HUMAN-ROBOT JOINT ACTION: COORDINATING ATTENTION, COMMUNICATION, AND ACTIONS

by

Chien-Ming Huang

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 10/07/2015

The dissertation is approved by the following members of the Final Oral Committee:

Bilge Mutlu (Chair), Department of Computer Sciences, University of Wisconsin–Madison
Maya Cakmak, Department of Computer Science & Engineering, University of Washington
Mark Craven, Department of Computer Sciences, University of Wisconsin–Madison
Xiaojin Zhu, Department of Computer Sciences, University of Wisconsin–Madison
Michael Zinn, Department of Mechanical Engineering, University of Wisconsin–Madison

To my dearest family,
Huei-Lung Huang, Chu-Ying Kuo, & Chien-Chung Huang,
for their unlimited love and support.

ACKNOWLEDGMENTS

A year spent in artificial intelligence is enough to make one believe in God.

— Alan Perlis

Foremost, I want to thank my advisor, Bilge Mutlu, for his wholehearted guidance and support. He has always encouraged me to push the boundaries of my research ideas to broaden my horizons. My work has greatly benefited from his keen insights into interaction design. Over the years, he has also shaped me as a researcher in how I tackle problems, present my research, and write scientific reports. I am indebted to him for my scientific training and career development. I also want to thank my thesis committee members—Maya Cakmak, Mark Craven, Xiaojin (Jerry) Zhu, and Michael Zinn—for their thoughtful feedback and comments. Stimulating discussions with them expanded my research view and pointed toward exciting future work related to this dissertation. In particular, Maya has been a good friend and mentor in my research journey since I was pursuing my Masters degree. She has been a great help in developing and organizing my research ideas. Conversations with her have always been refreshing and encouraging. Jerry and Mark have particularly provided artificial intelligence and machine learning perspectives to strengthen my research.

I want to extend my thanks to members of the Human-Computer Interaction Laboratory at the University of Wisconsin–Madison—Sean Andrist, Chris Bodden, Steve Johnson, Faisal Khan, Shiyu Luo, Margaret Pearce, Tomislav Pejsa, Irene Rae, Allison Sauppé, Daniel Szafir, and Zhi Tan—and research friends and fellow students—Jilana Boston, Jingjing Du, Faye Golden, Brandi Hefty, Jing Jing, Erdem Kaya, Yunkyung Kim, Ross Luo, Mark Orr, Jason Power, Catherine Steffel, Danielle Albers Szafir, Xiaoyu Wang, and Kohei Yoshikawa. Their company and support made this Ph.D. journey enjoyable and unforgettable.

I would also like express my gratitude to Andrea Thomaz and Takayuki Kanda. Andrea ignited my interests in HRI by introducing me to the field through our work at the Georgia Institute of Technology. Takayuki has been a kind mentor

and encouraged me to push the frontiers of human-robot interaction in the real world during my internship at ATR, Japan. I also enjoyed discussions with Dan Bohus on situated interaction that motivate future work of this research. Finally, this dissertation research would not have been possible without support from the National Science Foundation awards 1017952 and 1149970 and an equipment loan from Mitsubishi Heavy Industries.

Last but not least, I want to express my devoted love and thanks to my family and Chun-An Sun. Their love has been a solid foundation of support in my journey to completing my dissertation. My greatest appreciation goes to the two Holy Teachers who have led me through all the ups and downs of life.

CONTENTS

Contents iv

List of Tables vi

List of Figures vii

Abstract xv

1 Introduction 1
- *1.1 Central thesis* 3
- *1.2 Scope* 4
- *1.3 Research approach* 6
- *1.4 Research platforms* 8
- *1.5 Overview of contributions* 8
- *1.6 Dissertation overview* 10

2 Background 11
- *2.1 Coordination mechanisms of joint action* 11
- *2.2 Joint action between humans and robots* 18

3 Joint Attention: Using Multimodal Behaviors to Establish Perceptual Common Ground 24
- *3.1 Introduction* 24
- *3.2 Related work* 25
- *3.3 Study 1: initiating joint attention via gaze cues* 35
- *3.4 Study 2: effects of referential gestures on information delivery* 53
- *3.5 Study 3: learning multimodal behaviors for effective communication* 70
- *3.6 Summary* 86

4 Action Observation: From Observation to Prediction to Reaction 88

4.1 Introduction 88
4.2 Related work 89
4.3 Study 4: predicting user intentions to support anticipatory actions 93
4.4 Summary 126

5 Task-sharing & Action Coordination: Adapting Actions to Achieve Fluid Collaboration 128
5.1 Introduction 128
5.2 Related work 129
5.3 Study 5: designing adaptive strategies for object handovers 130
5.4 Summary 150

6 General Discussion 152
6.1 Lessons learned 152
6.2 Limitations 162
6.3 Future research 165

7 Conclusion 172
7.1 System contributions 172
7.2 Methodological contributions 174
7.3 Empirical contributions 175
7.4 Final remarks 177

References179

LIST OF TABLES

4.1 Summary of the quantitative evaluation of the effectiveness of different intention prediction approaches. 100
4.2 Statistical test results for objective measures, correlations between prediction accuracy and objective measures, and test results for subjective measures. 118
4.3 Descriptive statistics of objective measures broken down into correct (A-Correct) and incorrect (A-Incorrect) predictions as well as correct (A-N-Correct) and incorrect (A-N-Incorrect) neighboring predictions. . 119

5.1 Confidence scores for KNN-based prediction of receiver states. 139

6.1 Objective measures for estimating the quality of human-robot joint action, their descriptions, and in which present studies they were employed. 155
6.2 Subjective scales for estimating user experience in human-robot joint action, their descriptions, and in which present studies they were employed. 156

LIST OF FIGURES

1.1 The research approach used in this dissertation leverages models of human-human joint action to enable human-robot joint action. 5
1.2 Research platforms used in this dissertation. The Wakamaru robot was used in Studies 1–3 whereas the MICO robot was used in Studies 4–5. . 9

2.1 Human-robot joint action as the processes of *expression* and *prediction*. From a robot's point of view, the process of expression is to use behavioral means to reveal internal states to the human partner. The process of prediction is to understand and predict the partner's internal states by monitoring his/her behavioral signals. 19

3.1 A visual comparison of the conventional evaluation method used to date and the proposed multivariate evaluation method. Conventional approaches follow categorical manipulations in a small number of design variables, while the proposed multivariate evaluation method involves joint manipulation of all design variables. 34
3.2 The repertoire of robot behavior consists of a *social cue repository* and a set of *activity models*, allowing a selectively use of behavior to achieve specific interaction outcomes. 36
3.3 An example behavioral specification on synchronizing referential cues in speech and gaze. 37
3.4 A summary of how the constructs of Activity Theory informed the system design of the Robot Behavior Toolkit. 39
3.5 An example activity model in which the robot instructs a human partner to clear objects on a table. 41
3.6 The Robot Behavior Toolkit consists of three subsystems—the perceptual, cognitive, and behavior systems; two memories—the working memory and long term memory; and supporting components—the activity model and knowledge base. 42

3.7 Information flow between our Toolkit and ROS. Rounded squares represent *topics*. Solid and dashed lines denote *publishing* and *subscribing* to topics, respectively. Light dashed lines denote *service* communication between nodes. 42
3.8 An example behavior output generated by the Toolkit in XML (top) and in visual representation (bottom). 44
3.9 The setup of the storytelling (top) and collaborative work (bottom) tasks in the experiment. 46
3.10 Results on information recall, collaborative work, and perceptions. (*), (**), and (***) denotes $p < .05$, $p < .01$, and, $p < .001$, respectively. . . . 49
3.11 The data collection study in Study 2. Participants were asked to teach other participants about a topic they were trained on. The goal was to understand how participants used various gestures during their teaching sessions. 54
3.12 Top lexical affiliate categories for each type of representative gesture. Percentages represent the amount by which categories of lexical affiliates co-occur with each gesture type. For example, 42.9% of the lexical affiliates for iconic gestures are "action verbs." Note that categories might overlap. 55
3.13 A schematic of temporal alignment between gestures and lexical affiliates (not to scale). For instance, iconic gestures began on average 645 ms before and ended 555 ms after their corresponding lexical affiliates. . . 56
3.14 Distributions of targets for gesture-contingent gaze for each type of gesture. The human data showed four main gaze targets: the *recipient*, the narrator's *own gesture*, the *reference*, and *other*, non-task-relevant targets. *Traveling* represents transitions between these targets. Narrators gazed most toward references while displaying deictic gestures but split their gaze evenly between the recipient and references while displaying other types of gestures. 57

3.15 Examples of the four common types of gestures—*deictic*, *beat*, *iconic*, and *metaphoric* gestures—observed in human storytellers (top) and implemented into the robot (bottom). The narrator uses deictic gestures to point toward an object of reference, beat gestures before introducing a new concept, iconic gestures to depict a concrete object such as "a flat wooden surface," and metaphoric gestures to visualize abstract concepts such as "about six hours." 58
3.16 An example utterance from the robot's narration, including the lexical affiliate "a wooden or stone surface," its corresponding iconic gesture, and the gesture-contingent gaze behavior of looking toward the recipient. 59
3.17 An experimenter demonstrating the setup of the evaluation study. . . . 61
3.18 Summary of significant models. For each interaction outcome, models for both genders, ${R_A}^2$ values, and details of statistical tests for significant gesture predictors are presented. There was no significant model for narration duration for females. (†), (*), (**), and (***) denote $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively. Significance was assessed using two-tailed t-tests. β coefficients are standardized and therefore comparable across gestures. 64
3.19 A visual summary of the results, highlighting the predictive relationships between gestures and outcomes for females and males. Solid and dashed lines represent significant and marginal effects, respectively. Numbers on lines represent standardized β coefficients and p-values for predictors. 66
3.20 We used a learning-based approach to *model* how humans employ multimodal behaviors involving speech, gaze, and gestures during narration and *generate* multimodal behaviors for a humanlike robot to perform the same narration task. 71
3.21 Speech features for gestures. Features were identified through affinity diagramming of human data and by using the guideline suggested by McNeill (1992). An example of each type of feature is also provided. . 72

3.22 The proposed dynamic Bayesian network for modeling and generating multimodal behaviors. C denotes a latent cognitive process that directs verbal, involving speech (S), and nonverbal, involving gaze (Ga) and gestures (Ge), processes. 72
3.23 Our conceptualization of the process of generating speech, gaze, and gestures. Learning and inference for high-level features are performed at the *feature* level, while specific motions are defined at the *domain* level. 75
3.24 An example of the robot displaying speech, gaze, and gesture behaviors generated by the proposed learning-based approach. 76
3.25 Results on perceptions of the robot and retelling performance. Only significant results are marked. (NS), (†), (*), (**), and (***) denote $p > .10$, $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively. 80
3.26 Results on perceptions of the robot and retelling performance. Only significant results are marked. (NS), (†), (*), (**), and (***) denote $p > .10$, $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively. 81
3.27 Results on participants' use of positive and negative adjectives in describing the robot's behavior. Top three choices of adjectives and percentages of used positive and negative adjectives are listed. Values in parentheses indicate how many participants used the adjective to describe the robot. Only significant results are marked. (*) denote $p < .050$. 82

4.1 Data collection of dyadic interactions in a sandwich-making task. Top Left: Two participants, wearing gaze trackers, working together to make a sandwich. Top Right: The participant's view of the task space from the gaze tracker. Orange circle indicates their current gaze target. Botoom: The layout of ingredients on the table. The ingredients, from top to bottom, left to right, are *lettuce1*, *pickle1*, *tomato2*, *turkey*, *roast beef*, *bacon2*, *mustard*, *cheddar cheese*, *onions*, *pickle2*, *ham*, *mayo*, *egg*, *salami*, *swiss cheese*, *bologna*, *bacon1*, *peanut butter*, *lettuce2*, *pickle3*, *tomato1*, *ketchup*, *jelly*. . . . 95

4.2 Illustration of episodic prediction analysis. Each illustrated episode ends at the start of the verbal request. The top plot shows probabilities of glanced ingredients that may be chosen by a customer. Note that the plotted probability was with respect to each ingredient. By calculating the normalized probability across all ingredients, we can determine the likelihood of which ingredient will be chosen. The bottom plot shows the customer's gaze sequence. Ingredients are color coded. Purple indicates gazing toward the bread. Black indicates missing gaze data. An anticipation window is defined as the time period starting with the last change in the prediction and ending with the onset of the speech utterance. The beginning and end probabilities are the probabilities of the predicted ingredient at the beginning and end of the anticipation window. 99

4.3 Two main categories of correct predictions: one dominant choice (top) and the trending choice (bottom). Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Purple indicates gazing toward the bread and yellow indicates gazing toward the worker. Black indicates missing gaze data. 101

4.4 Examples of incorrect predictions. Red indicates the prediction made by the SVM-based predictor, whereas blue indicates the actual ingredient requested by the customers. Purple indicates gazing toward the bread whereas yellow indicates gazing toward the worker. Black indicates missing gaze data. 103

4.5 Examples of special gaze patterns. Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Blue lines indicate the ingredients that the customers picked. Red lines are our predictions. Purple indicates gazing toward the bread, whereas yellow indicates gazing toward the worker. Black indicates missing gaze data. 106

