis using gaze and pointing—makes a robot more understandable, increases the efficiency of task performance, and reduces the impact of errors from miscommunication [47]. When robots are providing instructions or referencing objects, people use robots' deictic gestures to improve their task speed and efficiency [3, 44, 115]. Robots can

even use deictic gaze to subtly influence people's selections of objects without those people realizing it [175].

Deictic NVB is an effective mechanism for robot communication. People rate robots that gesture along with their speech more highly than robots that do not show any nonverbal behavior [210], and multimodal deixis (for example, looking and pointing) is better than unimodal deixis (i.e., pointing or looking alone) [230]. Cooperative gestures are most effective when they are presented frontally and with machine-like "abrupt" motion [201]. Interestingly, the saliency of an object in a cluttered environment has only a small effect on people's interpretation of a robot's pointing behaviors [230].

Computational models of NVB allow robots to generate their own gaze and gestures in response to the context of the interaction. Some of these models are based on empirical examples for human performance, such as data-driven models of tutoring [8] and narration [116]. Others are based on contextual and semantic knowledge [111, 179].

In this chapter, we use the NVB model detailed in Chapter 8, which takes into account the user's perspective to select the correct deictic behavior for object references. Some robot behavior generators take a similar approach, modeling the user's perspective to select the most appropriate deictic behaviors for providing route directions [185], references to people nearby [158], or even object references as in the current work [111]. A robot that simulates human cognition when selecting deictic behaviors for spatial references can more effectively convey to people the region of space to which it refers [105]. Our model is different from prior work because it applies to object references, not to people or spatial areas, and because it uses both top-down and bottom-up cues from the scene to model the user's perspective. Both top-down and bottom-up attentional processes are important components of fluid joint action between people and robots [109].

9.3 Experiment

To evaluate the effect of nonverbal behavior on people's performance during interrupted tasks, we conducted an in-person human-robot interaction study. In this study, we test people's memory for a set of assembly instructions given by a robot. For some participants, the robot uses NVBs selected by the model detailed in Chapter 8 to augment its spoken instructions. We compare people's performance with or without NVBs and at various levels of task difficulty to evaluate how NVBs affect instruction recall and task efficiency.

9.3.1 Design

Experimental variables

This study has three between-subjects independent variables.

- *NVB* is "present" or "absent" depending on whether or not the robot displays nonverbal behaviors when providing the assembly instructions
- *Memorization load* is "low" or "high" depending on the number of steps in the assembly to be memorized
- *Interruption* is "present" or "absent" depending on whether or not the user is interrupted during their completion of the task

Therefore, this study has a 2 (NVB) $\times$ 2 (memorization) $\times$ 2 (interruption) design, which results in eight conditions. Participants are randomly assigned to one of these conditions.

The nonverbal behaviors in the NVB condition are *looking* and *pointing*. These behaviors are autonomously generated in real time in response to object references using the model described in Chapter 8. Details of the model implementation for this experiment are in Section 9.3.2.

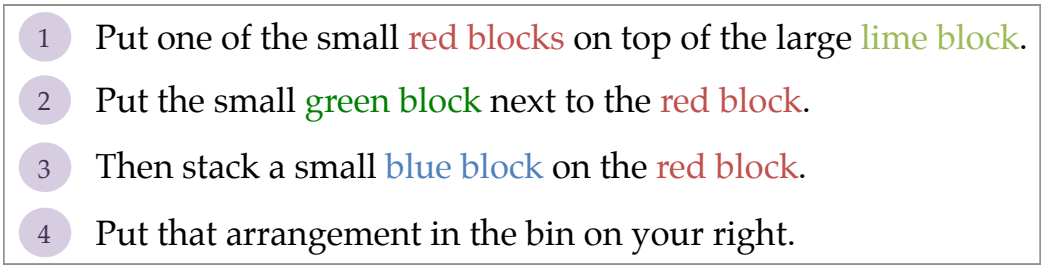


Figure 9.1: An example of the steps for the construction of one assembly in the high memorization condition. Assemblies in the high memorization condition have four steps, and those in the low memorization condition have three steps. Low memorization assemblies are generated by removing the third step of a high memorization assembly.

We employ two strategies for changing the difficulty of the task. The first is an increase in memorization load (Figure 9.1). Low memorization assemblies involve three steps for completion, and high memorization assemblies require four steps. In both cases, the final assembly step is always an instruction to place the assembly in a particular bin. The other assembly steps involve a subject, a spatial relation, and a target. For example, "put the small green block next to the red block" involves the *green block* (subject), *next to* (spatial relation), and *red block* (target).

The second strategy for changing task difficulty is interruption. In this study, an interruption involves completing a mental rotation test [14] (Figure 9.2). In the test, participants are shown pictures of a target shape and four possible rotations of that shape. They are asked to select the image that correctly represents what the target shape would look like when rotated. Participants who are interrupted complete eight such questions with a time limit of four minutes. We selected a mental rotation test as an interruption to try to interfere with people's spatial and visual memory for the assembly instructions.

The two difficulty manipulations provide different types of challenges. Increasing memorization load puts greater strain on working memory. Because each step of the task requires memorizing two objects (the subject and target), the number of object references to be memorized in each task goes from six in the low memorization

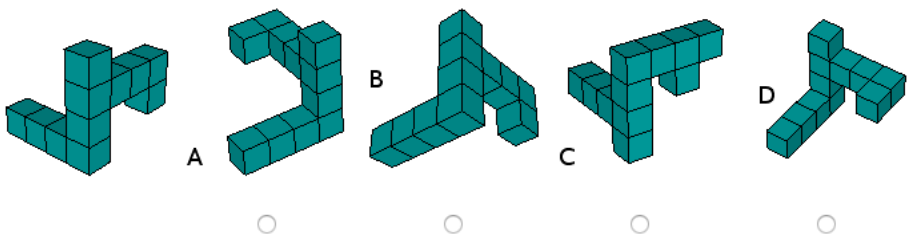


Figure 9.2: An example of a mental rotation question used in the interruption. The correct answer is B. Courtesy of [14].

condition to eight in the high memorization condition, approaching the 7 ± 2 limit to working memory [167]. The interruption, in contrast, presents an unexpected and rapid shift of attention. It was selected to mimic a distraction that might occur during any type of human-robot collaboration.

Measures

There are two objective measures and one subjective measure in this study. The first objective measure is *recall accuracy*, how well a participant follows a robot's instructions as measured by the number of correct steps the participant completes in each assembly. Each step is scored individually for accuracy, with one point awarded for each correct object or relation. Therefore, participants can receive partial credit for an assembly even if some of the steps are completed incorrectly. For example, if the step's instruction was "put one of the small red blocks on top of the large lime block," but the participant put a small blue block underneath the large lime block, they would lose two of the three possible points (for missing the red block and the "on top" relation), but would still be awarded one point (for involving the lime block).

The second objective measure is the *completion time*, how long it takes the participant to put the blocks together once they are given the instructions. Completion time is measured from the moment the robot finishes its instructions to the moment the participant indicates that they are done with the task (see Section 9.3.3 for details). Lower completion times mean more efficient interactions.

The subjective measure is people's perceptions of the robot. Specifically, we evaluate people's perceptions of the robot's animacy, anthropomorphism, intelligence, and likability, using the Godspeed survey [34]. This standardized human-robot interaction questionnaire has five or six Likert-scale questions for each of the four perception items we are studying.

9.3.2 Apparatus

The robot in this study is a 58 centimeter tall humanoid called Nao. We used two degrees of freedom in Nao's head to enact looking behaviors and six degrees of freedom in Nao's arms for pointing behaviors. Nao's speech was generated using the robot's built-in text-to-speech system.

Participants constructed different assemblies using eight brightly colored Mega Blocks. A Microsoft Kinect v2 sensor provided real-time sensing capabilities for the Nao, enabling it to detect the blocks in real time and to track their positions in 3D space. Each block had a fiducial marker attached, so that the Kinect could uniquely identify blocks using the augmented reality library ArUco [97]. The markers were attached along each block's edge, so the center of a marker did not represent the center of a block. To ensure that the Nao's deixis would be correctly aimed at the center of the blocks, the Kinect found block centers using color segmentation and blob detection techniques from the OpenCV library [45], matching the marker closest to a given block center as that block's identifier.

Because object detection and nonverbal behavior modeling occurred in real time, the NVBs in this study were not pre-scripted. The NVBs a particular participant saw depended on the block layout. Though all participants began with the same block layout, as they manipulated the blocks, the NVB to each block was re-calculated based on its new position.

In every condition, Nao shifted its weight slightly from foot to foot to simulate

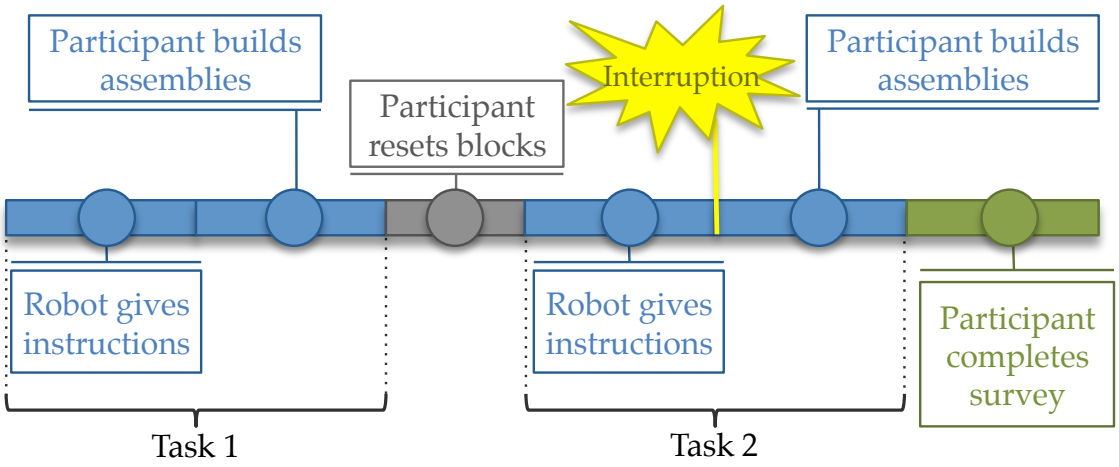


Figure 9.3: A timeline of the interaction. If the participant was in the interruption condition, an interruption occurred between the robot's instructions and the participant's assembly in task 2.

animacy when it was not providing task instructions. When performing computationally expensive actions like calculating saliency scores for each object, which required several seconds, Nao scanned left and right with its head to simulate looking at all of the objects on the table.

9.3.3 Methods

We collected data from 48 participants recruited from a university campus (mean age 26; 25 male, 21 female, 2 other or preferred not to respond). Participants were compensated $5 for this 30 minute study. Participants were randomly assigned to an NVB, memorization, and interruption condition.