4.6 I propose an "anticipatory control" method that enable robots to proactively plan and execute actions based on an anticipation of a human partner's task intent as inferred from their gaze patterns. 107
4.7 System diagram of the implemented system for anticipatory action preparation and execution. 108
4.8 The task setup of the human-robot interaction experiment in Study 4. The robot served as a "worker" to prepare a smoothie ordered by a human "customer." Between the robot and the user were a menu for the user to look for his/her choice and a workspace for the robot to prepare the order. 113
4.9 Tukey box plots of data from the objective measures. The extents of the box represent the the first and third quartiles. The line inside the box represents the second quartile (the median). The difference between the first and third quartiles is the interquartile range (IQR). The ends of the whiskers represent the first quartile minus 1.5 times IQR and the third quartile plus 1.5 times IQR. (***) denotes $p < .001$. 117
4.10 Tukey box plots of data from the manipulation check and subjective measures. The extents of the box represent the the first and third quartiles. The line inside the box represents the second quartile (the median). The difference between the first and third quartiles is the interquartile range (IQR). The ends of the whiskers represent the first quartile minus 1.5 times IQR and the third quartile plus 1.5 times IQR. (***) denotes $p < .001$. 120

5.1 I studied human-human handovers (top) in a household scenario, identified strategies that humans used for coordination, implemented them on an robotic manipulator, and evaluated their effectiveness in supporting coordination in human-robot handovers (bottom). 131

5.2 Average durations of the receiver's states in the regular and tasked conditions. Receivers took longer in the *retrieve* and *place* states, suggesting that these were the states in which the giver had to adapt their actions to the receiver's availability. Error bars indicate 95% confidence intervals. 134
5.3 Velocity profiles and example snapshots from adaptive coordination strategies—slowing down (top) and waiting (bottom)—displayed by givers in human-human handovers. 135
5.4 Average velocities of the giver's hand across different states in the regular and tasked conditions. Givers slowed down their actions during *retrieve*, *give*, and *retract* states. Error bars indicate 95% confidence intervals. . . 136
5.5 Cross-validation results of our KNN model in predicting receiver states using the *tasked* dataset. The dashed line indicates baseline accuracy (16.67%). 139
5.6 An example human-robot handover using our method for adaptive coordination that employed the *waiting* and *slowing-down* strategies. . . 140
5.7 Interaction plots and ANOVA test details for objective measures of team performance. 145
5.8 Interaction plots and ANOVA test details for subjective measures of user experience. 145
5.9 Data from measures of team performance. P, R, and A represent the *proactive*, *reactive*, and *Adaptive* coordination methods, respectively. Pairwise comparisons use a Bonferroni-adjusted α level of .008 for significance. Error bars indicate 95% confidence intervals. 147
5.10 Data from measures of user experience. P, R, and A represent the *proactive*, *reactive*, and *Adaptive* coordination methods, respectively. Pairwise comparisons use a Bonferroni-adjusted α level of .008 for significance. Error bars indicate 95% confidence intervals. 148

6.1 The quality of joint action largely depends on the *bandwidth of awareness* between the interaction partners. Two processes—expression and prediction—help increase the bandwidth of awareness. From a robot's point of view, the process of expression is to use behavioral means to reveal internal states to the human partner. The process of prediction is to understand and predict the partner's internal states by monitoring his/her behavioral signals. 152

6.2 A research approach towards human-robot joint action in the real-world. This dissertation research has focused on the laboratory modeling, development, and evaluation (Solid line). The research approach used in this dissertation can be augmented with field observation and deployment, as well as modeling and evaluation with multiple application scenarios, to enrich the applicability of the developed systems. Future research is outlined by dashed lines. 167

ABSTRACT

Successful integration of robots into human environments to support daily activities, such as assisting with household chores, promises to have positive impacts on people's quality of life. Such integration requires robots to serve, assist, and work with human users in a variety of joint actions. *Joint action* is a form of social interaction involving interaction partners who work cooperatively to coordinate attention, communication, and actions to achieve a shared goal. This dissertation seeks to design coordination mechanisms for robots to support human partners in such joint actions.

In a series of five studies, I explore how human-inspired coordination mechanisms can be realized for autonomous interactive robots. In Studies 1–3, I focus on designing gaze cues and gestures for robots to direct human partners' attention in supporting effective communication. In Study 4, I investigate how a robot might monitor a human partner's gaze cues to predict the partner's intent and how the robot can utilize the predicted intent to perform anticipatory actions to achieve efficient joint action with the partner. In Study 5, I explore how a robot can adapt to its partner's working pace by monitoring the partner's actions as well as task progress. Results from these studies show that the robots using the human-inspired coordination mechanisms can engage people in joint actions to elicit greater team performance (e.g., reduced idle time and task completion time) and enhanced user experience (e.g., positive perceptions of the robot and the interaction).

This dissertation contributes to enabling natural, effective joint actions between humans and robots. Throughout my investigation, I develop autonomous robot systems that manage real-time sensing, planning, and acting to support joint actions with humans. I explore innovative approaches to modeling, generating, and evaluating social behaviors for robots. Moreover, results from the studies provide insights to understanding human-human joint action and designing human-robot joint action. This dissertation further motivates future research toward integrating robots into everyday human environments to engage people in joint actions to create social, cognitive, and task benefits.

1 INTRODUCTION

Robots hold great promise in improving people's quality of life in a variety of ways (Christensen et al., 2009). As the demographic structure of our society is becoming more constrictive, characterizing long life expectancy and a low birth rate, robots are designed to provide assistance to people at nursing homes and help family members with domestic chores. Such service robots could support healthy aging to compensate for people's decreased physical or sensory capabilities. Moreover, robots could also play a significant role in applications of rehabilitation and therapy. They can leverage the characteristics of embodiment to participate in joint activities with patients with physical difficulties to aid rehabilitative exercises as well as to promote therapeutic actions, such as exhibiting social behaviors to engage with and invite social responses (e.g., mutual gaze) from people with autism. Furthermore, robots are developed to work alongside humans to create more productive and economical manufacturing environments. Robots are now deployed to work separately from humans in manufacturing. Co-worker robots not only can help with dangerous or repetitive tasks but also can work side-by-side with humans to utilize their relative strengths in completing joint tasks. For instance, robots could work on welding and painting tasks while human workers work on assembly tasks that require dexterous motor skills. Additionally, robots are envisioned to support educational training, participate in recreational activities, serve in public places such as airports and museums, and more. The time when robots become common alongside humans in support of various activities is soon arriving. The successful introduction of robots into our daily lives requires robots to engage in a variety of *joint actions* with humans.

Joint action can be characterized as interaction partners working cooperatively to coordinate attention, communication, and actions to achieve a shared goal (Clark, 1996; Sebanz et al., 2006). As social animals, humans are remarkably good at joint action. People develop skills and strategies to accomplish complex tasks collectively (Tomasello, 2009). People work as a team to improve task outcomes such as increased efficiency, to offload task and cognitive burden, and to achieve goals that are

otherwise infeasible to be achieved by each individual. For example, in professional automobile racing, pit crews swiftly replace tires, refuel, and carry out necessary repairs and adjustments in a limited period; surgeons, nurses, anesthetists, and assistants work seamlessly together to perform surgery; and firefighters respond collectively to emergencies and catastrophes. Yet, collective efforts are not only seen in professional activities. Day-to-day interactions frequently involve people working together as well. For instance, people coordinate actions in assembling furniture which involves passing tools and pieces of components to one another and moving the assembled furniture jointly to a desired location. Even conversation requires individuals to engage in joint action, coordinating conversational acts to achieve common ground (Clark and Brennan, 1991; Garrod and Pickering, 2004); for example, people coordinate attention and verbal and non-verbal behaviors with their food servers in ordering takeouts at deli shops.

While humans seem to engage in joint action effortlessly in everyday interactions, successful joint action involves careful employment of various coordination mechanisms. Increasing evidence from cognitive science and psychology has revealed key ingredients of successful joint action, including directing others' attention to a task-relevant target (Tomasello, 1995), anticipating others' intent and actions (Sebanz and Knoblich, 2009), and coordinating actions in space and time (Sebanz et al., 2006).

This dissertation aims to enable robots to engage in successful joint action with humans, contributing to the introduction of robots into human environments, by equipping them with human-style coordination mechanisms. In particular, I draw on the five core coordination mechanisms that support joint action, as outlined by Sebanz et al. (2006), in my investigations. These mechanisms are *joint attention*, *action observation*, *task-sharing*, *action coordination*, and *perception of agency*.

Joint attention involves following and directing an interaction partner's attentional focus (e.g., eye gaze) to an object or event of interest in the environment and is considered a key ingredient to social interaction (Tomasello, 1995). This mechanism helps establish common ground between partners in joint action (Clark and Brennan, 1991; Sebanz et al., 2006). Without proper common ground shared

by partners, the performance of joint action would be impaired (Clark and Krych, 2004). In addition to attending to a joint interest at the right time, people observe their interaction partners' actions to understand their action goals. Such action observation is particularly critical in inferring what the partners are going to do (Sebanz and Knoblich, 2009). Moreover, a common understanding of the shared task between partners facilitates predictions of what other partners would and should do next. Furthermore, people coordinate their actions with those of their partners in time and space to achieve common goals. As joint action becomes seamless and harmonious, interaction partners may perform actions around the same time that result in almost identical effects, causing partners not able to distinguish whose actions cause particular effects (Farrer and Frith, 2002).

This dissertation investigates how the coordination mechanisms of *joint attention*, *action observation*, *task-sharing*, and *action coordination*[1] can be realized in robots and how robots might employ these mechanisms to support joint action with humans to create greater team performance and improve user experience.

1.1 Central thesis

The central thesis of this dissertation is that **robots employing coordination mechanisms grounded in human-human joint action can interact effectively with humans in joint action, leading to improvements in team performance and user experience**. To examine this thesis, I conducted a series of investigations guided by prior research on joint action in cognitive science and psychology. My investigations consisted of five studies (Chapter 3–5) that focused on different coordination mechanisms of joint action (Sebanz et al., 2006). Each study involved characterizing human behavior in joint action quantitatively, implementing human-inspired behavioral characteristics for an interactive robot, and evaluating how the robot with the implemented characteristics might improve joint action with humans in a laboratory experiment.

[1]The mechanism of *perception of agency* in joint action is left for future research as this research focuses on sequential joint action (See the below discussion of research scope).

1.2 Scope

Joint action is a complex process that involves partners employing various coordination mechanisms to achieve common goals. While prior investigations in human-human joint action have been a source of inspiration for enabling joint action between humans and robots, human-robot joint action is far from natural and effective for applications in the real world, partly due to the incomplete knowledge of joint action and limitations of robotic technologies. As an effort towards the realization of natural, effective human-robot joint action, I focused on my investigation in the following ways.

In this dissertation, I focused on *situated* joint action in which partners interact with each other in a shared physical space. This type of colocated joint action is distinguished from joint action where partners are physically separated, for instance a phone conversation. Moreover, I focused on *cooperative* joint action but not competitive joint action. In cooperative joint action, people share common goals, such as cooperatively unloading groceries. In contrast, people might have competing goals in competitive joint action; for instance, players of the opposite team in a basketball game have different goals (e.g., win for their team) while engaged in the joint action of playing basketball. Furthermore, I did not consider game-theory style cooperation (Marsh et al., 2009) where participants cooperate with a conflicting goal.

I also concentrate my investigation on *dyadic sequential* joint action where two interaction partners cooperatively work together in a sequential manner to achieve a shared goal. This characterization is separated from one-to-many, many-to-one, or many-to-many joint actions. Finally, I studied *ordinary* joint action in which interaction partners did not need particular training to perform the joint action (e.g., unloading dishes or ordering a drink), different from specialized joint action where particular training is required, such as rowing and cheerleading (Marsh et al., 2009).

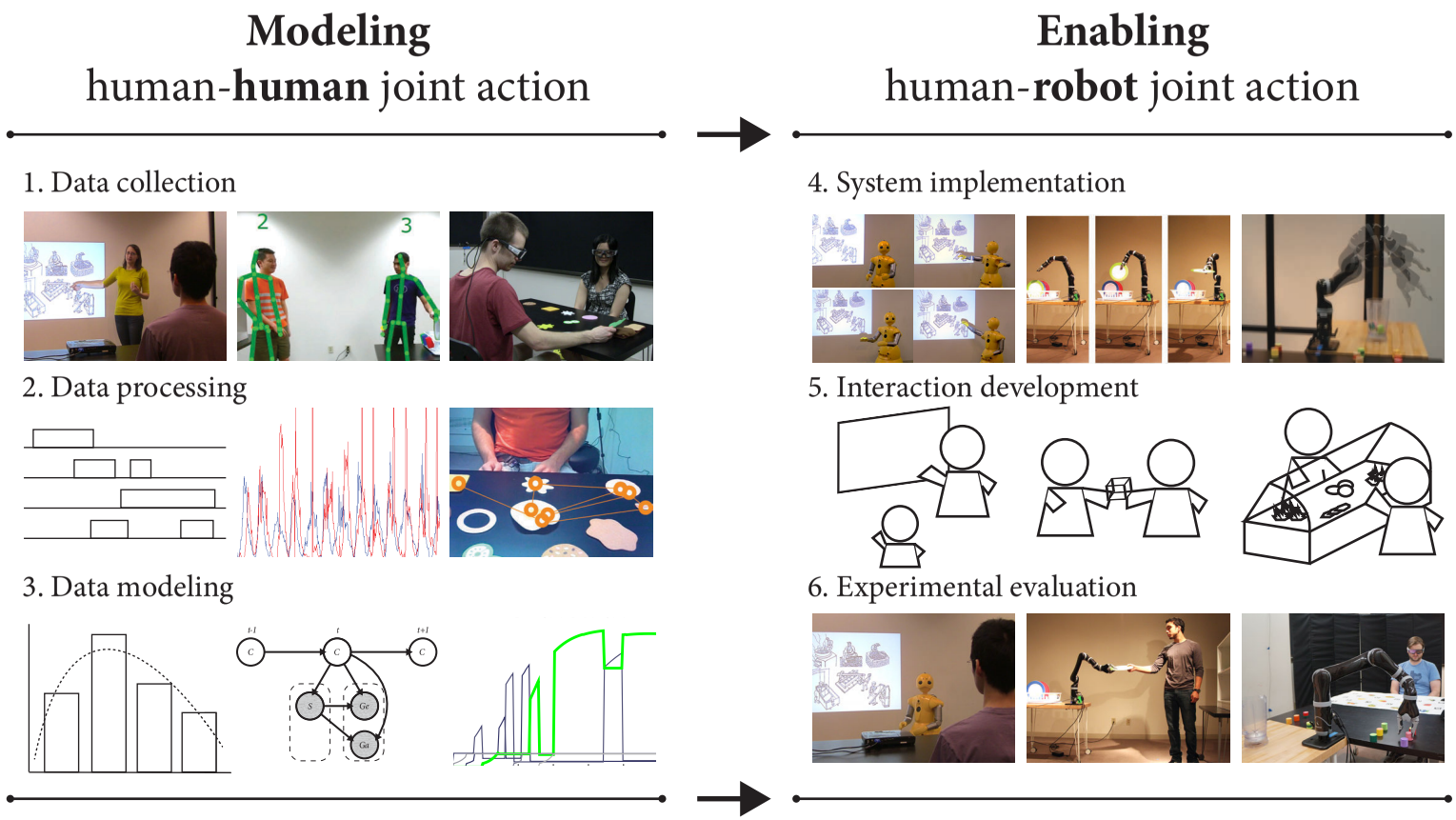


Figure 1.1: The research approach used in this dissertation leverages models of human-human joint action to enable human-robot joint action.

1.3 Research approach

I seek inspiration from human-human joint action–either scientific findings of human interaction (Study 1) or parameterized models of joint action (Studies 2-5)–to develop human-inspired robotic controllers that direct robots' behaviors toward achieving natural, effective human-robot joint action. The premise of this human-inspired approach to designing robotic systems is that users would be able to interact with the robotic systems intuitively if designed properly since people are already experts in engaging in joint action with others. Human-inspired approaches have been explored in prior research on designing effective human-robot interactions (Moon et al., 2014; Mutlu et al., 2012; Shi et al., 2013).

In particular, the human-inspired approach I followed in this dissertation can be summarized in six steps, as illustrated in Figure 1.1. The first three steps aim to obtain an understanding of human joint action while the last three steps intend to realize a robotic system that is capable of engaging in joint action effectively with human users.

Modeling human-human joint action

Step 1. *Data collection* — Human participants are invited to the laboratory to enact a scenario in which robots are envisioned to participate. The enactment sessions are recorded using devices such as high-definition cameras or Kinect sensors. The goal here is to collect data of human interaction that will later be used to develop a computational understanding of human behavior.

Step 2. *Data processing* — The collected data usually requires additional processing either to annotate with features of interest (e.g., behavioral signals such as gaze cues and gestures) or to reduce sensor noises.

Step 3. *Data modeling* — A variety of modeling techniques are explored and employed to develop a computational model that represents characteristics of the studied human behavior. A model can be constructed by simply extracting descriptive statistics of human behavior (e.g., average duration of a gesture or frequency of gaze cues) or by using artificial intelligence techniques such as probabilistic

graphical models and support vector machines. The resulting model generally involves behavioral parameters that can later be manipulated or used to operate robot behaviors.

Enabling human-robot joint action

Step 4. *System implementation* — An interactive robot system is developed to implement the obtained computational model of human-style behavior. The robot system usually involves various components that can manage real-time sensor inputs, make decisions about what, how, and when to act, and control a robot based on the decisions. The implementation usually involves using multiple threads to achieve real-time interactivity. Additionally, I utilize various software packages, including OpenCV, Microsoft Speech API, MoveIt!, and ROS, in developing systems that are powered by advances in robotics research.

Step 5. *Interaction development* — Interaction scenarios that represent applications in which robots have been envisioned to participate with human users are developed. These scenarios resemble human day-to-day interactions. The goal is to use these scenarios to study how well the implemented robot systems can engage in joint actions with humans.

Step 6. *Experimental evaluation* — Human participants, different from those in the data collection, are invited to the laboratory to interact with the developed system. The experimental task is generally similar to that in the data collection. The goal is to obtain an understanding of the performance of the system and how people might perceive the system and their interaction with it. This understanding is quantified using a variety of objective, subjective, and behavioral measures. Depending on the context, objective measures usually cover the dimensions of accuracy (e.g., predictive accuracy) and efficiency (e.g., task completion time); subjective measures can include users' perceptions of the robot in terms of different characteristics (e.g., intelligence, naturalness, and competency) and their experience interacting with the robot (e.g., engaged, pressured, and enjoyable); behavioral measures capture how users respond to the robot behaviorally (e.g., duration of locating a specific object).

As detailed in the above steps, my approach to enabling natural, effective human-robot joint action is inherently multidisciplinary, involving the use of various techniques in artificial intelligence, robotics, human-computer interaction, and data analytics. Moreover, as my investigations involve testing with human participants, the experimental protocols of the studies reported in this dissertation were reviewed and approved by the University of Wisconsin–Madison's Education and Social/Behavioral Science Institutional Review Board (IRB).

1.4 Research platforms

I used two robots—*Wakamaru*[2] and *MICO*[3]—with different characteristics in morphology and functionality in my investigations of human-robot joint action (Figure 1.2). The Wakamaru robot, developed by Mitsubishi Heavy Industries, has an anthropomorphized appearance with a head, two eyes, and two arms. It is approximately 1 meter tall. The humanlike characteristics allowed for joint attention cues, such as eye gaze and gestures. Thus, the Wakamaru robot was used in my investigation of joint attention in Studies 1–3.

The MICO robot, developed by Kinova robotics, is an one-arm robotic manipulator with a two-finger gripper. The arm has 6 DoF and has a kinematic structure different from human arms. While it moves in a robotic manner, its flexibility in reaching and grasping makes it useful for tasks involving physical manipulations. I used the MICO robot in Studies 4 and 5 to study joint action in contexts involving physical manipulations.

1.5 Overview of contributions

This dissertation contributes chiefly to the field of human-robot interaction (HRI) while also providing contributions to robotics and human-computer interaction (HCI). Throughout my investigation, I seek to make systems, methodological, and

[2]http://www.mhi-global.com/products/detail/wakamaru.html
[3]http://www.kinovarobotics.com

Wakamaru

MICO

Figure 1.2: Research platforms used in this dissertation. The Wakamaru robot was used in Studies 1–3 whereas the MICO robot was used in Studies 4–5.

empirical contributions to enabling human-robot joint action. While the contributions of this research are presented in Chapter 7, I here provide an overview. In this research, I developed several interactive robotic systems that were able to engage with people in joint tasks. These implementations of interactive robotic systems provided insight to the feasibility and limitations of current technologies in realizing effective human-robot teams. Moreover, I proposed new approaches to modeling, generating, and evaluating social behaviors for robots to communicate with people effectively in joint action. These innovative approaches serve as tools for future research on designing robot behaviors for effective human-robot interaction. Additionally, I conducted five studies to investigate how the implemented robotic systems might interact with people in joint action. Findings from these studies provided empirical evidence showing the effectiveness of human-inspired coordination mechanisms in creating greater team performance and user experience in human-robot joint action. Overall, this research contributes to the successful introduction of robots into human environments to serve, assist and work with people in everyday activities.

1.6 Dissertation overview

This chapter motivates, introduces the thesis and scope of, presents research approach and platforms used in, and highlights contributions of this dissertation research. The rest of this dissertation is organized as follows.

Chapter 2 provides background knowledge on human-human joint action, focusing on coordination mechanisms that support joint action as suggested by Sebanz et al. (2006). In addition, I review previous work on enabling joint action between humans and robots. Chapters 3–5 present five studies I conducted as my investigations towards human-robot joint action. In particular, Chapter 3 focuses on how robots might use the mechanism of *joint attention* to direct users' attention and communicate information to them effectively. It includes three studies focusing on using different behavioral cues to direct users' attention. Based on findings in the literature of human communication, Study 1 examines how gaze cue can be used to effectively direct attention of others. Study 2 investigates effects of different types of gestures in shaping human-robot interaction. Study 3 explores how speech, gaze, and gestures might be coordinated to produce coherently communicative behaviors.

Extending the mechanism of joint attention in knowing the attentional focus of the interaction partner, Chapter 4 studies the mechanism of *action observation* that monitors partners' gaze and predict their intentions in preparation of anticipatory actions for a service robot. Chapter 5 builds on the capacity of action observation and explores how *task-sharing* and *action coordination* might be realized and used to facilitate fluid joint action.

In Chapter 6, I discuss findings found in and lessons learned from the studies. I also discuss limitations of this dissertation that motivate future research. Finally, this dissertation is concluded by the system, methodological, and empirical contributions it makes to the field in Chapter 7.

2 BACKGROUND

In this chapter, I first review coordination mechanisms from psychology, cognitive science, and neuroscience and organize my review based on the mechanisms proposed by Sebanz et al. (2006) (Section 2.1). I then review previous research in the field of human-robot interaction and robotics that focused on enabling joint action between humans and robots (Section 2.2).

2.1 Coordination mechanisms of joint action

While research in cognitive and developmental psychology and neurosciences has intensively investigated how individuals are engaged in seamless, efficient joint action, the following review, drawing on Sebanz et al. (2006), focuses on five mechanisms of joint action: *joint attention*, *action observation*, *task-sharing*, *action coordination*, and *perception of agency*.

2.1.1 Joint attention

Joint attention—the process that leads interaction partners to visually attend to the same object or event in a shared environment—has been considered a fundamental skill to social interaction (Tomasello, 1995), language acquisition, and imitative learning (Baldwin, 1993). This process of attending to a joint interest helps establish perceptual common ground in joint action (Sebanz et al., 2006) and contributes to the *grounding* process that updates mutual belief to ensure successful communication and coordination in joint action (Clark and Brennan, 1991).

Humans develop skills of joint attention at an early age. Infants have been observed to follow eye gaze as early as the age of nine months (Baron-Cohen, 1997) and continue to enhance the ability as they grow. Along with the development of gaze following, infants learn how to manipulate adults' attention by pointing gestures before their first birthdays. Around the same period of time (i.e., 13 months), infants are able to use referential words to draw attention from adults (Kaplan and

Hafner, 2006). The developmental significance of joint attention has been not only found in the above pieces of evidence but also linked to social deficits, such as autism spectrum disorder (ASD) that can cause challenges in social communication and interaction (Baron-Cohen, 1997).

Joint attention can be divided into two parts—*initiating* and *responding to* joint attention (Mundy and Newell, 2007). Initiating joint attention is to direct the interaction partners' attention to the focus of interest using social cues, such as gaze and pointing gestures. This initiation of joint attention is important in joint action as it proactively aligns the perceptual common ground between interaction partners. Responding to joint attention, on the other hand, is to follow the interaction partners' attentional cues to the shared objects or events. This process also plays an important role in joint action. Without knowing what others are attending to, people are less likely to establish proper common ground and therefore prone to failures in joint action.

Joint attention naturally involves the use of multimodal behaviors. As shown in the review mentioned above, evidence in infant development suggested that at an early age infants are able to use gaze cues, pointing gestures, and verbal references in directing adults' attention (Bangerter, 2004; Baron-Cohen, 1997; Kaplan and Hafner, 2006). This use of multimodal behaviors becomes common in daily interaction (Ekman and Friesen, 1969b), where people employ multiple behaviors as communicative acts together to direct interaction partners' attention. For example, one may look at and point to an object while uttering "this one." In this example, the person coordinates multimodal behaviors in speech, gaze, and gestures in drawing others' attention to the object. These three communicative acts together produce rich behaviors to support effective communication and have been studied in a variety of contexts, such as narrations (Sidnell, 2006) and reenactments of previous events or experiences (Streeck, 2010).

In addition to pointing gestures, there are various forms of gestures that communicate and draw people's attention in joint action (Heath, 1992; Kendon, 1994; McNeill, 1992). Moreover, joint attention could be achieved by the use of broader demonstrative actions, such as "material signals"—signals that people use to indi-

cate things by manipulating material objects, using locations, and demonstrating actions (Clark, 2005).

The performance of joint action is impaired if partners are unable to establish joint attention properly. For instance, people took longer time to complete a joint task if the workspace is invisible to them (Clark and Krych, 2004). Arguably, the inefficiency was due to the failure in building perceptual common ground between the parties.

2.1.2 Action observation

Although joint attention allows people to know what interaction partners are attending to, action observation allows monitoring partners' actions to understand their action goals and infer what they are going to do (Bekkering et al., 2009; Sebanz and Knoblich, 2009). This process of action monitoring and goal inference contributes to action understanding, an important skill in social interaction. While different hypotheses have been proposed to explain how humans understand others' actions, increasing evidence in neuroscience has supported the "direct matching hypothesis." This hypothesis states that the capacity of action understanding is facilitated by mapping an observed action onto one's own motor representation of the same action (Rizzolatti et al., 2001). This account posits that people leverage their motor knowledge of the observed action, as if they were performing the action, to make sense of it, also known as *motor resonance*. In addition to facilitating action understanding, this process of direct action mapping has been linked to the capacity of imitation learning underlying human culture (Rizzolatti and Craighero, 2004).

At a lower level, action understanding is enabled by a neurophysiological mechanism known as the "mirror-neuron system" (Gallese and Goldman, 1998; Rizzolatti and Craighero, 2004), which utilizes *common coding*—using the same representations to encode one's own and others' actions—to understand observed actions (Sebanz and Knoblich, 2009). Common coding allows people not only to understand but also to predict others' actions and therefore enables monitoring and detection of others' errors during interactions, supported by the finding that the same brain

area is activated while processing both self-generated and observed errors (van Schie et al., 2004). In the same vein, this shared representation of actions allows people to predict the performance of an observed action. For instance, professional basketball players could predict the success of free shots more accurately and at an earlier time compared to coaches or sport journalists who arguably had similar visual experience not necessarily had the competent motor skill (Aglioti et al., 2008).