We used the same experimental setup in this study as in the in person validation study from Chapter 8. Thus, Figure 8.5 shows the room and object layout for this study.

Figure 9.3 provides a visual timeline of the interaction. Participants performed two construction tasks, one after the other. Each task was comprised of two assemblies. For each assembly, Nao provided a set of verbal instructions (Figure 9.1). For

participants in the NVB condition, these verbal instructions were augmented with simultaneous looking and pointing behaviors generated by the model described in Chapter 8.

There was a timer on the computer screen next to the participant. After Nao was done giving its instructions for the task, it told participants to press "start" on the timer and begin putting together the blocks, and to press "stop" when they were finished. Task completion time is measured from when the robot finishes its instructions to when the participant pressed "stop" on the timer, in order to account for time they spent thinking before pressing the "start" button.

For participants in the interruption condition, an interruption occurred just after Nao finished providing instructions to task 2 but before participants could start assembling the blocks. During the interruption, the experimenter came into the room, placed the robot in an idling mode by tapping its head once, and asked the participant to complete a mental rotation test (detailed in Section 9.3.1). The test itself had a four minute time limit, and the total interruption time was approximately five minutes, though it varied based on how quickly the participant completed the test. After the interruption, the robot was taken out of idling mode with a second head tap. It then prompted participants to begin the task 2 assembly.

Though we did not inform participants, task 1 is intended as a practice trial that allows people to familiarize themselves with the task and the robot. Results from task 2 are analyzed in Section 9.4.

At the end of the experiment, participants were asked to complete a questionnaire detailing their impressions of the robot and the task. They also provided demographic information at this time.

	Low Difficulty		High Difficulty	
	No Interruption	Interruption	No Interruption	Interruption
No NVB	0.976 (.04)	0.943 (.06)	0.675 (.21)	0.867 (.12)
NVB	0.929 (.11)	0.893 (.13)	0.917 (.05)	0.843 (.18)

Table 9.1: Average recall accuracy on task 2 for each of the eight conditions, written as mean (standard deviation).

	Low Difficulty		High Difficulty	
	No Interruption	Interruption	No Interruption	Interruption
No NVB	12.5 (2.6)	15.8 (2.0)	23.2 (5.3)	32.7 (7.9)
NVB	13.6 (2.1)	21.9 (5.3)	24.5 (6.5)	25.3 (10.9)

Table 9.2: Average completion time in seconds for task 2 in each of the eight conditions, written as mean (standard deviation).

9.4 Results

Two participants were excluded for noncompliance, so we examined data from 46 participants.

9.4.1 Objective Measures

To evaluate the behavioral effects of our manipulations, we examine the effect of the three experimental variables (memorization load, interruption, and NVB) on the two behavioral metrics (accuracy and completion time, both measured from the second task). Results are shown in Table 9.1 for accuracy and Table 9.2 for time.

We conducted a three-way analysis of variance (ANOVA) to measure the effects of our three independent variables on recall accuracy. The test revealed a statistically significant effect of memorization load ($F(1, 38) = 9.137, p = 0.004$) and a statistically significant interaction between memorization and NVB ($F(1, 38) = 4.713, p = 0.036$). Figure 9.4 illustrates this significant interaction. There was also a borderline significant interaction between interruption and NVB ($F(1, 38) = 3.397, p = 0.073$) and a

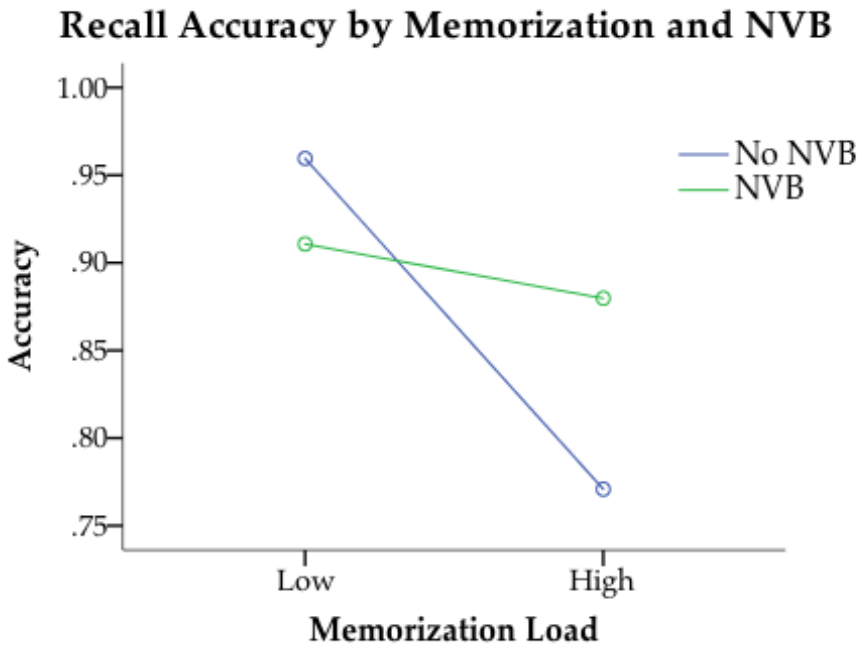


Figure 9.4: Accuracy of recall by memorization and NVB conditions. The interaction is significant ($p = 0.036$), indicating that NVB helped mitigate the difficulty of the task.

borderline significant three-way interaction among memorization, interruption, and NVB ($F(1, 38) = 3.278, p = 0.078$).

We investigate this three-way interaction with tests of simple effects, which reveal how one variable influences the others. First, we conduct a test for simple two-way interactions between interruption and NVB for each level of memorization. This simple two-way interaction yielded a significant effect for high memorization ($F(1, 38) = 6.554, p = 0.015$), but not for low memorization ($F(1, 38) = 0.001, p = 0.981$). This tells us that in the high memorization case, the effect of NVB on accuracy depends on whether an interruption occurs. Investigating further into the interaction, we run a test of simple simple main effects. We find a statistically significant effect of NVB on accuracy rates in the interruption absent condition ($F(1, 38) = 9.466, p = 0.004$) but not in any other conditions.

To evaluate the effect of our second objective measure, completion time, we con-

ducted a similar three-way ANOVA. For this test, we excluded the timing data from one participant whose response time (89 seconds) was an extreme outlier (> 3 SD from the mean). The test revealed a significant effect of interruption ($F(1, 37) = 8.629, p = 0.006$) and memorization ($F(1, 37) = 31.490, p < 0.001$). It also identified a borderline significant interaction between memorization and NVB ($F(1, 37) = 3.161, p = 0.084$) and a borderline significant three-way interaction between memorization, interruption, and NVB ($F(1, 37) = 3.362, p = 0.075$).

As with accuracy, we further investigate this three-way interaction with a test of simple effects. Testing for a simple two-way interaction between interruption and NVB did not yield significance for high memorization ($F(1, 37) = 2.639, p = 0.113$) or low memorization ($F(1, 37) = 0.917, p = 0.345$) conditions. However, a test of simple simple main effects showed a statistically significant influence of NVB on completion time for participants in the high memorization condition when an interruption occurred ($F(1, 37) = 4.330, p = 0.044$), but not without an interruption ($F(1, 37) = 0.101, p = 0.752$). In other words, in the high memorization condition, NVB mitigated the effects of the interruption on task completion time (Figure 9.5). There was no effect of NVB on the low memorization condition, either with or without an interruption.

9.4.2 Subjective Measures

Our subjective measure was user perception of the robot in terms of anthropomorphism, animacy, likability, and perceived intelligence. Each of these four items was measured by five or six Likert-type questions provided in a questionnaire. The items all had high internal consistency as determined by a Chronbach's alpha greater than 0.7.

We conducted a one-way ANOVA to determine if there were differences in people's responses to the four questionnaire items depending on which of the eight conditions

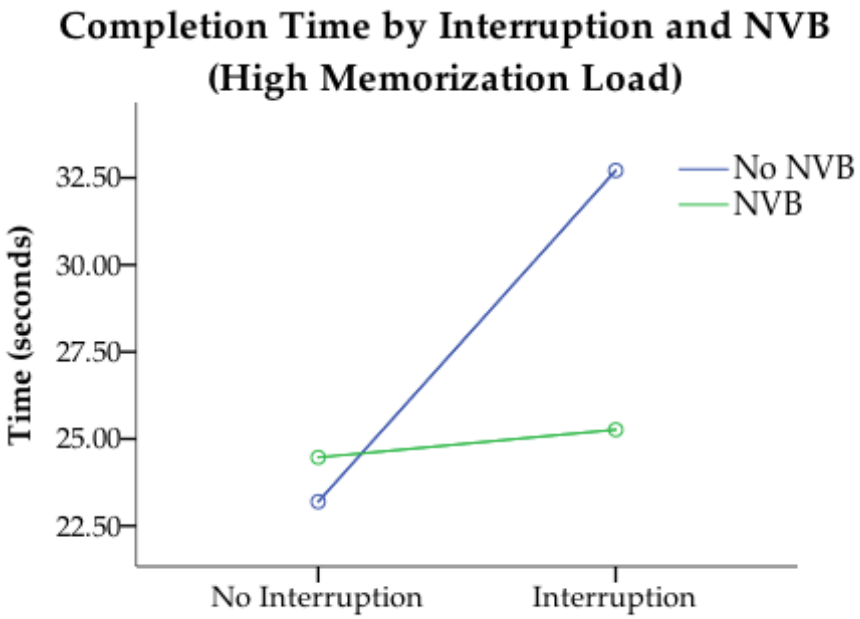


Figure 9.5: Completion times for interruption and NVB conditions, shown for the high memorization condition only. There is a significant simple simple main effect of NVB on completion time when an interruption occurs ($p = 0.044$) but not without an interruption.

they experienced. None of the experimental variables had statistically significant effects on these items (anthropomorphism: $F(7, 38) = 0.510, p = 0.821$; animacy: $F(7, 38) = 0.159, p = 0.992$; likeability: $F(7, 38) = 1.096, p = 0.385$; intelligence: $F(7, 38) = 1.621, p = 0.159$).

9.5 Discussion

Our first hypothesis predicted that task performance (in terms of recall accuracy and completion times) would improve when a robot used deictic NVBs to provide task instructions over when it only provided those instructions verbally. This hypothesis is not supported in the general case, because the results do not show a statistically significant effect of NVB across all memorization and interruption conditions.

However, there is a significant interaction effect between NVB and memorization

load for both recall accuracy and completion time. The interaction between NVB and interruption is also significant for completion time and borderline significant for recall accuracy. This indicates that for more difficult tasks—i.e., those with heavier memorization requirements or those in which the user's attention is distracted—NVBs do, in fact, have a positive effect on performance. Therefore, H2 is supported for both objective measures.

In short, the objective measure results from this study show that NVB has little effect on tasks that are already easy, but that when tasks become challenging, NVB improves task performance by increasing recall accuracy and decreasing completion times.

Our third hypothesis predicted that subjective evaluations of a robot's anthropomorphism, animacy, likability, and intelligence would be increased when the robot showed deictic NVBs while providing task instructions, over when the robot simply provided the instructions verbally. Our results do not support H3, because none of the items on the questionnaire reached significance.

This result is in contrast with other studies, which have shown that subjective perceptions of a robot are improved when the robot uses NVBs [116, 210]. While the current study only uses deictic NVBs, however, the previous studies also used expressive NVBs such as iconic or metaphoric gestures [164]. These types of gestures involve producing a visual representation of physical or abstract concepts, such as moving the hand up and down for "chopping" or signaling over the shoulder for "a long time ago." It may be possible that deictic behaviors, such as looking and pointing, do not elicit the same kind of perceptions of agency in a robot as other, more expressive gestures.

From the results, the difficulty manipulations used in this study ("low" or "high" memorization load, and "present" or "absent" interruption) seem approximately equally difficult. Recall accuracy is slightly worse for low memory, interrupted tasks

(89%) than for high memory, uninterrupted tasks (92%), while task completion times are slightly worse for high memory, uninterrupted tasks (24.5 seconds) than for low memory, interrupted tasks (21.9 seconds). One limitation of our study is that it only uses two levels for the two difficulty manipulations. Future work could investigate a range of difficulty levels to identify whether NVB helps even more with more difficult tasks and whether the effect plateaus at any point.

One novel feature of this study is the real time nonverbal behavior model that controlled the robot's actions. Because the model recalculated attention likelihood scores when blocks moved, the NVB a participant saw was specifically targeted toward the scene in front of them. As the results show, this NVB was effective in mediating the effects of a difficult task.

A primary principle of the behavior generation model is that too much NVB can be a hindrance to comprehension. This experiment did not evaluate this claim directly. A future study comparing NVBs produced by the model to other NVB generation models would elucidate how the scene-based model used here compares to other systems that potentially produce more NVBs during an interaction.

We do not claim that our model provides optimal behavior generation for spatial references. However, our model performs at least a subset of the optimal behaviors for nonverbal communication, as determined by the improvement of recall accuracy and completion times. Better results may be possible with a different behavior generation model, and future studies comparing such models would help identify what kinds of NVBs are useful in human-robot collaborations.

Additions to the model might improve its performance. For example, once a robot has named an object by pointing to it, the need to deictically refer to that object again may decrease for a short time afterward. This would add a "prior reference" factor in the likelihood equation, which would increase the likelihood of an object if it has been recently referenced. This factor can decay over time to capture the

temporal dynamics of attention. This and other modifications to the model could generate even more natural NVBs.

9.6 Summary

This chapter achieved two goals. First, it showed that the multimodal behavior model described in Chapter 8 was effective in a naturalistic human-robot interaction. The model successfully operated in a real-time collaboration and succeeded in improving people's performance on a joint construction task with a robot. This is promising for future application of this robot behavior model in new scenarios and with new robots.

Second, this chapter presented a novel HRI finding: that the benefit of nonverbal behavior depends on task difficulty. For easy tasks, NVB may not add much, but for harder tasks, NVB significantly improves performance. By manipulating the level of difficulty using interruptions and memorization load, but keeping the task domain and instructions the same, we showed that there are some conditions in which nonverbal behavior is more impactful than others. This finding opens the door to a possible categorization of task difficulty. Understanding what makes tasks difficult in this way would allow roboticists to apply nonverbal behaviors only when needed, reducing the amount of nonverbal behaviors performed extraneously while improving human-robot interactions when nonverbal behaviors are warranted.

10

Discussion*

This dissertation describes a body of work seeking to understand and improve upon the use of eye gaze and other nonverbal behaviors in socially assistive human-robot interactions. The studies and models described in Chapters 3 through 9 all revolve around how robots can use nonverbal communication to improve human partners' comprehension and task performance in natural, subtle ways. In this chapter, we highlight central themes that run throughout this work, and present some open questions for future research.

10.1 Central Themes

This dissertation covers a range of approaches to HRI research and application domains. In the previous chapters, we presented lab-based studies analyzing human responses to robot gaze measuring varied effects on perception and decision making; we developed both data-driven and heuristic models of nonverbal behavior; and we covered evaluation domains ranging from tutoring to collaborative manipulation.

Despite the variety, the studies and models in this dissertation all seek to answer

*Part of this chapter is in submission [6]

one central question: *How can eye gaze and other nonverbal behaviors be used by social robots to improve socially assistive human-robot interactions?*

In the process of answering this question, several themes emerged from this work. Here, we draw out and expand on these themes.

10.1.1 Social Behavior and Nonverbal Communication

The studies described in this dissertation reveal that nonverbal behavior from a robot is not necessarily understood by human viewers as informative or communicative. For example, in Chapter 6, a robot's referential gazes (its sorting suggestions) were generally ignored by users in the absence of any other nonverbal cue. Similarly, in Chapter 3, an analysis of micro-level behaviors showed that robot directional gaze did not elicit a reflexive attention shift the way human gaze did.

However, there may be ways to influence a viewer's interpretation of a robot's nonverbal behavior. For example, in Chapter 6, adding a nonverbal gesture (a handover delay), which was perceived as highly social, caused people to look at and comply with the robot's referential gaze significantly more frequently.

Other research has also found that the perception of a robot's gaze as socially communicative can be influenced by whether that robot has previously displayed social behavior. Infants who observed a robot engage in a socially communicative exchange with an adult were more likely to subsequently follow the robot's gaze than infants who did not see a social exchange [165]. This suggests that infants can view robots as entities whose gaze is meaningful, but only when they have prior experience to indicate that the robot is a social agent.

Adults also show a difference in gaze processing depending on their expectations. Studies modeled after the same reflexive cueing experiment as our study in Chapter 3 showed that the way a stimulus is presented affects whether it evokes the reflexive cueing effect that human faces do. When presented as a social stimulus (a face), the

image cued reflexive attention shifts and activated the superior temporal sulcus, a brain area that specializes in processing eyes [138, 203]. However, the same image presented as a non-social image (a car) fails to cue these reflexive responses and did not elicit as much activation in the superior temporal sulcus. These results suggest that cognitive processing of a stimulus as a face depends on prior expectations and context.

The social signals that serve as a requirement for nonverbal communication may succeed because they set up an expectation of the robot as an animate agent, whose gaze and other behaviors are intentional and meaningful. If this is the case, the communicative effect of a robot's nonverbal behavior could be mediated by people's perceptions of the robot's animacy.

Establishing robot animacy is still a challenge in HRI. Many behaviors can influence people's perceptions of a robot's animacy, from low-level motion patterns [94] to longer time scale behaviors like cheating [221]. However, researchers still don't know what features of a robot's behavior elicit these perceptions of animacy. Understanding how animacy relates to the interpretation of a robot as a social agent may improve the way nonverbal behaviors are used in HRI.

10.1.2 Varying Levels of Analysis

One important aspect of this dissertation research is that we analyze human behavior at a variety of levels on the spectrum from micro-scale to macro-scale. Micro-scale behaviors, such as eye saccades and reflexes, occur rapidly (within hundreds of milliseconds) and often over small distances (such as tiny shifts of gaze). These behaviors may require specialized tools for measurement, such as eye trackers. Chapter 3 involves measuring micro-scale responses to directional cues.

Macro-scale behaviors are measured over larger times and distances, for instance, how people elect to sort blocks the robot has given them, as in Chapter 6. Macro-

scale behaviors can generally be observed without specialized equipment, and may involve a holistic analysis of combined actions over an extended duration of minutes or even hours.

Between these ends of the spectrum are behaviors like identifying a target (Chapter 4) or selecting an object (Chapter 5), which occur over seconds rather than an entire interaction, but can be seen and measured at a fairly coarse level of analysis.

As described in Section 2.4 and Chapter 3, people's responses to robot gaze differ from their responses to human gaze in some ways when measured on a micro time scale. For instance, robot directional gaze does not cue the same reflexive response as human directional gaze in the first 500 milliseconds of exposure [2]. In the moments just before naming an object, people spend more time ensuring joint attention by looking at their partner's face than at the object if their partner is a robot, but more time looking at the object than at their partner's face if their partner is another human [269].

Conversely, there is much evidence that robot gaze has macro-scale behavioral effects that follow expected patterns from human-human interactions. For instance, robot gaze modeled after human behavior can successfully convey object references [3] and manage conversational turn-taking [24].

Presently, the bulk of research on the effects of robot gaze tends toward the macro-scale side of the spectrum rather than the micro-scale side. One reason is that measuring micro-scale effects requires carefully controlled environments and precise sensors, while macro-scale effects can be measured in more naturalistic settings with common tools like video cameras.

Further investigations into the differences between micro-scale and macro-scale effects of robot gaze is warranted. Understanding the disparity between people's responses at these various levels of analysis would reveal as much about human cognitive processing as it would about how to design effective robots for social interactions.

10.2 Open Research Questions

This section presents several open questions in the research on social eye gaze for human-robot interaction and discusses how each might be investigated. Some of these questions—such as the importance of a robot's gaze capabilities (Section 10.2.1)—have been investigated by researchers in HRI but require further exploration. Other questions have been addressed in different contexts, but minimally explored with respect to eye gaze, like the effect of attributions of agency (Section 10.2.2) and embodiment (Section 10.2.3) on human-robot interactions. Finally, one question addresses the scope of nonverbal behaviors in the broader field of HRI (Section 10.2.4).

10.2.1 What is the role of physical capability in eye gaze for HRI?

As discussed in Section 2.2.1, research on gaze in HRI is conducted on robots with a range of physical capabilities. These capabilities, which replicate the subtle effects of human gaze—such as pupil dilation, saccades, and expressive secondary features like eyebrows—can provide additional social cues during interaction, but they are difficult to implement on physical systems.

The role of these capabilities has not yet been fully characterized. For example, many of the robots currently used for HRI research (such as the Nao) have fixed eyes, and must move their entire head to indicate gaze shifts. Robots like Keepon take this restriction a step further, requiring entire body shifts to indicate gaze direction. (Both Nao and Keepon can be seen in Figure 2.1.) But head movements might be insufficient to communicate more subtle or rapid gazes. There is some evidence that head pose estimation from an RGB-D camera is an unreliable indicator of human gaze direction [136], likely because people orient to lateral visual targets through a combination of saccades and head turns [87]. It is not clear what information is lost

when robots do not mimic this biological capability.

Mapping gaze behavior from virtual agents, which have nearly unlimited capabilities, to physical robots, which are constrained by hardware, is not trivial [206]. Understanding the effect of each capability will allow researchers to avoid over-generalizing their findings from virtual agents to embodied robot interactions. It will also enable robot designers to selectively implement hardware capabilities for specific effects, minimizing robot costs and complexity.