Behavioral evidence of action observation, particularly eye gaze, has been well documented in the literature. In a block stacking task, observers' gaze fixations were found to predict performers' action goals (e.g., the locations of touch point and movement destination) (Flanagan and Johansson, 2003). Specifically, the observers' gaze led the performers' actions. Moreover, the observers' gaze patterns were similar to those when they were performing the same task themselves, indicating that the processes of observation and action involved the uses of similar motor representations, action plans, and neural mechanisms (van Schie et al., 2004). These results provided behavioral evidence to support the direct matching hypothesis as well as the phenomenon of motor resonance. Similarly, the close coordination and predictive relationship between eyes and hand movements were found in individual object manipulation (Johansson et al., 2001). It was found that people monitored key kinematic events to ensure subgoal completion and oversee the task progress when they transported objects. This behavioral manifestation indicated how eye gaze might indicate task intention and actions.

Action observation to understand and predict intention and action is essential in achieving successful joint action.

2.1.3 Task-sharing

While action observation allows an understanding of a partner's action intent from visual features, a shared task representation enables partners to know what to expect from each other ahead of time (Sebanz et al., 2006) and therefore is particularly important in planning coordination between partners in joint action (Knoblich et al., 2011). *Task-sharing* and *action observation* are closely linked, yet different

in their uses and influences in joint action. Action observation utilizes a shared representation of an observed action in inferring action intent, also known as motor resonance, whereas task-sharing leverages a shared representation of a task to anticipate partners' actions even without observing their actions (Sebanz et al., 2005). A shared task representation consists of a set of rules specifying associations between the stimulus conditions and expected actions.

Findings from neuroscience and cognitive science have provided evidence for the existence of a shared task representation and how people might utilize it to anticipate their partners' actions (Knoblich et al., 2011). For example, studies have revealed that people formed shared representations of tasks and considered their partners' action options in their action planning, even when the tasks did not require a consideration of their partners (Sebanz et al., 2003). Moreover, if an action was predictable (e.g., a known association between a stimulus and an action) and its stimulus was provided, the readiness potential (RP)—an electrophysiological signal indicating motor preparation (Libet et al., 1983)—was found to be present prior to the observation of that action (Kilner et al., 2004). This finding suggested that a shared representation of task enabled forecasting an expected action. Task-sharing, therefore, allows anticipation of upcoming actions rather than simple reaction to observed actions.

Evidence from neuroimaging studies further showed that knowing the other partner's task activated not only areas associated with the human action system but also areas associated with mental state attribution while making predictions of the partner's actions (Ramnani and Miall, 2004). The activation in areas for mental state attribution indicated that the partner was viewed as an intentional agent. This result led to a hypothesis suggesting that a representation of *intentional relations* might be involved in associating stimuli and actions (Sebanz and Frith, 2004), because the formation of such an intentional relation might also activate areas associated with mental state attribution. An intentional relation considers an agent, an object, and how the agent is related to the object (Barresi and Moore, 1996).

Shared action and task representations can work together to anticipate the partner's actions and identify potential task errors in joint action.

2.1.4 Action coordination

By understanding and anticipating another's actions and tasks, action observation and task-sharing allow for further coordination of actions between interaction partners to achieve seamless joint action. For example, in facilitating collaboration, people perform complementary actions to those of their partners, presumably driven by the joint goals inferred from the understanding of their partners' actions and tasks (Bekkering et al., 2009; Sebanz et al., 2006).

During joint action, people engage in "planned coordination," which involves them sharing a joint goal and understanding their contributions toward achieving the goal (Knoblich et al., 2011). It was suggested that planned coordination is enabled by shared task representations and joint perceptions. Joint perceptions involve directing attention to a task-relevant focus and taking each other's perspective to understand what each other can or cannot see in the space. Additionally, this perception mechanism might serve as a method to incorporate perceived capacity of the partner into one's own action planning (Marsh et al., 2006). Beside the spatial alignment of common ground via joint perceptions, people coordinate their actions temporally to their partners' actions to produce fluid workflow. The key to such temporal coordination is anticipation, which as shown by Knoblich and Jordan (2003) could be achieved by incorporating the timing of the partner's actions into one's own action planning.

In addition to planned coordination, "emergent coordination" also facilitates joint action (Knoblich et al., 2011). Emergent coordination occurs spontaneously, and usually subconsciously, between interaction partners resulting in behavioral synchronization and mimicry, including similar walking patterns (van Ulzen et al., 2008) and synchronized body sway in conversation (Shockley et al., 2003). Such unconscious synchronization and mimicry in behaviors is independent of any shared plans and joint goals between partners. Moreover, this behavioral tendency

has been shown to have positive social consequences, such as increased liking, in interactions (Chartrand and Bargh, 1999).

As enabled by mechanisms of joint attention, action observation, and task-sharing, action coordination brings together individual efforts of the partners to create joint effects of the coordinated action, leading to the completion of the joint task.

2.1.5 Perception of agency

Perception of agency in joint action concerns the ability to distinguish whose actions cause particular effects. Also known as "agency judgement" and "consciousness of action" (Jeannerod, 1999; Georgieff and Jeannerod, 1998), this problem arises in joint action usually when interaction partners perform actions around the same time that result in almost identical effects (Farrer and Frith, 2002). Prior research has shown that people had difficulties in attributing agency of an observed movement if that movement was the same as what they were performing (Daprati et al., 1997). People could be misled to believe that they had performed an action, which was in fact performed by a confederate (Wegner and Wheatley, 1999). Conversely, people could also attribute action agency to others when they actually performed the action (Wegner et al., 2003).

The confusion about action effect has been hypothesized to link to the formation of shared representations during joint action (Jeannerod, 1999). Georgieff and Jeannerod (1998) hypothesized a "who" system that governs how people might attribute agency to an action. Neuroimaging evidence supported the account of the who system and found that agency attribution is related to how people know their actions being unfolded in space and time (Frith et al., 2000). This evidence also supported how perception of agency might be related to shared representations that are offered by action observation and task-sharing.

People continuously attribute agency to actions in social interaction (Farrer and Frith, 2002). Being able to correctly attribute agency of action is particularly important in joint action as it allows action feedback so that one can adjust one's

action to elicit intended joint effects. More research is needed to address perception of agency in joint action, especially when joint action becomes fluid and seamless, and interaction partners are well immersed and experiencing the feeling of flow (Csikszentmihalyi, 1999).

In this dissertation, I draw on insights from *joint attention*, *action observation*, *task-sharing*, and *action coordination* to investigate how robots might leverage these coordination mechanisms to achieve effective interaction with humans during joint action. Below, I review prior research on human-robot joint action.

2.2 Joint action between humans and robots

Over the years, interactive robotic systems have been developed to investigate how joint action between humans and robots might be realized and to study how joint action might shape interaction outcomes, such as improved task performance and increased user experience, in various applications. In this section, I review and discuss the design space and applications of human-robot joint action. Prior work focusing on realizing the specific coordination mechanisms discussed in the previous section will be reviewed in respective chapters that investigates the particular mechanism.

2.2.1 Design space of human-robot joint action

I characterize the design space of human-robot joint action using two processes—*expression* and *prediction* (Figure 2.1). Expression concerns how robots can communicate their internal states (e.g., attention, intention, and affect) to humans. Being able to communicate internal states effectively helps the partners to adjust and coordinate their actions during joint action. Perception concerns how robots might understand their interaction partners and the relevant world in order to act adaptively.

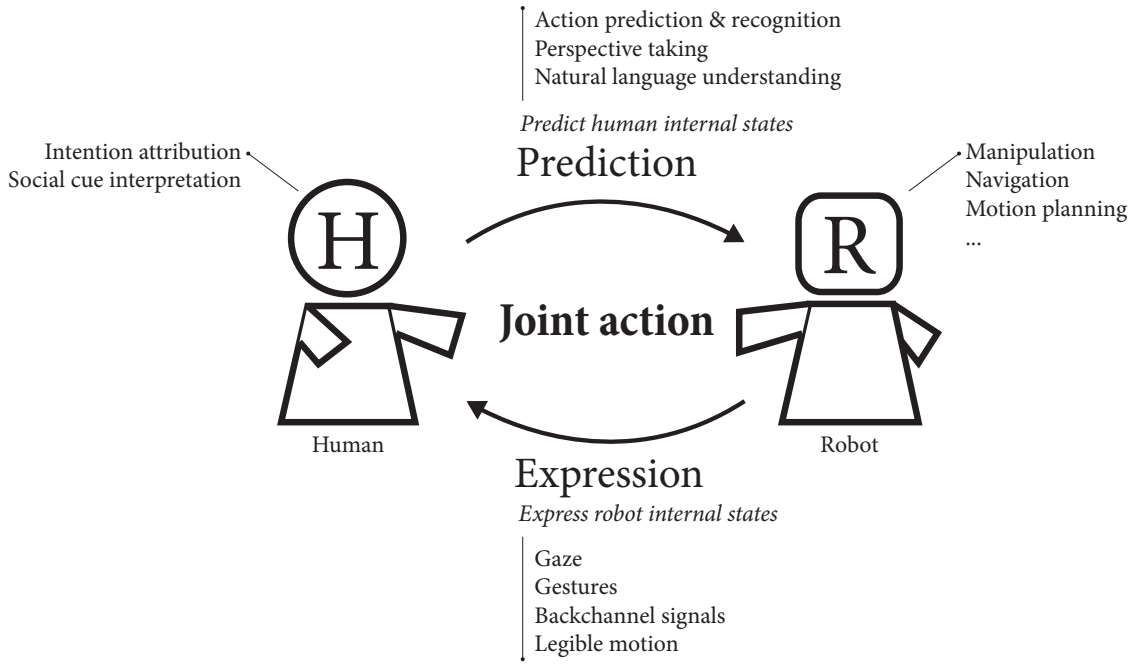


Figure 2.1: Human-robot joint action as the processes of *expression* and *prediction*. From a robot's point of view, the process of expression is to use behavioral means to reveal internal states to the human partner. The process of prediction is to understand and predict the partner's internal states by monitoring his/her behavioral signals.

Expression

There has been a large body of work studying how robots might better express themselves using behavioral characteristics during interactions with humans. While behavioral characteristics such as postures, facial expressions, proxemics, and more also facilitate the process of expression, I only highlight prior studies that explored the use of gaze cues, gestures, backchannel signals, and legible motion in manifesting robot states.

Gaze — In addition to playing an important role in human interaction (Argyle and Cook, 1976), gaze cues have been shown to facilitate communication and coordination between humans and robots. For example, a robot could use its gaze cues to regulate participant roles and improve subjective feeling of engagement while conveying information to humans (Mutlu et al., 2009a). Robot gaze cues have

also been used to improve human compliance in a joint manipulation task (Admoni et al., 2014). By employing humanlike gaze cues, robots were demonstrated to improve fluency and perceived quality of interaction in a robot-human handover task (Moon et al., 2014).

Gestures — Gestures are a natural mode of communication and support spatial coordination of attention in interaction (McNeill, 1992). It has been shown that robots could improve how quickly people cooperated with them by showing stylized gestures (Riek et al., 2010). Research has also explored how robots might couple speech with gestures in joint tasks with human users to improve users' perceptions of the robots, such liveliness (Salem et al., 2012) and naturalness (Sugiyama et al., 2007b).

Backchannel signals — Backchannel signals are behavioral cues used to convey continuous engagement in interaction. When working in teams, robots using backchannel signals, such as nods, reduce users' cognitive load and were perceived to be more engaged in the joint task (Jung et al., 2013). Yamazaki et al. (2008) showed that simple gaze movement along with nodding at proper timings encouraged users to engage in their interaction with a guide robot. Verbal backchannel signals (e.g., "um") have been utilized to create lifelike impression in robots (Kanda et al., 2007a).

Legible motion — Legible motion is motion that effectively conveys its intent to observers and can be realized through exaggerating the intended motion (Dragan and Srinivasa, 2013) or showing contrastive configurations in the motion (Cakmak et al., 2011b). Such effective communication of motion intent has been demonstrated to lead to fluid human-robot collaboration (Dragan et al., 2015).

Prediction

The opposite side of expression is prediction, dealing with how robots can understand human partners during joint action through various social skills, such as intention prediction, action recognition, and perspective taking.

Action prediction & recognition — Prior research in human-robot collaboration has explored how robots can predict users' action intent through monitoring their motion trajectories. Such intention prediction allowed robots to select appropriate

actions in response to users' actions (Koppula and Saxena, 2013). For example, predicted intentions were utilized to minimize interference with human coworkers in a shared workspace (Mainprice and Berenson, 2013) and to avoid reaching toward the same object in a collaborative task (Pérez-D'Arpino and Shah, 2015). Moreover, predicted intentions could be used to direct where a robot should look to increase its situated awareness (Ognibene and Demiris, 2013; Ognibene et al., 2013). Furthermore, a robot could lower the chance of experiencing, even avoid, bad consequences by predicting its users' misusing behavior (Brscić et al., 2015).

Previous research has also explored how robots could utilize a user's gaze to understand his or her intention during a collaborative assembly task. For instance, Sakita et al. (2004) utilized the duration of gaze fixations to predict whether or not a user was searching for a particular piece. The prediction was then used to guide a co-worker robot to help locate the piece for the user. In addition, research has also explored how gesture recognition might benefit human-robot interaction (Sugiyama et al., 2007a; Nickel and Stiefelhagen, 2007; Brethes et al., 2004).

Perspective taking — Perspective taking is the ability to view, understand, and reason a situation from another's point of view and has been shown to play an important role in facilitating joint action between humans and robots. In particular, perspective taking has been shown to help a robot to solve collaborative problems with a human partner (Trafton et al., 2005) and to learn better from a human teacher by taking the perspective of the teacher (Berlin et al., 2006). Additionally, perspective taking allowed robots to plan actions for efficient interactions with humans (Marin-Urias et al., 2008) and to recognize human partners' actions more accurately (Johnson and Demiris, 2005).