10.2.2 What underlies the difference in micro- and macro-scale responses to robot gaze?

Researchers do not yet understand the mechanisms that underlie human response to social robot eye gaze, and specifically, why the differences between micro- and macro-scale responses to robot gaze (discussed in Section 10.1.2) emerge. These differences may be artifacts of the experiments, they may arise from people's expectations of the robot, or they may have some other cause. Investigating the source of these differences would enable researchers to develop robot gaze that has a specific, targeted effect on human behavior.

Environmental cues may play a role in how people respond to robot eye gaze. Micro-scale experiments are often well controlled because of the precision required to measure small changes of behavior. As described in section Section 2.2.3, these kinds of studies have the disadvantage of reduced ecological validity. It may be that these artificial settings affect people's natural responses to robots by reducing how much importance is attributed to their eye gaze.

Another possibility is that there exists a difference in automatic versus conscious processing of robots. People may not automatically attribute importance to robot eye gaze (as measured on a micro level, where responses tend to be reflexive), but context and their own expectations lead them to treat robot gaze as a meaningful stimulus

when it is consciously processed (which can be seen in macro-scale measurements). Most studies in this review operate on the macro level, and find that robot eye gaze has an effect on human behavior. In contrast, the studies described in Section 2.4.2 focus on either micro-scale measurements or on investigations of infant behavior, both of which involve human responses performed rapidly and with little high-level cognitive control. On these levels, people seem to respond to robot gaze as though it has no social significance.

As described in Section 10.1.2, this automatic processing may be manipulable by changing expectations and context. It would be informative to evaluate whether the way an experimenter presented the robot affected how people processed that robot's eye gaze. Specifically, could people's response to a robot's eye gaze be changed just by whether the experimenter indicated that the robot was a social agent? Or would the perception of agency need to be established through experience with the robot?

10.2.3 Under what conditions is embodiment important for the success of a robot's gaze behavior?

Gaze in HRI has been explored using both virtual agents and physically embodied robots (Section 2.2.1). By virtue of being animated, virtual agents provide hypothetically unlimited realism in their gaze behaviors. However, as described in Section 2.2.2, interactions with physically embodied robots may lead to different human performance than interactions with virtual robots or videos of robots. Physically embodied robots have been shown to increase cognitive learning gains [153] and compliance with robot instructions [31], though this effect does not hold in all studies [135, 196].

Though researchers have investigated the effects of physical embodiment in cognitive tasks, there is little research on whether embodiment influences the effect of gaze in human-robot interactions. We do not yet know under what conditions, if any, physical robot embodiment influences the processing of robot eye gaze. Questions

(a) Baxter (b) IROMEC robot [84] (c) Chester [252]

Figure 10.1: Robots with physical bodies but animated eyes provide an interesting edge case when exploring the effects of embodiment on eye gaze in HRI (Section 10.2.3).

regarding physical embodiment for gaze in HRI include: is there a difference in the emotional expressivity of virtual and physical eyes? Do people feel attention from virtual eyes as they do from physically embodied eyes? Can embodied eyes communicate a robot's internal states through subtle cues as effectively as virtual eyes?

Robots with physical embodiment but animated eyes, such as Baxter, the IROMEC robot [84], and Chester [252], present an interesting test case for the effect of embodiment on eye gaze (Figure 10.1). These robots may help separate the effects of physical eyes, as opposed to physical bodies, when examining eye gaze in human-robot interactions.

One concern about virtual eyes is that flat, two-dimensional displays sometimes create a powerful illusion, commonly called the "Mona Lisa effect," that the eyes on the display are following the viewer regardless of viewing angle [12]. Because gaze direction is an important indicator of attention and spatial references, the Mona Lisa effect limits the ability for such systems to communicate. One approach to dispel this effect is to back-project a virtual face onto a contoured three-dimensional surface [11, 72, 146]. Back-projected technology provides the flexibility of animated eyes with the appearance of a more embodied system.

10.2.4 What domains within HRI can benefit from robot nonverbal communication?

Social gaze and other robot nonverbal behavior will become increasingly useful as it moves out of the lab and into natural human environments like homes and schools. Real-world human-robot interactions involve the complex interplay of many processes, from perception to planning to language and motion generation. Eye gaze and gestures are subtle, continuous cues that can be used to augment and support these other processes during a human-robot interaction.

The future applications of social gaze and gesture for robotics will depend on interdisciplinary research that allows real time perception and processing in these dynamic, unpredictable real-world environments. This technology requires the incorporation of work from HRI with other fields for real-time perception and information processing, such as computer vision and natural language processing.

Natural Language Processing

Tellex *et al.* have developed a model for understanding natural language commands to robots performing navigation and manipulation [240, 241]. The model first grounds the components of the natural language command to specific objects, locations, or actions in the environment. This grounding operates exclusively on verbal inputs. Incorporating referential eye gaze into the grounding model would potentially increase the confidence of symbol groundings by providing additional, multimodal command input.

Referential nonverbal behavior like eye gaze and pointing could disambiguate between two similar objects. For example, if there are two available groundings for the word "truck," looking at the intended reference during the command "put the pallet on the truck" clarifies the reference without needing additional referential speech such as "the one on the left." This can potentially increase efficiency by requiring less

verbal expressiveness from the user and less language processing from the system.

Knowledge about human nonverbal communication could also increase the speed (and thereby the efficiency) of the interaction. Because people naturally fixate on objects about one second before they verbally reference them (see Chapter 2), gaze could be used for pre-processing, allowing the system to eliminate some potential groundings before the whole command is even received.

Learning from Demonstration

Learning from demonstration (LfD) is an approach to robot learning in which the robot develops a policy for how to complete a task by watching demonstrations of that task being performed [29]. LfD has been used widely in numerous robotics domains [26].

Some researchers have already explored the effects of eye gaze in LfD (see Chapter 2.5.3). They have found that robot eye gaze acts as a feedback system for human teachers, revealing the robot's knowledge and focus of attention. This subtle but natural feedback mechanism leads to teaching that has fewer errors and less repetition of material [117], leading to more effective LfD interactions. Further research can explore how to best apply nonverbal behavior during LfD, including when and how the robot should use gaze, gesture, and other behaviors to indicate its mental state.

Legibility and Predictability of Motion

When collaborating with a robot, it is important that a robot's motion clearly reflect its intentions and future action. Legibility and predictability of a robot's motion trajectories can be mathematically defined [79]. The equations for legibility and predictability model the user's inferences between motion trajectories and goal locations.

People use gaze behavior to perform similar inferences about where a collaborator will reach. As discussed in Chapter 2.5.3 and shown in Chapter 5, people can recognize

and respond to eye gaze that indicates spatial references, successfully predicting the target of their partner's reference [4, 44]. Such expressive nonverbal behavior reveals mental states, improving task performance [47].

Incorporating eye gaze into equations for predictability and legibility would allow robots to take advantage of this natural, subtle, communicative behavior from people. Combining eye gaze with motion trajectories would generate multimodal robot behavior that is even more communicative than motion trajectories alone.

10.3 Summary

This dissertation presented a variety of studies and models, unified under the goal of understanding and developing nonverbal behaviors for socially assistive robots. Two major themes emerged from this work. The first is the importance of social behavior and expectations of animacy in order for nonverbal communication to be interpreted as meaningful. The second is the need to analyze human responses to robot behavior at a variety of scales, from micro-level to macro-level, to attain a complete picture of how nonverbal behaviors affect human-robot interactions. The open questions in this section outline continuing areas of research and highlight the broad impact that nonverbal behavior can have on the field of HRI.

11

Conclusion

Nonverbal communication is an important part of typical human-human interactions. In this dissertation, we investigated how to leverage that subtle, natural channel of communication for social robots that interact with people in domains like tutoring and collaborative manufacturing. The main contributions of this dissertation are a set of models for understanding and generating nonverbal behaviors for socially assistive human-robot interactions. Additionally, this dissertation contributed four novel studies that analyzed distinct aspects of nonverbal behavior in human-robot interactions.

The dissertation began with these laboratory-based HRI experiments that analyzed specific components of eye gaze and gesture in well-controlled interactions. These experiments analyzed human responses to robot behavior at various levels, from micro scale, millisecond-level measurements of response times (Chapter 3) to increasingly more macro scale measurements of attention recognition (Chapter 4), object reference recognition (Chapter 5), and behavioral compliance (Chapter 6).

These human-robot interaction studies contribute to nonverbal behavior modeling by providing key insights about the conditions in which people respond—or don't—to gaze and other nonverbal behaviors from a robot. The experiment in Chapter

3, modeled after a well-studied psychophysical task, shows that robot faces are not necessarily cognitively processed like human faces. The work in Chapter 4 suggests that to best convey a sense of attention, a robot should use short glances rather than long, infrequent stares. Nonverbal behaviors can augment speech without hindering performance when the two modes of communication are in conflict (Chapter 5), providing support for the use of gaze to augment spoken references. Sometimes, however, the communicative effect of one nonverbal channel (like eye gaze) is dependent on a second nonverbal channel (such as gesture), as seen in Chapter 6.

The dissertation applied these insights to modeling multimodal behavior with the dual goal of understanding human behavior and generating better robot behavior. We started by constructing a data-driven model of nonverbal behavior from a human-human tutoring interaction (Chapter 7). This model, which could both recognize the context of nonverbal behavior and suggest an appropriate nonverbal behavior to match a particular context, provided a high-level view of the nonverbal behaviors that people use in collaboration. The focus then narrowed to a single type of communication—object references—and we developed a model to generate appropriate referential behavior that is flexible enough to be used in a variety of scenes and with a variety of robot capabilities (Chapter 8). We evaluated this model in a naturalistic human-robot collaboration (Chapter 9), which revealed that nonverbal behavior is more effective as the task difficulty increases.

Bibliography

[1] *A Roadmap for U.S. Robotics: From Internet to Robotics*. Robotics Virtual Organization, 2013.

[2] Henny Admoni, Caroline Bank, Joshua Tan, and Mariya Toneva. Robot gaze does not reflexively cue human attention. In L. Carlson, C. Hölscher, and T. Shipley, editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci)*, pages 1983–1988, Austin, TX USA, 2011. Cognitive Science Society.

[3] Henny Admoni, Christopher Datsikas, and Brian Scassellati. Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci)*, pages 104–109, 2014.

[4] Henny Admoni, Anca Dragan, Siddhartha Srinivasa, and Brian Scassellati. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 49–56, 2014.

[5] Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. Are you looking at me? Perception of robot attention is mediated by gaze type and group size. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 389–396, 2013.

[6] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: A review. *Journal of Human Robot Interaction*. In submission.

[7] Henny Admoni and Brian Scassellati. Robot gaze is different from human gaze: Evidence that robot gaze does not cue reflexive attention. In *Proceedings of the "Gaze in Human-Robot Interaction" Workshop at HRI 2012*, Boston, MA USA, 2012.