Natural language understanding — Speaking in a natural, unconstrained way promises an intuitive interaction for human-robot joint action. Research has explored how robots might understand directions (Kollar et al., 2010) and commands (Tellex et al., 2011) that were given by humans in natural language.

The review here is not meant to be exhaustive; rather, it outlines research topics on enabling transparent joint action through the processes of expression and perception. In addition, to realize joint action between humans and robots, there are

functional skills that robots need to have and are not fully discussed here. These skills include, for example, how to properly manipulate physical objects (Chan et al., 2012), navigate in dense crowds (Trautman, 2013), manage interactions in a crowded place (Kidokoro et al., 2013), and plan motion and task with respect to human users (Alami et al., 2006; Sisbot et al., 2007).

2.2.2 Applications of human-robot joint action

Socially interactive robots (Fong et al., 2003) have been developed to use in a variety of applications to support effective human-robot joint action (Bauer et al., 2008; Breazeal et al., 2004). In this section, I highlight applications where robots with the ability to engage in joint action have been demonstrated to facilitate human partners during interactions.

Healthcare — Robots have been shown to benefit the domain of healthcare and aid in tasks, such as bed baths for patient hygiene (King et al., 2010). A recent study has demonstrated how nurses might use a direct physical interface to work with assistant robots in the context of patient care (Chen and Kemp, 2010). It was shown that being able to engage in direct physical joint action improved the nurses' task performance and experiences working with the robotic assistant.

Service in the public — Robots have been designed for and deployed in public places, such as museums and supermarkets, to serve people. For example, human-inspired friendly behaviors were designed for a humanoid robot to engage people in a museum setting (Huang et al., 2014). When deployed in a science museum for two months, an interactive robot using various engaging behaviors was shown to positively influence people's visit experience (Shiomi et al., 2006). In the context of super market, robots have been equipped with interaction skills to distribute flyers to people (Shi et al., 2013), to provide recommendation and route information to customers (Shiomi et al., 2009), and to encourage people to approach for assistance (Hayashi et al., 2012).

Manufacturing — Robots have also demonstrated their potential in working with human co-workers side-by-side to improve collaboration in manufacturing

settings (e.g., Lenz et al. (2008)). In a recent study, Wilcox et al. (2013) proposed an algorithm that generates scheduling policy to allow a robot to adapt to a human co-worker's working style. Nikolaidis et al. (2013) provided insights into how robots may generate risk-aware motion when collaborating with human workers in a close proximity.

Entertainment — Robots that can engage in joint action with humans have also been used for various entertaining cases. For instance, "Shimon," a robotic marimba player (Hoffman and Weinberg, 2010), could engage people in improvisatory music play (Hoffman and Weinberg, 2011). Drawing on theories of acting, Hoffman et al. (2008) designed a robotic system that participated in live performance with two human actors on stage. These fluid musical or theatrical performances were joint actions that involved careful coordination of attention and actions among interaction partners.

Education — Robots hold promise in providing educational support. In a two-week field deployment, a robot was found to improve students' learning if they were actively engaged in interacting with the robot during the two weeks (Kanda et al., 2004). In addition to educational benefits, the robots deployed in elementary schools were able to establish social relationships with children (Kanda et al., 2004, 2007b). Learning social interaction and establishing social relationships are considered important aspects in social learning.

Household — Household is another domain where robots have been envisioned to participate in joint action with people (Beer et al., 2012; Sung et al., 2009). For example, there has been a need to have robots to retrieve household objects for people (Choi et al., 2009).

In summary, successful joint action requires interaction partners to carefully coordinate attention, communication, and actions during interaction. To enable such coordination between humans and robots, prior research has explored how various behavioral modalities and social skills might be used by robots to express themselves and to perceive their interaction partners. Successful human-robot joint action has shown promise in many daily applications in which robots assist, serve, and work with people to increase task performance and user experience.

3 JOINT ATTENTION: USING MULTIMODAL BEHAVIORS TO ESTABLISH PERCEPTUAL COMMON GROUND

3.1 Introduction

As one of the fundamental mechanisms supporting joint action, *joint attention* brings interaction partners to a shared attentional focus and helps establish perceptual common ground (Clark and Brennan, 1991; Sebanz et al., 2006). In this chapter, I focus on *initiating* joint attention that directs the partners' attention to a task-relevant place using behavioral cues (Mundy and Newell, 2007). Initiating joint attention can be accomplished via multimodal behaviors, including gaze cues, gestures, and verbal references (Bangerter, 2004; Baron-Cohen, 1997; Kaplan and Hafner, 2006). These three communicative acts together produce rich behaviors to support effective communication.

In this chapter, I describe three studies investigating how a robot might use different behavioral cues to guide a person's attention to achieve effective communication. Study 1 examines how properly timed gaze cues can utilize people's propensity to follow gaze direction to increase task performance[1] (Section 3.3). Study 2 investigates effects of various types of gestures on effective communication[2] (Section 3.4). Study 3 explores how gaze cues, gestures, and speech might be temporally coupled to produce natural, coherent communicative behaviors[3] (Section 3.5).

In addition to studying how a robot can use behaviors in multiple channels to create effects of joint attention, I explore different approaches to modeling, generating, and evaluating social behaviors for robots in Studies 1–3. In particular, Study 1 introduces the *Robot Behavior Toolkit*, an implementation of the proposed framework to generate coherent robot behaviors. Study 2 proposes an evaluation approach that explores complex relationships among behavioral variables and how

[1] Study results presented in this chapter were also published in Huang and Mutlu (2012, 2013b).
[2] Study results presented in this chapter were also published in Huang and Mutlu (2013a, 2014b).
[3] Study results presented in this chapter were also published in Huang and Mutlu (2014a).

they contribute to interaction outcomes. Study 3 presents a data-driven approach to modeling temporal dynamics in multimodal behaviors.

Results presented in this chapter not only contribute to the understanding of how robots can initiate joint attention with humans and produce coherent behaviors for effective communication, but also lead to various methodological innovations for modeling, generating, and evaluating social robot behaviors. Next, I review relevant work on the topics of human-robot joint attention as well as different approaches to modeling, generating, and evaluating social behaviors for human-robot interaction.

3.2 Related work

3.2.1 Joint attention in human-robot joint action

Given its importance in human-human joint action (Section 2.1.1), joint attention has been widely studied and established as a critical component in human-robot interaction (Breazeal et al., 2005; Huang and Thomaz, 2011; Imai et al., 2003; Scassellati, 1999; Sidner et al., 2004; Thomaz et al., 2005). One thread of this research focused on how a robot might exhibit gaze cues in response to the joint attention initiated by its human partner (Breazeal et al., 2005; Huang and Thomaz, 2011; Scassellati, 1999). These studies highlighted that robots responding to joint attention, especially through gaze, revealed their internal mental states to their human partners and thus facilitated the alignment of mutual understanding. Mutual understanding between the interaction partners was shown to reduce the number of task errors and task time in human-robot interaction (Breazeal et al., 2005; Huang and Thomaz, 2011).

Another thread in this research sought to enable proactive alignment and maintenance of joint attention with human partners (Imai et al., 2003; Scassellati, 1999; Sidner et al., 2004; Thomaz et al., 2005). These works explored how robots can use different behavioral modalities, including gaze and gestures, to manipulate their human partners' attention. Together, prior research has established the importance of joint attention in human-robot joint action.

While prior studies in human-robot interaction have demonstrated the effectiveness of multimodal behaviors in manipulating the interaction partner's attentional focus, how behaviors in multiple channels should be temporally coupled to create a coherent presentation for a robot was unexplored. Additionally, how effects of initiating joint attention using multimodal behaviors may go beyond "causing an attentional shift in the partner" (e.g., looking at what the robot was looking at). How can robots leverage this ability of manipulating people's attention to create higher-level social and cognitive benefits for their human partners? Studies presented in this chapter seek to address these unexplored questions.

3.2.2 Multimodal communication using speech, gaze, & gestures

Speech, gaze, and gestures are three common modalities used for communication. Here, I review the roles of these three behaviors in human communication and how they have been designed to support effective communication between humans and robots.

Multimodal communication in humans

Humans naturally use multimodal behaviors involving speech, gaze, and gestures during interaction. While *speech* serves as the main channel to convey information, *gestures* are widely used in conversations across different tasks, cultures, and age groups (Goldin-Meadow, 1999) to illustrate imagery (McNeill, 1992), draw attention from other participants (Heath, 1992), disambiguate unclear speech, and supplement speech with additional information (Kendon, 1994).

For speakers, gestures convey information in support of their speech, enabling them to reinforce or supplement spoken content (Kendon, 1994; McNeill, 1992). For recipients, the accompanying information facilitates the comprehension of the spoken content (Kendon, 1994; Goldin-Meadow, 2003). These benefits are observed across a broad range of communicative settings including narrative (McNeill, 1992), conversational (McNeill, 1992; Kendon, 1994), and instructional (Lozano and Tversky, 2006) communication and across cultures (McNeill, 1992; Graham, 1975; Kendon, 2004). The ability to interpret gestures also has developmental significance

(Goldin-Meadow, 2003), making gestures particularly important for teaching and learning (Roth, 2001). For example, education research has shown that gestures help young children recall information (Thompson et al., 1998) and learn new algebraic concepts (Alibali and Nathan, 2007). Moreover, gestures contribute to the shaping of students' perceptions of the teacher and keep them motivated and engaged in the classroom (Richmond, 2002).

Gaze also supplements speech, facilitating turn-taking in conversation (Kendon, 1978), signaling conversational roles (Goodwin, 1981), and regulating intimacy between participants (Argyle and Cook, 1976). These three channels make up a rich repertoire of social behaviors (Ekman and Friesen, 1969b) that play a critical role in effective communication in settings such as narration (Sidnell, 2006) and reenactments of previous events or experiences (Streeck, 1993).

Research in human communication has extensively studied the relationship between speech and gesture (Kendon, 1980; McNeill, 1992), particularly the four common types of gestures: (1) *deictic* gestures, which involve pointing toward a shared reference, (2) *iconic* gestures that illustrate concrete objects or actions, (3) *metaphoric* gestures, which use concrete metaphors to represent abstract concepts such as time, and (4) *beat* gestures, which are rhythmic movements that mark the structure of the speaker's discourse. Gestures and speech are tightly linked at both semantic and structural levels. Also referred to as *representative* gestures, deictic, iconic, and metaphoric gestures are closely related to the semantics of speech through *lexical affiliates*—the words or phrases with which gestures co-express semantic meaning (Schegloff, 1984). Beat gestures are linked to speech at a structural level and indicate significant points in speech, such as connecting discontinuous parts in speech and introducing a new topic (McNeill, 1992). Research has also studied the relationship between speech and gaze (Griffin, 2001; Kendon, 1978; Meyer et al., 1998), identifying a close coupling between these two channels, such as a tendency to gaze toward referents before they refer to them in speech (Griffin, 2001; Meyer et al., 1998) and the use of gaze cues to signal opportunities for exchanging conversation floor (Kendon, 1978).

Multimodal communication in robots

Previous research in human-robot interaction has explored the development of mechanisms for achieving natural, effective multimodal behaviors for robots, such as the development of models of gaze for displaying appropriate head movements at meaningful speech points for a museum guide robot (Yamazaki et al., 2008), aligning gaze shifts with discourse structure for a storytelling robot (Mutlu et al., 2006), and signaling task preferences along with an intentionally delayed in joint manipulation (Admoni et al., 2014). These examples highlight the importance of temporal alignment among different modalities of behavior to improve human-robot collaboration, perceptions of the robot, and overall user experience.

Prior research has also recognized the importance of gestures as a key mechanism for human-robot communication, particularly exploring how robots might (1) display humanlike gestures to support their speech and (2) use gestures in specific patterns to improve human-robot interaction. Research on realizing gestures for robots has been focused primarily on a subset of the four typical types of gestures (i.e., deictic, iconic, metaphoric, and beat gestures), such as how robots might use deictic gestures to give directions to their users (Okuno et al., 2009) or to learn a task from their users (Sugiyama et al., 2007a). This line of research also includes novel approaches to and control architectures for generating gestures (Bremner et al., 2009; Ng-Thow-Hing et al., 2010; Salem et al., 2012). For instance, Salem et al. (2012) developed a control architecture for deictic and iconic gestures for a humanoid robot performing in a human-robot joint task. Similarly, Bremner et al. (2009) introduced a gesture production approach based on actuator end-point and trajectory to generate open-hand gestures. Finally, Ng-Thow-Hing et al. (2010) proposed a probabilistic model for synchronizing gestures and speech and conducted a video-based evaluation to explore how manipulating model parameters might produce gestures with different levels of expressivity.

Another body of work in human-robot interaction involves investigating how robot gestures affect people's perceptions of the robot and their experiences. This thread of research has shown that gestures positively shape participants' affective states (Narahara and Maeno, 2007), behavioral responses to the gestures (Riek et al.,

2010), engagement in the interaction (Bremner et al., 2011), and perceptions of the robot when gestures and speech are mismatched (Salem et al., 2012). Such research involved laboratory studies in which participants performed a task with a physical robot (Bremner et al., 2011; Salem et al., 2012) or video-based studies in which participants observed robots performing gestures (Riek et al., 2010; Narahara and Maeno, 2007; Ng-Thow-Hing et al., 2010).

These lines of research advance knowledge as to how robots might use gestures toward improving human-robot interaction and highlight promise that gestures hold for shaping user experience and improving people's perceptions of robots. However, for robots to realize the full potential of using gestures, a better understanding of how they might selectively use different types of gestures to target improvements in specific outcomes in human-robot interaction is needed. Moreover, how gestures should be integrated with gaze and speech to produce natural, coherent behaviors requires further exploration.

3.2.3 Approaches to modeling, generating, & evaluating behaviors for robots

In this section, I review approaches that have been used to model, generate, and evaluate social behaviors in the context of human-robot interaction. Success in these approaches promises effective communication between humans and robots.

Approaches to modeling social behaviors

Inspection-based approaches — Inspection-based approaches involve examining the data of human interaction and explicitly specifying "rules" that characterize certain aspects of social behavior. For instance, in studying how a robot can use gaze cues to signal participant roles in triad, Mutlu et al. (2009a) recruited people to engaged in triadic conversations and extracted specifications of gaze cues (e.g., how much time to spend looking at one participant over the other) from the recruited participants. These specifications were then used to direct how a robot might use gaze cues to signal participant roles in a similar interaction context.

In Study 2, I employed a similar inspection-based approach to explore how a robot might display coherent multimodal behaviors involving speech, gaze, and gestures (Section 3.4). I explicitly specified the semantic link between speech and gestures and empirically obtained parameters to quantify the temporal speech-gesture and gaze-gesture alignments. Specifically, the speech-gesture associations were hand-coded by identifying the lexical affiliate for each gesture according to literature on human communication (McNeill, 1992) and quantifying the alignment between them. I also modeled the link between gaze and gestures by obtaining distributions of where the speaker looked while performing different types of gestures.

While this inspection-based approach has demonstrated its effectiveness in capturing saliences in behavioral variables and relationship for reproducing modeled behaviors, it requires a deliberate process to extract social specifications, including identifying behavioral variables, obtaining quantified parameters for the variables, and specifying any relationships among the variables. Additionally, while such inspection-based approaches might be feasible for modeling a small number of behaviors from small datasets, this feasibility diminishes when a larger number of behaviors or large datasets are considered. Furthermore, the models built by inspection-based approaches are highly sensitive to the decisions made and inspection methods used by the researcher in the modeling process.

Learning-based approaches — An alternative approach to the careful process of inspecting social dynamics is a learning-based approach. A learning-based approach automatically learn the behavioral parameters and potential relationships from data of human interaction. Learning-based approaches have commonly been used to build predictive models of human behavior and to control behaviors of embodied conversational agents (e.g., Lee and Marsella (2012); Morency (2010); Otsuka et al. (2007)). These approaches frequently use probabilistic graphical models (PGMs) for their support for modeling complex relationships under uncertainty.

Based on this type of learning-based approach, Study 3 seeks to use PGMs to represent human multimodal behaviors, learn model parameters from annotated

data on human behaviors, and draw on the learned model to achieve natural, humanlike robot behaviors (Section 3.5).

Approaches to generating social behaviors

In social interaction, humans employ a large number of social cues, including linguistic, vocal, and nonverbal cues, and adapt their use of these cues to the context of the interaction, particularly the activity in which they are engaged and the specific goals of the social interaction (Ekman and Friesen, 1969b). How can robots similarly draw on a large repertoire of social cues and selectively use these cues to achieve effective interactions?

Previous research on enabling robots to generate social behavior has followed two main approaches. The first approach has sought to replicate human social cues and mechanisms in robots based on a *scientific* understanding of human social behavior. For instance, Breazeal and Scassellati (1999) drew on findings from developmental psychology to develop a robot that displayed social behaviors that infants display toward their caregivers, such as facial and prosodic expressions of emotion. Mutlu et al. (2012) developed conversational gaze cues for robots based on an understanding of the cues that participants in human conversations use to effectively manage speaking turns and conversational roles.

The second approach has drawn on an *artistic* interpretation of human social behavior, particularly on principles of drama and character animation, to develop social cues for robots. For instance, Takayama et al. (2011) explored how principles of character animation such as anticipation and follow-through (Thomas et al., 1995) might guide the design of robot social behaviors that are more readable and appealing. Researchers have also explored how principles from drama and acting might enable robots to display social behaviors in live, improvised performances that are consistent with the robot's character and the goals of the interaction (Hoffman et al., 2008; Bruce et al., 2000).

Prior work has also explored how robots might generate in a systematic way the broader range of social behaviors that humans use in communication by integrating verbal, vocal, and nonverbal cues based on affective or cognitive models of

interaction. This research has developed several frameworks and architectures for generating social behavior (Breazeal and Scassellati, 1999; Holroyd et al., 2011). For instance, rather than designing specific cues or expressions for the robot, Breazeal and Scassellati (1999) used models of human emotion as the basis for systematically generating a wide range of affective social acts. Holroyd et al. (2011) developed a model that systematically generates behavioral cues, such as directed gaze and mutual gaze, adjacency pairs, and backchannel behaviors, to facilitate engagement in human-robot interaction.

A small number of existing architectures involve mechanisms that enable the robot to make decisions about selectively using social cues toward generating social behaviors that are appropriate for the context of the interaction or that help achieve the goals of the interaction. Following the *Hierarchical Task Network* formalism, Montreuil et al. (2007) built on *Joint Intention Theory* (Cohen and Levesque, 1990) to enable the robot to use a set of high-level "social rules" to assess the social appropriateness of its actions given a social interaction scenario. Duffy et al. (2005) proposed the *Social Robot Architecture* that enabled the robot to display "reactive" and "deliberative" behaviors toward satisfying the goals of the interaction. The *Pattern-based Mixed-initiative (PaMini)* dialog framework developed by Peltason and Wrede (2010) specified a set of high-level "interaction patterns" that generates dialog behaviors that are appropriate for a given task context and dialog input from human counterparts. Finally, the *BonSAI* framework developed by Siepmann and Wachsmuth (2011) specifies a set of "informed strategies" that represent how the robot might behave in response to input from human counterparts.

While the work reviewed above outlines the design space for enabling robots to systematically generate social behaviors toward supporting interaction goals, it was unclear how low-level specifications for social behavior might be linked to high-level task control to achieve specific goals. In Study 1, I seek to fill this specific gap by introducing formalisms to create low-level specifications for social behavior, including speech, vocal, and nonverbal behaviors, and to define the context and goals of the interaction (Section 3.3).

Approaches to evaluating social behaviors

To evaluate the effectiveness of designed social behaviors, researchers and designers have employed experimental methods that are commonly used in research in human factors and the behavioral sciences to assess how categorical manipulations in a small set of design variables might affect outcomes of interest. For instance, to assess how properly timed gaze aversion—breaking mutual gaze to look away from the other interaction partner—might shape outcomes of human-robot interaction, Andrist et al. (2014) designed two gaze behaviors—averting gaze at "good" and "bad" timings, respectively—and compared the manipulations to a baseline behavior where the robot did not avert its gaze during interaction. The behavior of gaze aversion was itself composite and could be directed away to up, down, left, and right. While this approach provided a sound, empirical basis for predicting the effects of the timing and presence of gaze aversion on various outcomes of human-robot interaction, it did not inform design decisions regarding the frequency with which the robot should avert its gaze nor how different directions of aversion might elicit different effects and to what extend.

To address this limitation, I propose a novel, *multivariate* evaluation method in Study 2 (Section 3.4). My method builds on multiple regression to model the relationship between a set of design variables and an interaction outcome of interest as a linear system in order to determine which design variables predict the interaction outcome and to what extent these predictors affect it. The proposed approach is different from prior uses of regression-based analysis in studying interactions with robot systems, which focus on how emergent aspects of the interaction between the user and the interactive robot system, such as the number of repeated inquiries by the user (Foster et al., 2009) or the total duration of the interaction (Peltason et al., 2012), might predict user experience with the robot system. In contrast, my proposed method involves directly manipulating each design variable to vary across a continuous range and assessing its relative contribution to predicting the outcome variable. Directly manipulating design variables in this way provides the designer with analytical and practical advantages, including understanding how the full

range of design variables affect the robot's effectiveness and how robots might go beyond human capabilities in using the designed interactive behaviors.

Additionally, the proposed method enables a more practical investigation of the rich design space for interaction through experimental trials with human participants. As the design space gets large, conventional approaches for evaluating design variables quickly become infeasible. For instance, if the designer is interested in combining four different behaviors in an interactive robot system and evaluating how these behaviors affect the robot's interaction with its users, even the simplest categorical manipulations would create $2^4 = 2 \times 2 \times 2 \times 2 = 16$ unique conditions to compare against each other, requiring a very large number of experimental trials and statistical comparisons. Due to this impracticability, design studies to date have investigated the effects of the presence or absence of a single variable, such as gaze aversion behavior (Andrist et al., 2014), or a group of variables, such as response time, approach velocity, and proximity used altogether (Huang et al., 2014). These

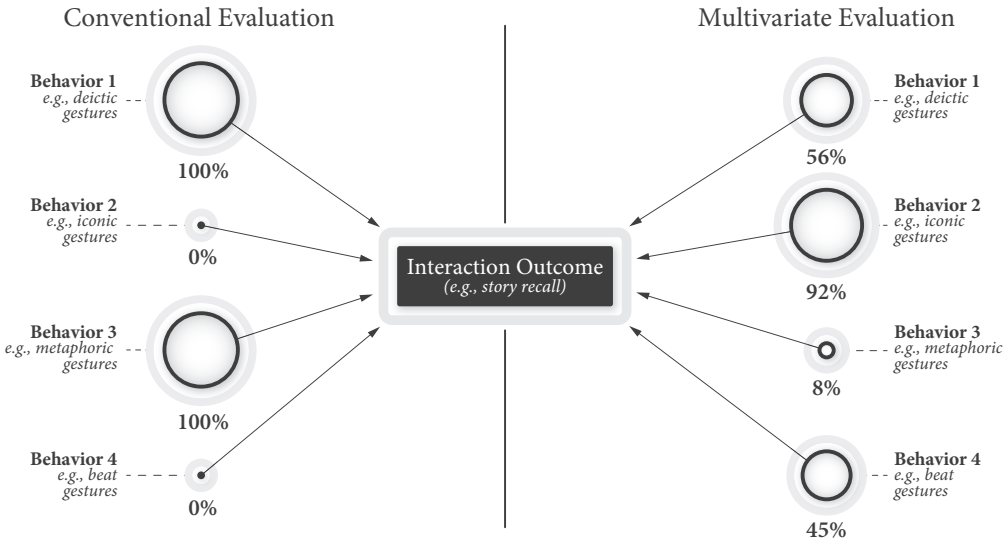


Figure 3.1: A visual comparison of the conventional evaluation method used to date and the proposed multivariate evaluation method. Conventional approaches follow categorical manipulations in a small number of design variables, while the proposed multivariate evaluation method involves joint manipulation of all design variables.

studies, however, provide either a very narrow understanding of the design space or lack the fidelity required to fine-tune the robot's use of each designed behavior to improve the desired outcomes of the interaction. The method I propose seeks to address these limitations by enabling the designer to systematically vary the use of each design variable across its full range, such as how frequently the robot might use gaze aversion behaviors, and model a linear combination of all design variables considered in the study in each experimental trial. This method provides a more holistic view of the effects of all the design variables involved in the study on the interaction between the robot and its user. Figure 3.1 provides a visual comparison of the proposed multivariate evaluation approach and the conventional method used in evaluating robot systems. Next, I present the first study in my investigation of joint attention.

3.3 Study 1: initiating joint attention via gaze cues

This study focuses on how a robot can use gaze cues along with verbal references to direct a user's attention to a task-relevant places and how such attentional directions might affect outcomes of joint tasks. Below, I first describe how referential gaze and speech cues are linked for initiating joint attention (Section 3.3.1). I then present the *Robot Behavior Toolkit* that I implemented to enable the production of coherent multimodal behaviors and how the Toolkit can generate coherent gaze and speech in this study (Section 3.3.2). Finally, an experimental evaluation is presented to show effects of a robot initiating joint attention on people's task performance and user experience (Section 3.3.3).

3.3.1 User modeling

Instead of empirically modeling how people carefully align gaze cues with speech in initiating joint attention with others, I turned to research in human communication, which has extensively studied how human gaze and utterances are highly coupled in time. In particular, it has been reported that referential gaze precedes linguistic

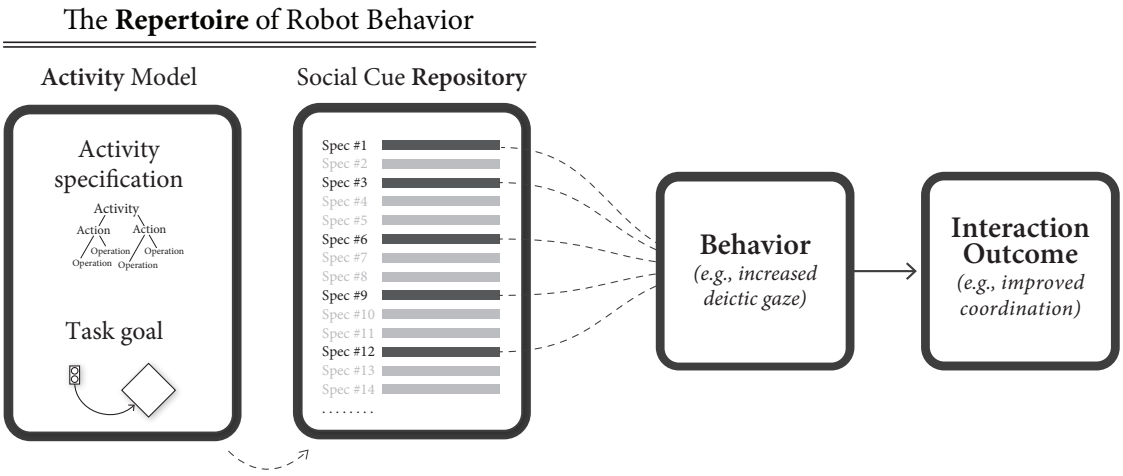


Figure 3.2: The repertoire of robot behavior consists of a *social cue repository* and a set of *activity models*, allowing a selectively use of behavior to achieve specific interaction outcomes.

reference by approximately 800 to 1000 milliseconds (Griffin, 2001; Meyer et al., 1998), meaning that people tend to look at what they are about to verbally refer to in the environment before the onset of the actual verbal reference. This close coupling relationship provides power referential cues to direct another's attention. Next, I present a the Robot Behavior Toolkit that allowed the realization of coupling referential gaze and speech cues for an interactive robot.