[8] Henny Admoni and Brian Scassellati. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proceedings of the 6th ACM International Conference on Multimodal Interaction (ICMI)*, Istanbul, Turkey, 2014.

[9] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. Robot nonverbal behavior improves task performance in difficult collaborations. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[10] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[11] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems*, volume 7403 of *Lecture Notes in Computer Science*, pages 114–130. Springer Berlin Heidelberg, 2012.

[12] Samer Al Moubayed, Jens Edlund, and Jonas Beskow. Taming Mona Lisa: Communicating gaze faithfully in 2d and 3d facial projections. *ACM Transactions on Interactive Intelligent Systems*, 1(2), January 2012.

[13] Samer Al Moubayed and Gabriel Skantze. Perception of gaze direction for situated interaction. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction (Gaze-In)*, pages 1–6, 2012.

[14] Alaska Research Group. Spatial ability test. `http://psychometrics.akresgr.org/spatialtest/`. Accessed: 2015-09-01.

[15] Aldebaran. `https://www.aldebaran.com/en`, July 2015.

[16] Jacopo Aleotti, Vincenzo Micelli, and Stefano Caselli. Comfortable robot to human object hand-over. In *21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 771–776, 2012.

[17] Martha W. Alibali and Mitchell J. Nathan. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. *Video Research in the Learning Sciences*, pages 349–365, 2007.

[18] John R Anderson, Michael Matessa, and Christian Lebiere. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, 12(4):439–462, 1997.

[19] Sean Andrist, Wesley Collier, Michael Gleicher, Bilge Mutlu, and David Shaffer. Look Together: Analyzing Gaze Coordination with Epistemic Network Analysis. *Frontiers in Psychology*, 6(1016), 2015.

[20] Sean Andrist, Bilge Mutlu, and Michael Gleicher. Conversational Gaze Aversion for Virtual Agents. In Eds. R. Aylett, B. Krenn, C. Pelachaud, H. Shimodaira, editor, *Intelligent Virtual Agents*, volume LNCS 8108, pages 249–262, 2013.

[21] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look Like Me: Matching Robot Personality via Gaze to Increase Motivation. In *Proceedings of the ACM Annual*

Conference on Human Factors in Computing Systems (CHI), Seoul, Republic of Korea, April 2015. ACM.

[22] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. A head-eye coordination model for animating gaze shifts of virtual characters. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, Santa Monica, California, 2012. ACM Press.

[23] Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. Designing Effective Gaze Mechanisms for Virtual Agents. In *Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI)*, pages 705–714, Austin, Texas, 2012. ACM Press.

[24] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 10th International Conference on Human-Robot Interaction (HRI)*. ACM, 2014.

[25] Arduino. `http://arduino.cc`, September 2012.

[26] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[27] Michael Argyle. Non-verbal communication in human social interaction. In R. A. Hinde, editor, *Non-verbal communication*. Cambirdge University Press, Oxford, England, 1972.

[28] Michael Argyle and Mark Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Oxford, England, 1976.

[29] Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *International Conference on Machine Learining (ICML)*, 1997.

[30] Gérard Bailly, Stephan Raidt, and Frédéric Elisei. Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6):598–612, June 2010.

[31] Wilma A. Bainbridge, Justin W. Hart, Elizabeth S. Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3:41–52, 2011.

[32] Adrian Bangerter. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419, June 2004.

[33] Simon Baron-Cohen. *Mindblindness: An essay on Autism and Theory of Mind*. MIT Press, Cambridge, MA, 1995.

[34] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1:71–81, 2009.

[35] Anna Batki, Simon Baron-Cohen, Sally Wheelwright, Jennifer Connellan, and Jag Ahluwalia. Is there an innate gaze module? Evidence from human neonates. *Infant Behavior and Development*, 23:223–229, 2000.

[36] Andrew P. Bayliss, Giuseppe di Pellegrino, and Steven P. Tipper. Sex differences in eye gaze and symbolic cueing of attention. *The Quarterly Journal of Experimental Psychology*, 58A(4):631–650, 2005.

[37] E.T. Bekele, Uttama Lahiri, AR. Swanson, J.A Crittendon, Z.E. Warren, and Nilanjan Sarkar. A Step Towards Developing Adaptive Robot-Mediated Intervention Architecture (ARIA) for Children With Autism. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(2):289–299, 2013.

[38] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Towards a Humanoid Museum Guide Robot that Interacts with Multiple Persons. In *Proceedings of the 5th IEEE/RAS International Conference on Humanoid Robots 2005*, pages 418–423. IEEE, December 2005.

[39] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. Towards a humanoid museum guide robot that interacts with multiple persons. In *Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robotics*, pages 418–423, 2005.

[40] Kirsten Bergmann and Stefan Kopp. Modeling the production of coverbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*, 24:530–551, 2010.

[41] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *IEEE International Conference on Robotics and Automation (ICRA '00)*, volume 1, pages 348–353, 2000.

[42] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, page 1, 2010.

[43] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, January 2013.

[44] Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerrard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. I Reach Faster When I See You Look: Gaze Effects in Human-Human and

Human-Robot Face-to-Face Cooperation. *Frontiers in neurorobotics*, 6(May):1–11, January 2012.

[45] Gary Bradski et al. The opencv library. *Doctor Dobbs Journal*, 25(11):120–126, 2000.

[46] Cynthia Breazeal, Guy Hoffman, and Andrea Lockerd. Teaching and working with robots as a collaboration. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 04)*, pages 1030–1037. IEEE Computer Society, 2004.

[47] Cynthia Breazeal, Cory D. Kidd, Andrea L. Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 708–713, 2005.

[48] Cynthia Breazeal and Brian Scassellati. A context-dependent attention system for a social robot. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1999.

[49] Cynthia Breazeal and Brian Scassellati. How to build robots that make friends and influence people. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '99)*, volume 2, pages 858–863, Kyongju, South Korea, 1999.

[50] Davina Bristow, Geraint Rees, and Christopher D. Frith. Social interaction modifies neural response to gaze shifts. *Social Cognitive and Affective Neuroscience*, 2:52–61, 2007.

[51] Rechele Brooks and Andrew N. Meltzoff. The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, 38(6):958–966, 2002.

[52] Frank Broz, Hagen Lehmann, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. In *Proceedings of the "Gaze in Human-Robot Interaction" Workshop at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, Boston, MA, USA, March 2012.

[53] Allison Bruce. The role of expressiveness and attention in human-robot interaction. In *2002 IEEE International Conference on Robotics and Automation*, pages 38–42, Washington, DC, USA, 2002. IEEE Press.

[54] George Butterworth and Shoji Itakura. How the eyes, head and hand serve definite reference. *British Journal of Developmental Psychology*, 18(1):25–50, 2000.

[55] M. Cakmak, S. S. Srinivasa, J. Forlizzi, and S. Kiesler. Human preferences for robot-human hand-over configurations. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '11)*, pages 1986–1993, 2011.

[56] Maya Cakmak, Siddhartha S. Srinivasa, Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Using spatial and temporal contrast for fluent robot-human hand-overs. In *6th International Conference on Human-Robot Interaction (HRI '11)*, page 489, 2011.

[57] Andrew J Calder, Andrew D Lawrence, Jill Keane, Sophie K Scott, Adrian M Owen, Ingrid Christoffels, and Andrew W Young. Reading the mind from eye gaze. *Neuropsychologia*, 40(8):1129–1138, 2002.

[58] Joseph N Cappella and Catherine Pelachaud. Rules for Responsive Robots: Using Human Interactions to Build Virtual Interactions. *Stability and Change in Relationships*, pages 325–354, 2002.

[59] Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 34(4), April 2000.

[60] Justine Cassell, Obed Torres, and Scott Prevost. Turn taking vs. Discourse Structure: How best to model multimodal conversation. In *Machine Conversations*, pages 143–154. Kluwer, 1998.

[61] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: the behavior expression animation toolkit. *Life-Like Characters*, pages 477–486, 2004.

[62] Wesley P. Chan, Chris A.C. Parker, H.F. Machiel Van der Loos, and Elizabeth A. Croft. Grip Forces and Load Forces in Handovers: Implications for Designing Human-Robot Handover Controllers. In *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013)*, pages 9–16, 2012.

[63] Katarzyna Chawarska and Frederick Shic. Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 39(12):1663–1672, 2009.

[64] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues. In *7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 293–300, Boston, MA, USA, March 2012. ACM.

[65] Jung Ju Choi, Yunkyung Kim, and Sonya S. Kwak. Have you ever lied?: the impacts of gaze avoidance on people's perception of a robot. In *Proceedings of the 8th ACM/IEEE international Conference on Human-Robot Interaction (HRI 2013)*, pages 105–106, March 2013.

[66] Herbert H. Clark. Coordinating with each other in a material world. *Discourse Studies*, 7(4):507–525, October 2005.

[67] Jacob Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[68] R Alex Colburn, Michael F Cohen, and Steven M Drucker. The Role of Eye Gaze in Avatar Mediated Conversational Interfaces. Technical report, Microsoft Research, 2000.

[69] Kerstin Dautenhahn, Chrystopher L. Nehaniv, Michael L. Walters, Ben Robins, Hatice Kose-Bagci, N. Assif Mirza, and Mike Blow. KASPAR–a minimally expressive humanoid robot for human-robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4):369–397, December 2009.

[70] P. Ravindra S De Silva, Katsunori Tadano, Azusa Saito, Stephen G. Lambacher, and Mastake Higashi. Therapeutic-assisted robot for children with autism. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3561–3567, 2009.

[71] Frederic Dehais, Emrah Akin Sisbot, Rachid Alami, and Mickael Causse. Physiological and subjective evaluation of a human-robot object hand-over task. *Applied Ergonomics*, 42:785–791, 2011.

[72] Frédéric Delaunay, Joachim de Greeff, and Tony Belpaeme. Towards retro-projected robot faces: An alternative to mechatronic and android faces. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man)*, pages 306–311, Toyama, Japan, 2009.

[73] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[74] MA Diftler, JS Mehling, ME Abdallah, NA Radford, LB Bridgewater, AM Sanders, RS Askew, DM Linn, JD Yamokoski, FA Permenter, BK Hargrave, R Platt, RT Savely, and RO Ambrose. Robonaut 2-The First Humanoid Robot in Space. In *2011 IEEE International Conference on Robotics and Automation (ICRA '11)*, pages 2178–2183, 2011.

[75] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5):377–398, May 2012.

[76] M.W. Doniec, Ganghua Sun, and B. Scassellati. Active learning of joint attention. In *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots*, pages 34–39, Dec 2006.

[77] Paul E. Downing, Chris M. Dodds, and David Bray. Why does the gaze of others direct visual attention. *Visual Cognition*, 11(1):71–79, 2004.

[78] Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013.

[79] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308, March 2013.

[80] Jon Driver, Greg Davis, Paola Ricciardelli, Polly Kidd, Emma Maxwell, and Simon Baron-Cohen. Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, 6(5):509–540, 1999.

[81] Aaron Edsinger and Charles C. Kemp. Human-Robot Interaction for Cooperative Manipulation: Handing Objects to One Another. In *16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2007)*, pages 1167–1172, 2007.

[82] NJ Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24:581–604, 2000.

[83] David Feil-Seifer and Maja J. Matarić. Defining socially assistive robotics. In *Proceedings of the 9th International IEEE Conference on Rehabilitation Robotics*, 2005.

[84] Ester Ferrari, Ben Robins, and Kerstin Dautenhahn. Therapeutic and educational objectives in robot assisted play for children with autism. In *IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man)*, pages 108–114, 2009.

[85] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4):143–166, March 2003.

[86] Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L. Hill, Jon Oberlander, and Alois Knoll. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 295–302, New York, NY, USA, 2008. ACM.

[87] Edward G Freedman and David L Sparks. Coordination of the eyes and head: movement kinematics. *Experimental brain research*, 131(1):22–32, 2000.

[88] Chris Kelland Friesen and Alan Kingstone. The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, 5(3):490–495, 1998.

[89] Chris Kelland Friesen, Chris Moore, and Alan Kingstone. Does gaze direction really trigger a reflexive shift of spatial attention? *Brain and Cognition*, 57:66–69, 2005.

[90] Chris Kelland Friesen, Jelena Ristic, and Alan Kingstone. Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2):319–329, 2004.

[91] Simone Frintrop, Erich Rome, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception*, 7(1):6:1–6:39, January 2010.

[92] Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694–724, 2007.

[93] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. Messages embedded in gaze of interface agents — impression management with agent's gaze. In *SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves (CHI)*, page 41, New York, New York, USA, 2002. ACM Press.

[94] Tao Gao, George E. Newman, and Brian J. Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59:154–179, 2009.

[95] Maia Garau, Mel Slater, Simon Bee, and MA Sasse. The impact of eye gaze on communication using humanoid avatars. In *2001 Conference on Human Factors in Computing Systems (SIGCHI 2001)*, pages 309–316, Seattle, WA USA, 2001. ACM New York, NY, USA, ACM Press.

[96] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environ-

ment. *Proceedings of the conference on Human factors in computing systems - CHI '03*, 5(1):529, 2003.

[97] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[98] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt. Interacting in time and space: Investigating human-human and human-robot joint action. In *19th International Symposium in Robot and Human Interactive Communication*, pages 252–257, 2010.

[99] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419 – 429, 1999.

[100] Zenzi M. Griffin and Kathryn Bock. What the Eyes Say About Speaking. *Psychological Science*, 11(4):274–279, July 2000.

[101] Elena Corina Grigore, Kerstin Eder, Anthony G. Pipe, Chris Melhuish, and Ute Leonards. Joint Action Understanding improves Robot-to-Human Object Handover. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, November 2013. IEEE.

[102] Erdan Gu and Norman Badler. Visual attention and eye gaze during multiparty conversations with distractions. In Jonathan Gratch, Michael Young, Ruth Aylett, Daniel Ballin, and Patrick Olivier, editors, *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 193–204. Springer Berlin Heidelberg, 2006.

[103] Jaap Ham, Raymond H. Cuijpers, and John-John Cabibihan. Combining robotic persuasive strategies: The persuasive power of a storytelling robot that

uses gazing and gestures. *International Journal of Social Robotics*, 7:479–487, 2015.

[104] Joy E. Hanna and Susan E. Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57:596–615, 2007.

[105] Yasuhiko Hato, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. Pointing to space: Modeling of deictic interaction referring to regions. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, HRI, pages 301–308, Piscataway, NJ, USA, 2010. IEEE Press.

[106] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[107] Jari K. Hietanen and Jukka M. Leppänen. Does facial expression affect attention orienting by gaze direction cues? *Journal of Experimental Psychology: Human Perception and Performance*, 29(6):1228–1243, December 2003.

[108] Jari K. Hietanen, Lauri Nummenmaa, Mikko J. Nyman, Riitta Parkkola, and Heikki Hämäläinen. Automatic attention orienting by social and symbolic cues activates different neural networks: An fMRI study. *NeuroImage*, 33:406–413, 2006.

[109] Guy Hoffman and Cynthia Breazeal. Robotic partners' bodies and minds: An embodied approach to fluid human-robot collaboration. In *Proceedings of the 2006 AAAI Workshop: Cognitive Robotics*. AAAI Press, 2006.

[110] Matthew W. Hoffman, David B. Grimes, Aaron P. Shon, and Rajesh P.N. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19:299–310, 2006.

[111] Aaron Holroyd, Charles Rich, Candace L. Sidner, and Brett Ponsler. Generating connection events for human-robot collaboration. In *20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 241–246. Ieee, July 2011.

[112] Bruce M. Hood, J. Douglas Willen, and Jon Driver. Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2), March 1998.

[113] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '12)*, pages 25–32, 2012.

[114] Chien-Ming Huang and Bilge Mutlu. Modeling and Evaluating Narrative Gestures for Humanlike Robots. In *Proceedings of Robotics: Science and Systems*, June 2013.

[115] Chien-Ming Huang and Bilge Mutlu. The Repertoire of Robot Behavior: Designing Social Behaviors to Support Human-Robot Joint Activity. *Journal of Human-Robot Interaction*, 2(2):80–102, June 2013.

[116] Chien-Ming Huang and Bilge Mutlu. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 57–64. ACM, 2014.

[117] Chien-Ming Huang and Andrea L. Thomaz. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *20th IEEE International Symposium on Robot and Human Interactive Communication (2011 RO-MAN)*, pages 65–71, Atlanta, GA USA, 2011.

[118] Markus Huber, Markus Rickert, Alois Knoll, Thomas Brandt, and Stefan Glasauer. Human-Robot Interaction in Handing-Over Tasks. In *17th IEEE In-*

ternational Symposium on Robot and Human Interactive Communication (RO-MAN 2008), pages 107–112, 2008.

[119] Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3):598–612, 1995.

[120] M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase. Robot mediated round table: Analysis of the effect of robot's gaze. In *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2002)*, pages 411–416, 2002.

[121] Carlos T. Ishi, ChaoRan Liu, Hiroshi Ishiguro, and Norihiro Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 293–300. IEEE, 2010.

[122] Akira Ito, Shunsuke Hayakawa, and Tazunori Terada. Why robots need body for mind communication-an attempt of eye-contact between human and robot. In *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, pages 473–478, Kurashiki, Okayama Japan, September 2004. IEEE.

[123] Laurent Itti, Nitin Dhavale, and Frederic Pighin. Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. In Bruno Bosacchi, David B. Fogel, and James C. Bezdek, editors, *SPIE*, volume 5200, pages 64–78, 2004.

[124] Laurent Itti, Nitin Dhavale, and Frederic Pighin. Photorealistic Attention-Based Gaze Animation. In *ICME*, pages 521–524, 2006.

[125] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12):1489–1506, 2000.

[126] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, pages 194–203, March 2001.

[127] R S Johansson, G Westling, A Bäckström, and J R Flanagan. Eye-hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917–6932, September 2001.

[128] W. Lewis Johnson, Jeff W. Rickel, and James C Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11:47–78, 2000.

[129] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*, pages 1555–1566. ACM, 2013.

[130] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. Friendly social robot that understand human's friendly relationships. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2215–2222, 2004.

[131] SH Kang, Jonathan Gratch, and Candy Sidner. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. *Proceedings of the 11th International Conference on Autonomous Agents and Mutliagent Systems (AAMAS)*, pages 4–8, 2012.

[132] Daphne E. Karreman, Geke D.S. Ludden, Elisabeth M.A.G. van Dijk, and Vanessa Evers. How can a tour guide robot influence visitors engagement, orien-

tation and group formations? In M. Salem, A. Weiss, P. Baxter, and K. Dautenhahn, editors, *4th International Symposium on New Frontiers in Human-Robot Interaction*, Canterbury, UK, April 2015.

[133] Daphne E. Karreman, Gilberto Sepúlveda Bradford, Betsy van Dijk, Manja Lohse, and Vanessa Evers. What Happens When a Robot Favors Someone ? In *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 157–158, 2013.

[134] Charles C Kemp, Aaron Edsinger, and Eduardo Torres-Jara. Challenges for robot manipulation in human environments. *IEEE Robotics and Automation Magazine*, 14(1):20, 2007.

[135] James Kennedy, Paul Baxter, and Tony Belpaeme. Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, 7(2):293–308, 2015.

[136] James Kennedy, Paul Baxter, and Tony Belpaeme. Head pose estimation is an inadequate replacement for eye gaze in child-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 35–36. ACM, 2015.

[137] Sonu Chopra Khullar and Norman I Badler. Where to Look? Automating Attending Behaviors of Virtual Human Characters. *Autonomous Agents and Multi-Agent Systems*, 4(1-2):9–23, 2001.

[138] Alan Kingstone, Christine Tipper, Jelena Ristic, and Elton Ngan. The eyes have it!: An fMRI investigation. *Brain and Cognition*, 55:269–271, 2004.

[139] Nathan Kirchner, Alen Alempijevic, and Gamini Dissanayake. Nonverbal robot-

group interaction using an imitated gaze cue. *Proceedings of the 6th international conference on Human-robot interaction (HRI)*, page 497, 2011.

[140] Chris L. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, July 1986.

[141] Heather Knight and Reid Simmons. Estimating Human Interest and Attention via Gaze Analysis. In *International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, May 2013.

[142] Jacqueline M. Kory, Sooyeon Jeong, and Cynthia L. Breazeal. Robotic learning companions for early language development. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*, pages 71–72, New York, NY, USA, 2013. ACM.

[143] Spyros Kousidis and David Schlangen. The Power of a Glance : Evaluating Embodiment and Turn-Tracking Strategies of an Active Robotic Overhearer. In *Proceedings of the 2015 AAAI Spring Symposium: Turn-Taking and Coordination in Human-Machine Interaction*, pages 36–43. AAAI Press, 2015.

[144] Hideki Kozima, Marek P. Michalowski, and Cocoro Nakagawa. Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics*, 1:3–18, 2009.

[145] Yoshinori Kuno, Kazuhisa Sadazuka, Michie Kawashima, Keiichi Yamazaki, Akiko Yamazaki, and Hideaki Kuzuoka. Museum guide robot based on sociological interaction analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, pages 1191–1194, 2007.

[146] Taakaki Kuratate, Yosuke Matsusaka, Brennand Pierce, and Gordon Cheng. "Mask-bot": A life-size robot head using talking head animation for human-

robot communication. In *IEEE-RAS International Conference on Humanoid Robots*, pages 99–104, Bled, Slovenia, 2011.