3.3.2 System implementation

Instead of hard coding how gaze cues should be coupled with speech references in a robot's behavior, I developed a system that generates coherent robot behaviors to achieve interaction goals. Drawing on the idea that humans select from their repertoire the social acts that help them achieve specific task or communicative goals in social interaction, I proposed a *repertoire of robot behavior*, consisting of a *social cue repository*, a collection of computational specifications of social behavior, and a set of *activity models*, characterizations of social situations that guide the robot in picking and choosing social acts from the repository that best support task or

communication goals in a given situation (Figure 3.2). I implemented the *Robot Behavior Toolkit*[4] as an instantiation of the repertoire of robot behavior.

Social cue repository

The social cue repository emulates the broad range of social acts that humans employ in interaction (Ekman and Friesen, 1969b) and consists of specifications of social acts or mechanisms derived from research on human communication. Each specification includes information on the social cues involved in the social act or mechanism (e.g., eye contact), the interdependencies that the cues might have with other social cues (e.g., contingencies between speech and gaze), and the goals that the robot might achieve using the cue (e.g., increased affiliation).

Figure 3.3 illustrates an example specification, represented in XML format. A *behavior type* tag describes the behavioral channels involved in the specification. Each

[4]This is an open-source software available on the project website. http://hci.cs.wisc.edu/robot-behavior-toolkit/

```
<rules>
  <rule id=`1`>
    <editor>editor_1</editor>
    <edit_time>timestamp_1</edit_time>
    <references>reference_id</references>
    <description>
      The referential gaze typically precedes the onset of corresponding
      linguistic reference by approimate 800 msec to 1000 msec.
    </description>
    <behavior_type>gaze</behavior_type>
    <trigger>linguistic_reference</trigger>
    <behavior>
      precede(toward(gaze, artifact), linguistic_reference, rand(800,
      1000))
    </behavior>
    <outcomes>task</outcomes>
  </rule>
  ...
```

Figure 3.3: An example behavioral specification on synchronizing referential cues in speech and gaze.

specification may have multiple *triggers* that activate the specified behaviors. The ensuing *behaviors* are described in the form of a *function*. The example illustrated in Figure 3.3 includes three arguments for the *precede* function: the behavior specified in the first argument temporally precedes that specified in the second argument by the temporal amount specified in the third argument. In the illustrated case, the action of gazing toward a referent precedes the corresponding linguistic reference by 800–1000 milliseconds, a specification derived from research on human gaze (Griffin, 2001; Meyer et al., 1998). Each specification also includes a set of interaction *outcomes* that the specified behavior might elicit, such as improving performance in the task at hand or enhancing how the robot is perceived by its human counterparts.

Activity model

How do people determine what behaviors to employ from their repertoires in social interaction? Ekman and Friesen (1969b) argue that, among other factors, the *social context or setting* that the individual is in, the *communicative goals* that the individual wishes to achieve, and the *inter-relationships* among the social acts shape people's use of social behaviors. The social context or setting might characterize the *physical environment* such as a domestic environment or a workplace, the *organization of the interaction* such as dyadic interaction or group setting, the *relative statuses of the participants* such as a supervisor or a subordinate, and the *roles of the participants* such as a speaker or a bystander.

Psychological research has proposed a number of paradigms for studying and representing context and setting in human interaction including *Activity Theory* (Leontjev, 1978), *situated action models* (Lave, 1988), and *distributed cognition* (Hollan et al., 2000). In brief, Activity Theory offers a theoretical framework and a set of principles and constructs to study and represent human activity as a complex, socially situated phenomenon (Leontjev, 1978). Situated action models describe human activity and interaction as emergent, improvisatory, and contingent (Lave, 1988). Finally, distributed cognition considers humans and artifacts as equivalent agents that make up a cognitive system toward achieving an overall system goal

(Hollan et al., 2000). Nardi (1996) provides a more detailed comparison of these three frameworks.

Among these frameworks, I argue that Activity Theory offers the richest representation for human-robot interaction and thus adopt it as the theoretical basis for system design. The paragraphs below describe the five core principles and constructs in Activity Theory and explain how they inform the design of the proposed architecture.

Construct 1: consciousness — At the core of Activity Theory is the construct of *consciousness*, which unifies attention, intention, memory, reasoning, and speech (Vygotsky, 1979). This construct motivated the implementation of specific representations for attention and intention and a consideration of interdependencies between social acts, particularly speech.

Construct 2: object-orientedness — The construct of *object-orientedness* considers material artifacts, plans of action, or common ideas to be "shared for manipulation and transformation by the participants of the activity" (Kuutti, 1996). This concept

	Activity Theory Constructs	**System Design**
❶	Consciousness	Using *context model* and *activity model* to represent attention and intention
		Viewing *speech* as an accompanying channel to other behavioral channels
		Using *memories* for cognitive processing
❷	Object-orientedness	Representing objects as *motives* in the activity model
❸	Hierarchical structure	Three layers: *activity*, *action*, and *operation*
❹	Internalization & externalization	Internalization — forming a *context model* of current setting
		Externalization — generating *behavioral specifications*
❺	Mediation	Internal tool — the *Social Cue Repository*
		External tool — sensor data

Figure 3.4: A summary of how the constructs of Activity Theory informed the system design of the Robot Behavior Toolkit.

informed the design of a representation for shared artifacts and goal-oriented actions in the system.

Construct 3: hierarchical structure — Activity Theory suggests that human activity follows a *hierarchical structure* in which activity is organized into three layers: activity, action, and operation. An activity consists of a series of actions that share the same motive. Each action has a defined goal and a chain of operations that are procedures performed under a set of conditions. The design of the activity model component of the system reflects this hierarchy.

Construct 4: internalization and externalization — The construct of *internalization and externalization* captures cognitive processes in human activities; internalization involves transforming external actions or perceptions into mental processes, whereas externalization is the process of manifesting mental processes in external actions. Internalization and externalization are analogies to the processes of forming a representation of the situation and of displaying social behavior, respectively.

Construct 5: mediation — The final construct in Activity Theory is *mediation*. Several external and internal "tools" that people might use in an activity, such as physical artifacts, and cultural knowledge or social experience that an individual has acquired might mediate human activity. For instance, the social cue repository serves as an internal tool that mediates activities between humans and robots.

These five constructs inform the design of the Robot Behavior Toolkit (Figure 3.4). Figure 3.6 illustrate the implementation of the Toolkit. The *perceptual system* takes sensor data as input and preprocesses it to an internal structure for later processing. The *cognitive system* uses the external information from the perceptual system and internal information from the *activity model* to form a context model of the current situation. The *behavior system* uses the context model to guide behavior formation based on behavioral specifications from the *social cue repository*. The output of the Toolkit is high-level behaviors defined in an XML format for the robot to execute (e.g., Figure 3.8).

In particular, the "activity model" is an implementation of the constructs of object-orientedness and hierarchical structure, while the other three constructs are realized in the other system components. In the activity model, all actions

```
<Activity id=`1`>
  <Motive>clear(table)</Motive>
  <Description>Clear objects on table</Description>
  <Participants>Self, User1</Participants>
  <Action id=`1`>
    <Outcome>Task</Outcome>
    <Goal>disappear(object)</Goal>
    <Description>
      Instruct User1 to categorize object
    </Description>
    <Operation type=`utterance`>
      <Condition>present(User1)</Condition>
      <Condition>
        known(the blue object with two pegs)
      </Condition>
      <Condition> known(the blue box)</Condition>
      <Info turn=`end`>
        Could you help me put the blue object with
        two pegs into the blue box, please?
      </Info>
    </Operation>
    ...
```

Figure 3.5: An example activity model in which the robot instructs a human partner to clear objects on a table.

associated with the activity are governed by a high-level *motive*, which each action helps fulfill by achieving its corresponding goal. Actions involve operations that are constrained by a set of conditions and that can be executed when these conditions are met. Actions have predefined *outcomes* such as "task performance" and "rapport" that specify the orientation of an action. For instance, a *task performance* outcome indicates that the action will affect task performance. Outcomes help the robot prioritize its behaviors toward achieving specific task or communication goals. Figure 3.5 provides an example activity model represented in XML format. In the example, the robot has the motive of clearing the objects on the table. To fulfill this motive, the robot instructs the user to clear the table by moving the objects to boxes. The specified "task" outcome in this example would prompt the robot to prioritize and display social behaviors that would contribute to the completion of the task.

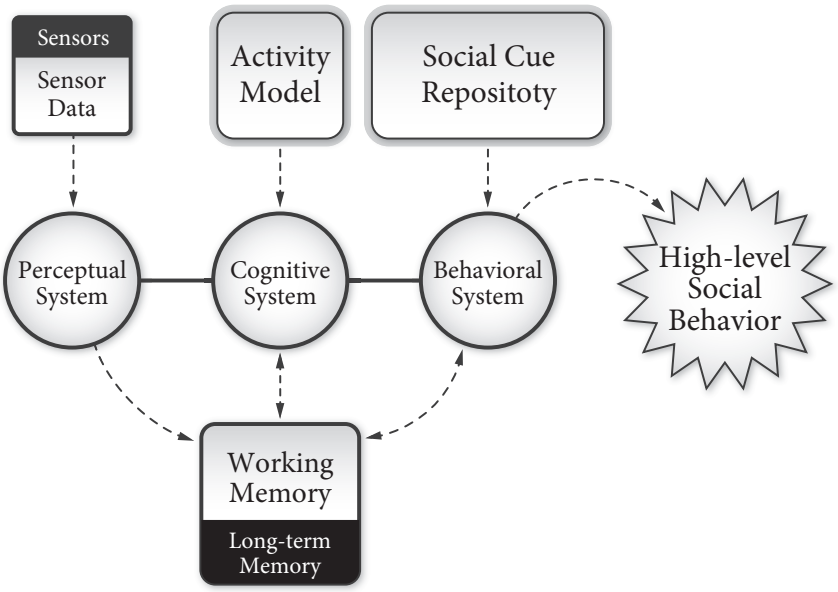


Figure 3.6: The Robot Behavior Toolkit consists of three subsystems—the perceptual, cognitive, and behavior systems; two memories—the working memory and long term memory; and supporting components—the activity model and knowledge base.

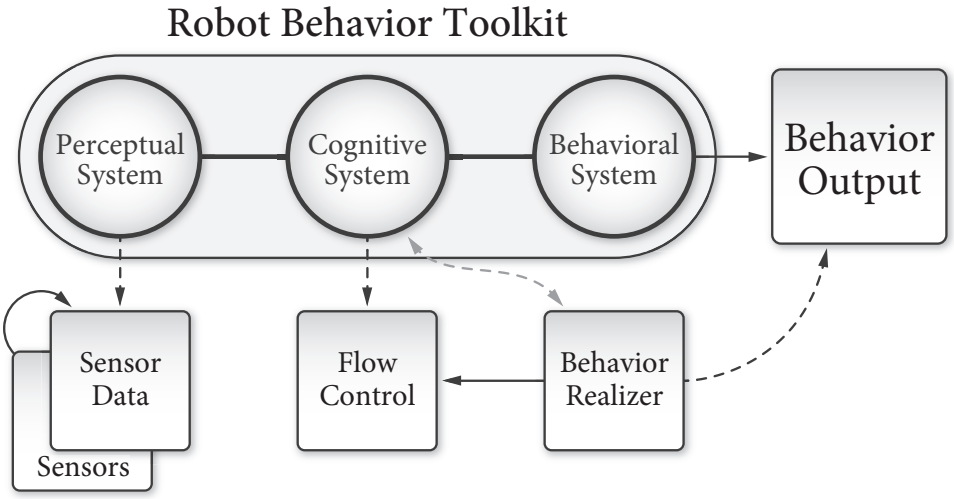


Figure 3.7: Information flow between our Toolkit and ROS. Rounded squares represent *topics*. Solid and dashed lines denote *publishing* and *subscribing* to topics, respectively. Light dashed lines denote *service* communication between nodes.

Generating social behavior

To enable the robot to display social behaviors, the system combines information from the activity model, trigger annotations, and sensor information to query the social cue repository to retrieve behaviors that would best satisfy the specified outcome for the activity. The system consolidates the resulting behaviors into an XML representation for execution. Figure 3.8 illustrates the representation for an example behavior. The Robot Behavior Toolkit is integrated with the ROS framework (Quigley et al., 2009) to support the use of available range of sensor devices and robot platforms. To date, the implementation has been tested with the Wakamaru robot and a Gazebo simulation of a PR2 robot. To use the Toolkit with a new robotic platform, researchers are required to provide a ROS *node* for that platform to subscribe to the *topic* to which the Toolkit outputs its generated behaviors (Figure 3.7).

3.3.3 Experimental evaluation

This experimental evaluation aimed to assess how a robot could leverage the coupling relationship between gaze and speech to manipulate people's attention to achieve effective communication. Below, I describe details of this evaluation.

Hypotheses

I developed three hypotheses on how the robot's properly timed gaze and speech behaviors, as generated by the Robot Behavior Toolkit, might affect social, cognitive, and task outcomes in joint tasks against different baseline behaviors.

Hypothesis 1: Participants will recall the information that the robot presents to them more accurately when the robot employs the behaviors generated by the toolkit than they will when the robot employs alternative behaviors. The basis of this hypothesis is the finding from gaze literature that gaze cues clarify what is being referred to in speech and improve story comprehension (Richardson and Dale, 2005).