[147] Bojana Kuzmanovic, Alexandra L. Georgescu, Simon B. Eickhoff, Nadim J. Shah, Gary Bente, Gereon R. Fink, and Kai Vogeley. Duration matters: Dissociating neural correlates of detection and evaluation of social gaze. *NeuroImage*, 46(4):1154–1163, 2009.

[148] M F Land and M Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559–65, January 2001.

[149] Stephen R. H. Langton and Vicki Bruce. Reflexive visual orienting in response to the social attention of others. *Visual Cognition*, 6(5):541–567, 1999.

[150] Jina Lee, Stacy Marsella, David Traum, Jonathan Gratch, and Brent Lance. The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In C. Pelachaud, editor, *Intelligent Virtual Agents*, volume LNAI 4722, pages 296–303. Springer-Verlag, 2007.

[151] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. Are physically embodied social agents better than disembodied social agents? the effects of physical embodiment, tactile interaction, and people's loneliness in humanrobot interaction. *International Journal of Human-Computer Studies*, 64(10):962–973, October 2006.

[152] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Maya Cakmak, and Siddhartha Srinivasa. Predictability or Adaptivity? Designing Robot Handoffs Modeled from Trained Dogs and People. In *6th International Conference on Human-Robot Interaction (HRI '11)*, pages 179–180, 2011.

[153] Daniel Leyzbeg, Sam Spaulding, Mariya Toneva, and Brian Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proceed-*

ings of the 34th Annual Conference of the Cognitive Science Society (CogSci), 2012.

[154] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, pages 423–430, New York, NY, USA, 2014. ACM.

[155] Zheng Li and Xia Mao. EEMML: the emotional eye movement animation toolkit. *Multimedia Tools and Applications*, 60(1):181–201, May 2012.

[156] Zheng Li and Xia Mao. Emotional eye movement generation based on Geneva Emotion Wheel for virtual agents. *Journal of Visual Languages & Computing*, 23(5):299–310, October 2012.

[157] Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceedings of the Seventh Annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12)*, pages 285–292, Boston, MA, USA, March 2012. ACM Press.

[158] Phoebe Liu, Dylan F. Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. It's not polite to point: Generating socially-appropriate deictic behaviors towards people. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–274, Piscataway, NJ, USA, 2013. IEEE Press.

[159] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 4, pages 3475–3480, 2004.

[160] Sandra C. Lozano and Barbara Tversky. Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language*, 55:47–63, July 2006.

[161] Max Lungarella and Giogio Metta. Beyond gazing, pointing, and reaching: A survey of developmental robotics. In *EPIROB '03*, pages 81–89, 2003.

[162] C. Neil Macrae, Bruce M. Hood, Alan B. Milne, Angela C. Rowe, and Malia F. Mason. Are you looking at me? eye gaze and person perception. *Psychological Science*, 13(5):460–464, 2002.

[163] Nikolaos Mavridis. A review of verbal and non-verbal human-robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

[164] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, 1992.

[165] Andrew N. Meltzoff, Rechele Brooks, Aaron P. Shon, and Rajesh P. N. Rao. "Social" robots are psychological agents for infants: a test of gaze following. *Neural Networks*, 23:966–972, 2010.

[166] Marek Michalowski, Kyle Machulis, and Mark Gasson. BeatBots MyKeepon GitHub Repository. `https://github.com/beatbots/MyKeepon`, July 2013.

[167] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 101(2):343–352, 1956.

[168] AJung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K. X. J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. Meet me where i'm gazing: How shared attention gaze affects human-robot

handover timing. In *Proceedings of the 9th ACM/IEEE Interational Conference on Human-Robot Interaction (HRI)*, pages 334–341, Bielefeld, Germany, 2014. ACM.

[169] Chris Moore and Phil Dunham. *Joint attention: Its origins and role in development*. Psychology Press, 2014.

[170] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Lausanne, Switzerland, March 2011.

[171] Robin Murphy, Jessica Gonzales, and Vasant Srinivasan. Inferring social gaze from conversational structure and timing. *Proceedings of the 6th international conference on Human-robot interaction (HRI '11)*, pages 209–210, 2011.

[172] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids '06)*, pages 518–523, 2006.

[173] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. Conversational gaze mechansims for humanlike robots. *ACM Transactions on Interactive Intelligent Systems*, 1(2), January 2012.

[174] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Human Robot Interactions (HRI'09)*, pages 61–68, La Jolla, California, USA, March 2009. ACM.

[175] Bilge Mutlu, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Nonverbal leakage in robots: Communication of intentions through

seemingly unintentional behavior. In *Human Robot Interactions (HRI'09)*, pages 69–76, La Jolla, California, March 2009. ACM.

[176] Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, December 2003.

[177] Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[178] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):5:1–5:24, March 2008.

[179] V. Ng-Thow-Hing, Pengcheng Luo, and S. Okita. Synchronized gesture and speech production for humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4617–4624, Oct 2010.

[180] Hai Nguyen, Cressel Anderson, Alexander Trevor, Advait Jain, Zhe Xu, and Charles C. Kemp. El-E: An assistive robot that fetches objects from flat surfaces. In *Robotic Helpers Workshop at HRI '08*, 2008.

[181] Kai Nickel and Rainer Stiefelhagen. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing*, 25(12):1875–1884, 2007.

[182] Aline Normoyle, Jeremy B. Badler, Teresa Fan, Norman I. Badler, Vinicius J. Cassol, and Soraia R. Musse. Evaluating perceived trust from procedurally animated gaze. In *Proceedings of Motion on Games*, MIG '13, pages 119:141–119:148, New York, NY, USA, 2013. ACM.

[183] Catharine Oertel, Marcin Włodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. Gaze patterns in turn-taking. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, volume 3, pages 2243–2246, Portland, OR, September 2012.

[184] Yuko Okumura, Yasuhiro Kanakogi, Takayuki Kanda, Hiroshi Ishiguro, and Shoji Itakura. *Cognition*, 128(2):127–33, August 2013.

[185] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Providing route directions: Design of robot's utterance, gesture, and timing. In *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 53–60, March 2009.

[186] Kazuhiro Otsuka, Yoshinao Takemae, and Junji Yamato. A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th International Conference on Multimodal Interfaces*, ICMI '05, pages 191–198, New York, NY, USA, 2005. ACM.

[187] Amit Kumar Pandey, Muhammad Ali, and Rachid Alami. Towards a Task-Aware Proactive Sociable Robot Based on Multi-state Perspective-Taking. *International Journal of Social Robotics*, 5(2):215–236, 2013.

[188] Catherine Pelachaud and Massimo Bilvi. Modelling gaze behavior for conversational agents. In *Intelligent Virtual Agents (IVA)*, volume LNAI 2792, pages 93–100. Springer-Verlag, 2003.

[189] Kevin A Pelphrey, Jeffrey D Singerman, Truett Allison, and Gregory McCarthy. Brain activation evoked by perception of gaze shifts: the influence of context. *Neuropsychologia*, 41(2):156 –170, 2003.

[190] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. A model of attention and interest using gaze behavior. In Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist, editors, *Intelligent Virtual Agents*, volume 3661 of *Lecture Notes in Computer Science*, pages 229–240. Springer Berlin Heidelberg, 2005.

[191] Ulrich J Pfeiffer, Bert Timmermans, Gary Bente, Kai Vogeley, and Leonhard Schilbach. A non-verbal Turing test: differentiating mind from machine in gaze-based social interaction. *PloS one*, 6(11):e27591, January 2011.

[192] Nadine Pfeiffer-Lessmann, Thies Pfeiffer, and Ipke Wachsmuth. An operational model of joint attention - timing of gaze patterns in interactions between humans and a virtual human. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, pages 851–856, Austin, TX USA, 2012. Cognitive Science Society.

[193] Ac Pierno, C Becchio, Mb Wall, At Smith, L Turella, and U Castiello. When gaze turns into grasp. *Journal of Cognitive Neuroscience*, 18(12):2130–2137, 2006.

[194] Karola Pitsch, Anna-Lisa Vollmer, and Manuel Mühlig. Robot feedback shapes the tutor's presentation: How a robot's online gaze strategies lead to micro-adaptation of the human's conduct. *Interaction Studies*, 14(2):268–296, 2013.

[195] Isabella Poggi, Catherine Pelachaud, and Fiorella De Rosis. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169–181, 2000.

[196] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. Comparing a

Computer Agent with a Humanoid Robot. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 145–152, 2007.

[197] Aditi Ramachandran, Alexandru Litoiu, and Brian Scassellati. Shaping productive help-seeking behavior during robot-child tutoring interactions. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016.

[198] Nathan Ratliff, Matt Zucker, J. Andrew (Drew) Bagnell, and Siddhartha Srinivasa. CHOMP: Gradient optimization techniques for efficient motion planning. *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[199] Paola Ricciardelli, Emanuela Bricolo, Salvatore M. Aglioti, and Leonardo Chelazzi. My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. *NeuroReport*, 13(17):2259–2264, December 2002.

[200] Charles Rich and Brett Ponsler. Recognizing engagement in human-robot interaction. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI '10)*, pages 375–382, 2010.

[201] Laurel D. Riek, Tal-Chen Rabinowitch, Paul Bremner, Anthony G. Pipe, Mike Fraser, and Peter Robinson. Cooperative gestures: Effective signaling for humanoid robots. In *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 61–68, March 2010.

[202] Jelena Ristic and Alan Kingstone. Taking control of reflexive social attention. *Cognition*, 94:B55–B65, 2005.

[203] Jelena Ristic, Laurent Mottron, Chris Kelland Friesen, Grace Iarocci, Jacob A. Burack, and Alan Kingstone. Eyes are special but not for everyone: The case of autism. *Cognitive Brain Research*, 24:715–718, 2005. Short communication.

[204] Wolff-Michael Roth. Gestures: Their role in teaching and learning. *Review of Educational Research*, 71(3):365–392, 2001.

[205] Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hörnstein, José Santos-Victor, and Rolf Pfeifer. Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 962–967, May 2008.

[206] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum*, pages 1–28, 2015.

[207] Martin Saerbeck, Tom Schut, Christoph Bartneck, and Maddy D Janse. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the 28th international conference on Human factors in computing systems (CHI)*, pages 1613–1622, 2010.

[208] Daisuke Sakamoto, Takayuki Kanda, Tetsuo Ono, Hiroshi Ishiguro, and Norihiro Hagita. Android as a telecommunication medium with a human-like presence. In *2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 193–200. IEEE, 2007.

[209] K. Sakita, K. Ogawara, S. Murakami, K. Kawamura, and K. Ikeuchi. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 1, 2004.

[210] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[211] Satoru Satake, Takayuki Kanda, Dylan F. Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to Approach Humans?-Strategies for Social Robots to Initiate Interaction. *Journal of the Robotics Society of Japan*, 28(3):327–337, 2010.

[212] Allison Sauppé and Bilge Mutlu. Robot Deictics: How Gesture and Context Shape Referential Communication. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI)*, pages 342–349, 2014.

[213] Brian Scassellati. Mechanisms of shared attention for a humanoid robot. In *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, volume 4, page 21, 1996.

[214] Brian Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for metaphors, analogy, and agents*, pages 176–195. Springer, 1999.

[215] Brian Scassellati. How social robots will help us to diagnose, treat, and understand autism. In Sebastian Thrun, Rodney Brooks, and Hugh Durrant-Whyte, editors, *Robotics Research*, volume 28 of *Springer Tracts in Advanced Robotics*, pages 552–563. Springer Berlin / Heidelberg, 2007.

[216] Brian Scassellati, Henny Admoni, and Maja Matarić. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294, 2012.

[217] Boris Schauerte and Gernot A. Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pages 6:1–6:8, New York, NY, USA, 2010. ACM.

[218] Atsushi Senju, Yoshikuni Tojo, Hitoshi Dairoku, and Toshikazu Hasegawa. Reflexive orienting in response to eye gaze and an arrow in children with and without autism. *Journal of Child Psychology and Psychiatry*, 45(3):445–458, March 2004.

[219] Julie Shah and Cynthia Breazeal. An empirical analysis of team coordination behaviors and action planning with application to human-robot teaming. *Human factors*, 52(2):234–245, 2010.

[220] Frederic Shic, Brian Scassellati, David Lin, and Katarzyna Chawarska. Measuring context: The gaze patterns of children with autism evaluated from the bottom-up. In *IEEE 6th International Conference on Development and Learning (ICDL)*, pages 70–75, July 2007.

[221] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. No fair!! an interaction with a cheating robot. In *5th ACM/IEEE International Conference on Human-Robot Interaction*, pages 219–226, 2010.

[222] Elaine Short, Katelyn Swift-Spong, Jillian Greczek, Aditi Ramachandran, Alexandru Litoiu, Elena Corina Grigore, David Feil-Seifer, Samuel Shuster, Jin Joo Lee, Shaobo Huang, Svetlana Levonisova, Sarah Litz, Jamy Li, Gisele Ragusa, Donna Spruijt-Metz, Maja Mataric, and Brian Scassellati. How to train your dragonbot: Socially assistive robots for teaching children about nutrition through play. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, August 2014.

[223] Candace L. Sidner, Cory D. Kidd, Chrisopher Lee, and Neal Lesh. Where to look: A study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI '04)*, pages 78–84, New York, NY, USA, 2004. ACM.

[224] Emrah Akin Sisbot and Rachid Alami. A Human-Aware Manipulation Planner. *IEEE Transactions on Robotics*, 28(5):1045–1057, October 2012.

[225] Mihaela Sorostinean, Francois Ferland, Thi-Hai-Ha Dang, and Adriana Tapus. Motion-oriented attention for a social gaze robot behavior. In *Proceedings of the International Conference on Social Robotics (ICSR)*, pages 310–319. Springer, 2014. LNAI 8755.

[226] T.P. Spexard, M.. Hanheide, and G.. Sagerer. Human-Oriented Interaction With an Anthropomorphic Robot. *IEEE Transactions on Robotics*, 23(5):852–862, October 2007.

[227] Siddhartha S. Srinivasa, Dmitry Berenson, Maya Cakmak, Alvaro Collet, Mehmet R. Dogar, Anca D. Dragan, Ross A. Knepper, Tim Niemueller, Kyle Strabala, Mike Vande Weghe, and Julius Ziegler. HERB 2.0 : Lessons Learned From Developing a Mobile Manipulator for the Home. *Proceedings of the IEEE*, 100(8):2410–2428, 2012.

[228] Vasant Srinivasan. *High Social Acceptance of Head Gaze Loosely Synchronized With Speech for Social Robots*. PhD thesis, Texas A & M University, 2014.

[229] Vasant Srinivasan and Robin R Murphy. A Survey of Social Gaze. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 253–254, 2011.

[230] Aaron St. Clair, Ross Mead, and M. J. Matarić. Investigating the effects of visual saliency on deictic gesture production by a humanoid robot. In *IEEE RO-MAN*, pages 210–216, July 2011.

[231] Maria Staudte and Matthew W Crocker. Visual Attention in Spoken Human-Robot Interaction. In *Proceedings of the 4th ACM/IEEE International Confer-*

ence on Human-Robot Interaction (HRI 09), pages 77–84, La Jolla, California, USA, March 2009. Saarland University, ACM Press.

[232] Maria Staudte and Matthew W Crocker. Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, 120:268–291, August 2011.

[233] Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of ACM SIGGRAPH*, pages 506–513, New York, NY, USA, 2004. ACM.

[234] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S. Srinivasa. Learning the communication of intent prior to physical collaboration. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 968–973. Ieee, September 2012.

[235] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, Siddhartha S. Srinivasa, Maya Cakmak, and Vincenzo Micelli. Toward Seamless Human-Robot Handovers. *Journal of Human-Robot Interaction*, 2(1):112–132, 2013.

[236] Dan Szafir and Bilge Mutlu. Pay attention! Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems (CHI 2012)*, pages 11–20, Austin, TX USA, May 2012.

[237] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 69–76, Lausanne, Switzerland, March 2011. ACM Press.

[238] Adriana Tapus, Maja Matarić, and Brian Scassellati. Socially assistive robotics: The grand challenges in helping humans through social interaction. *IEEE Robotics and Automation Magazine*, pages 35–42, March 2007.

[239] Adriana Tapus, Andreea Peca, Amir Aly, Cristina Pop, Lavinia Jisa, Sebastian Pintea, Alina S Rusu, and Daniel O David. Children with autism social engagement in interaction with Nao, an imitative robot A series of single case experiments. *Interaction Studies*, 13(3):315–347, 2012.

[240] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Approaching the Symbol Grounding Problem with Probabilistic Graphical Models. *AI Magazine*, pages 64–76, 2011.

[241] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[242] Kristinn Thórisson. Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue With People. In *Proceedings of the First International Conference on Autonomous Agents (AGENTS 97)*, Marina Del Rey, CA USA, 1997. ACM.

[243] Kristinn R Thórisson. Face-to-face communication with computer agents. In *AAAI Spring Symposium on Believable Agents*, pages 86–90, Stanford University, CA, August 1994.

[244] Kristinn R Thórisson. Layered Modular Action Control for Communicative Humanoids. In *Computer Animation '97*, pages 134–143, 1997.

[245] Christine M. Tipper, Todd C. Handy, Barry Giesbrecht, and Alan Kingstone. Brain responses to biological relevance. *Journal of Cognitive Neuroscience*, 20(5):879–891, 2008.

[246] Jason Tipples. Orienting to counterpredictive gaze and arrow cues. *Perception and Psychophysics*, 70(1):77–87, 2008.

[247] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child Development*, 57(6):1454–1463, 1986.

[248] JG Trafton and MD Bugajska. Integrating Vision and Audition within a Cognitive Architecture to Track Conversations. *Proceedings of the 3rd International Conference on Human-Robot Interaction (HRI '08)*, pages 201–208, 2008.

[249] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[250] Jochen Triesch, Christof Teuscher, Gedeon O. Deák, and Eric Carlson. Gaze following: why (not) learn it? *Developmental Science*, 9(2):125–147, 2006.

[251] Jef A van Schendel and Raymond H Cuijpers. Turn-yielding cues in robot-human conversation. In K. Dautenhahn M. Salem, A. Weiss, P. Baxter, editor, *4th International Symposium on New Frontiers in Human-Robot Interaction*, Canterbury, UK, April 2015.

[252] Marynel Vázquez, Aaron Steinfeld, Scott E. Hudson, and Jodi Forlizzi. Spatial and other social engagement cues in a child-robot interaction: Effects of a sidekick. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 2014.

[253] Roel Vertegaal and Yaping Ding. Explaining effects of eye gaze on mediated group conversations: amount or synchronization? In *Proceedings of the 2002*

ACM Conference on Computer Supported Cooperative Work (CSCW), pages 41–48, New York, NY, USA, 2002. ACM.

[254] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2001)*, volume 3 of *CHI '01*, pages 301–308. ACM, Acm, 2001.

[255] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2014.

[256] Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan Schaal. Overt visual attention for a humanoid robot. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 4, pages 2332–2337, October 2001.

[257] Michael von Grünau and Christina Anston. The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, 24:1297–1313, 1995.

[258] Kazuyoshi Wada and Takanori Shibata. Living with seal robots—its sociophsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5):972–980, October 2007.

[259] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. Embodiment and human-robot interaction : A task-based perspective. In *16th IEEE International Conference on Robot & Human Interactive Communication*, pages 872–877, Jeju, Korea, August 2007. IEEE.

[260] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, 2006.

[261] Ning Wang and Jonathan Gratch. Don't just stare at me! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, pages 1241–1249, Atlanta, GA USA, April 2010. ACM.

[262] Jeremy M Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.

[263] Tian Xu, Hui Zhang, and Chen Yu. Cooperative gazing behaviors in human multi-robot interaction. *Interaction Studies*, 14(3):390–418, 2013.

[264] Akiko Yamazaki, Keiichi Yamazaki, Matthew Burdelski, Yoshinori Kuno, and Mihoko Fukushima. Coordination of verbal and non-verbal actions in human-robot interaction at museums and exhibitions. *Journal of Pragmatics*, 42:2398–2414, 2010.

[265] Keiichi Yamazaki, Michie Kawashima, Yoshinori Kuno, Naonori Akiya, Matthew Burdelski, Akiko Yamazaki, Hideaki Kuzuoka, Liam Bannon, Ina Wagner, Carl Gutwin, Richard Harper, and Kjeld Schmidt. Prior-to-request and request behaviors within elderly day care: Implications for developing service robots for use in multiparty settings. In L. Bannon, I. Wagner, C. Gutwin, R. Harper, and K. Schmidt, editors, *ECSCW*, pages 61–78, Limerick, Ireland, September 2007. Springer.

[266] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI)*, pages 140—-145, Nagoya, Japan, November 2007. ACM Press.

[267] Yuichiro Yoshikawa and Kazuhiko Shinozawa. Responsive Robot Gaze to Interaction Partner. In *Robotics: Science and Systems*, 2006.

[268] Yuichiro Yoshikawa, Kazuhiko Shinozawa, Hiroshi Ishiguro, Norihiro Hagita, and Takanori Miyamoto. The effects of responsive eye movement and blinking behavior in a communication robot. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pages 4564–4569, October 2006.

[269] Chen Yu, Paul Schermerhorn, and Matthias Scheutz. Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems*, 1(2), January 2012.

[270] Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi. Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems*, 44(2):157–168, April 2014.

[271] Minhua Zheng, AJung Moon, Elizabeth A. Croft, and Max Q.-H. Meng. Impacts of Robot Head Gaze on Robot-to-Human Handovers. *International Journal of Social Robotics*, 2015.