Hypothesis 2: Participants' performance in a collaborative sorting task will be higher when the robot employs behaviors generated by the toolkit than it will when

```
<behaviors>
  <channel type=`gaze`>
    <action endTime=`214.5` startTime=`0` target=`unspecified`/>
    <action endTime=`1160` startTime=`214.5` target=`the green
      object with one peg`/>
    <actoin endTime=`2735.4` startTime=`1160` target=`unspecified`/>
    <action endTime=`3597` startTime=`2735.4` target=`the red box`/>
    <action endTime=`4308` startTime=`3597` target=`unspecified`/>
    <action endTime=`4963` startTime=`4308` target=`listener`/>
  </channel>
  <channel type=`speech`>
    Could you help me put the green object with one peg into the red
    box, please?
  </channel>
</behaviors>
```

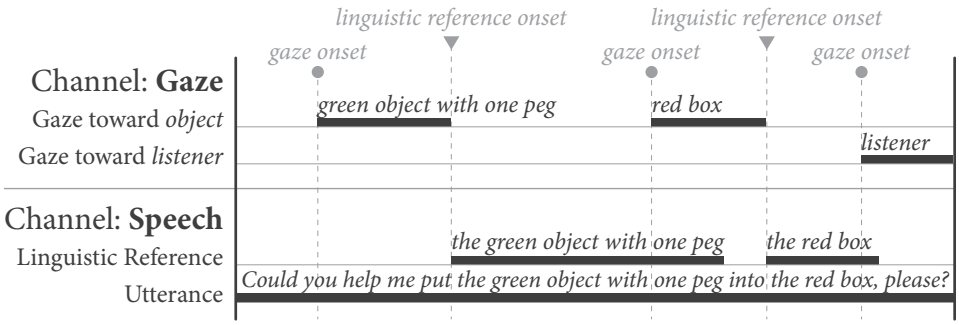


Figure 3.8: An example behavior output generated by the Toolkit in XML (top) and in visual representation (bottom).

the robot employs alternative behaviors. This hypothesis builds on prior work in human-robot interaction that suggested that appropriately timed gaze cues of a robot facilitate the effective locating of information among distractions (Staudte and Crocker, 2009).

Hypothesis 3: Participants will evaluate the robot as more natural, likable, and competent when it employs behaviors generated by the toolkit than they will when the robot employs alternative behaviors. This prediction follows findings from the prior research that gaze cues cold shape the favorability of the robot (Mutlu et al., 2006, 2009a).

Experimental design, task, & conditions

I tested the above hypotheses in a laboratory experiment, which involved two human-robot interaction scenarios in order to increase the generalizability of the findings across task contexts. In the first scenario, the robot told participants the story of the 12 signs of the Chinese Zodiac (see top picture in Figure 3.9). In its story, the robot referred to a set of cards that were laid on a table located between the robot and the participant. The cards showed pictures of the 12 animal characters and the figure mentioned in the story. The second scenario involved a collaborative categorization task (see bottom picture in Figure 3.9). In the task, the robot instructed the participants to categorize a set of colored lego blocks into different colored boxes. There were 15 blocks with different colors, sizes, and heights and two colored boxes laid on the table located between the robot and the participant. The participant did not know how the each block should be categorized and had to wait for instructions from the robot to place each block into its respective box. I used a pre-recorded human voice for the robot's speech and modulated its pitch to create a gender-neutral voice.

I manipulated the specifications in the knowledge base of the toolkit to create the following four conditions for both tasks:

(1) **Humanlike**: The robot exhibited gaze behaviors generated by the toolkit using the following social behavioral rules:

- Referential gaze precedes linguistic reference by approximately 800 to 1000 milliseconds (Griffin, 2001; Meyer et al., 1998).
- The speaker looks toward the listener at the end of a turn (Duncan, 1972).
- The speaker looks toward the person whom he/she is greeting (Kendon and Ferber, 1973).

(1) **Delayed**: The robot showed the same behaviors as it did in the *humanlike* condition except that the behaviors were delayed, resembling the timings of the listener as opposed to that of the speaker, e.g., referential gaze following the onset of the linguistic reference by approximately 500 to 1000 milliseconds (Fischer, 1998).

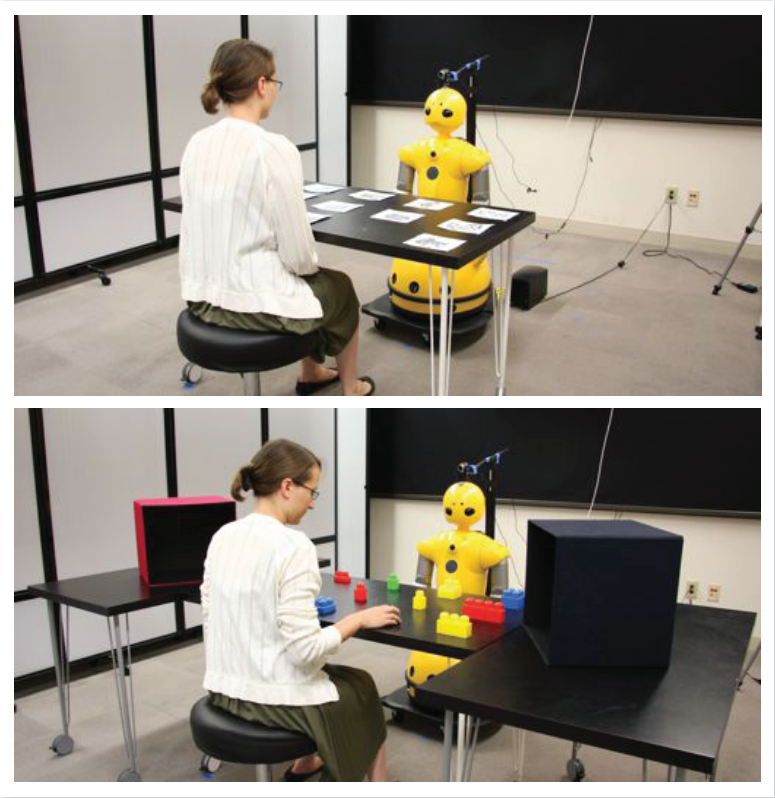


Figure 3.9: The setup of the storytelling (top) and collaborative work (bottom) tasks in the experiment.

(3) **Incongruent**: The robot followed the timings in the *humanlike* condition, but looked toward an object that was different from what was referred to in the linguistic reference.

(4) **No-gaze**: The robot did not display any gaze behaviors other than tracking the participant's face.

In all conditions, the robot tracked the participant's face when the specified gaze behavior involved looking toward the listener. The linguistic references in the robot's speech were manually marked.

The study followed a between-participants design. Participants were randomly assigned to one of the four conditions. There were four male and four female

participants in each condition. The first and second task involved a total of one and eight trials, respectively. In each trial of the second task, the order in which the robot referred to the objects was randomized.

Procedure

At the beginning of the study, the experimenter provided the participants with a brief introduction of the goals of the study and obtained informed consent. After the first task, the participants took a three-minute break while the experimenter prepared for the second task. After completing the second task, the participants were asked to complete a recall test about the story. They then filled out a post-experiment questionnaire. At the end of the study, the experimenter interviewed and debriefed the participants. The experiment took approximately 30 minutes. The participants received $5 for their participation.

Measures

The two independent variables in our study were the manipulation in the behavioral specifications for the robot's gaze behavior and participant gender. We measured two types of dependent variables, objective and subjective.

Objective. Following the storytelling task, I measured the participants' recall of the details of the robot's story. A total of 10 questions were asked in the recall test. All questions were related to the order in which the characters were presented in the story about the signs of the Chinese Zodiac. The questions followed true-or-false, multiple-choice, or multi-select formats. An example question is provided below.

> Q: The Dragon is before the Rabbit in the Zodiac cycle.
> A: "False"

Following the collaborative work task, I measured the time that the participant took to locate objects to which the robot referred. In particular, I measured the time between the end of the linguistic reference and one of the following cases: (1) the participant's last gaze toward the object before moving the object, (2) the participant touching the object, or (3) the participant reaching for the object. This measure

served as a measure of task performance and captured how fast the participants located the information needed to complete the task.

Subjective. I used a post-experiment questionnaire to measure participants' perceptions of the robot in terms of naturalness of behavior, likability, and competence. The questionnaire also included several manipulation-check questions. Seven-point rating scales were used in all questionnaire items. The naturalness scale, which consisted of seven items, measured how participants perceived the naturalness of the robot's behavior. The likability scale consisted of 10 items, which measured participants' ratings of the likability of the robot. The competence scale, which consisted of 14 items, measured the participants' perceptions of the robot's competence in the task and its overall competence. Item reliabilities for naturalness (Cronbach's $\alpha = .79$), likability (Cronbach's $\alpha = .90$), and competence (Cronbach's $\alpha = .85$) scales were sufficiently high.

Participants

A total of 32 participants were recruited for this evaluation study. All participants were native English speakers from the Madison, Wisconsin area with an average age of 24.9 years, ranging between 18 and 61. Average familiarity with robots among the participants was relatively low (M=3.25, SD=1.67) and verage familiarity with the experimental tasks was also low (M=2.13, SD=1.21) when measured by seven-point rating scales.

Results

I used one-way analyses of variance (ANOVA) to analyze data from the manipulation checks and two-way analyses of variance for objective and subjective measures.

Manipulation checks. To test whether the manipulation in the robot's gaze behavior was successful, I asked participants whether the robot's gaze seemed to be random, whether the timing of when the robot looked toward objects seemed right, whether the timing of when the robot referred to an object and looked toward it matched, and whether the robot's gaze and speech were synchronized. The results showed that the participants were able to identify the differences across conditions

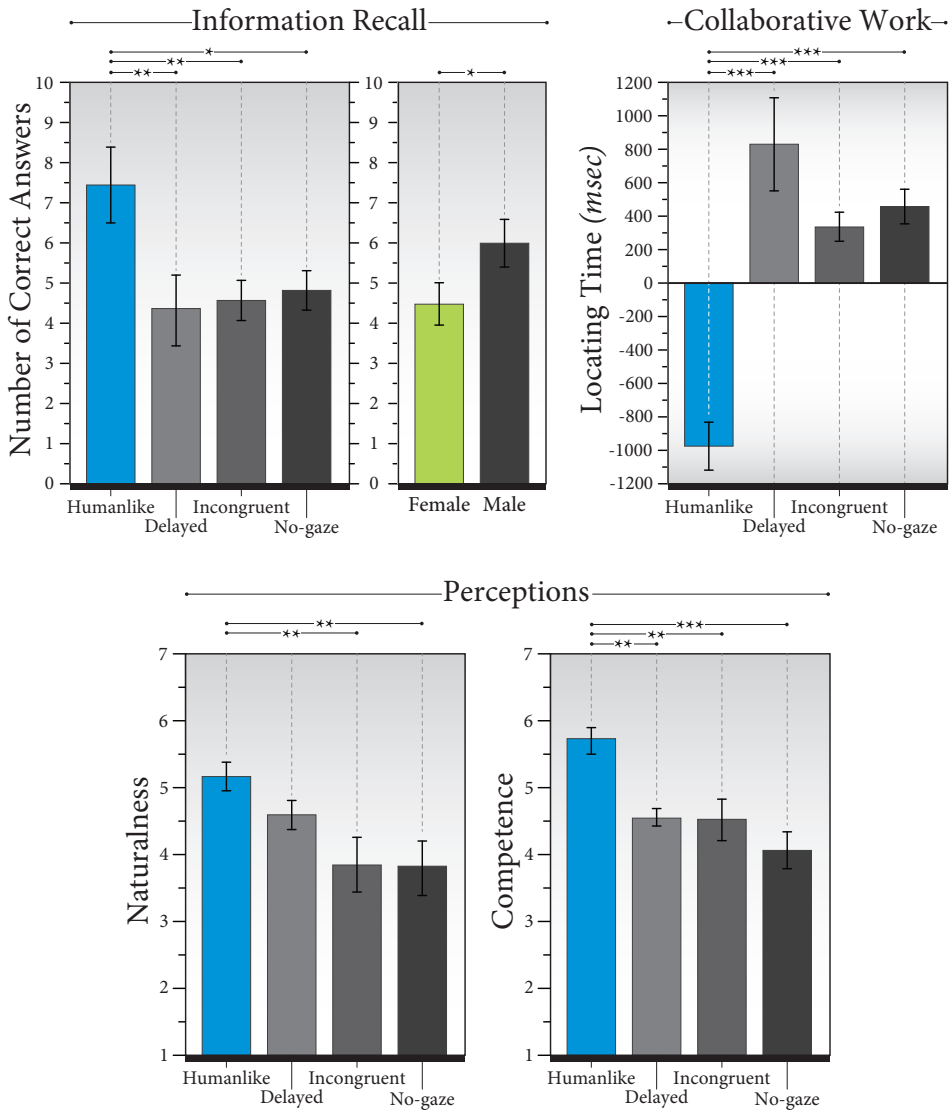


Figure 3.10: Results on information recall, collaborative work, and perceptions. (*), (**), and (***) denotes $p < .05$, $p < .01$, and, $p < .001$, respectively.

in the majority of these measures; the gaze manipulation had a significant effect on whether the participants found the robot's gaze to be random, $F(3, 28) = 3.55$, $p = .027$, $\eta_p^2 = 0.275$, whether the timing of when the robot looked toward objects seemed right, $F(3, 28) = 11.83$, $p < .001$, $\eta_p^2 = 0.559$, whether they thought that the timing of when the robot referred to an object and looked toward it matched, $F(3, 28) = 33.42$, $p < .001$, $\eta_p^2 = 0.782$, and whether they found the robot's gaze and speech to be synchronized, $F(3, 28) = 4.96$, $p = .007$, $\eta_p^2 = 0.347$. There was no main effect of the manipulation on whether the participants thought that the robot looked toward them at the right time, $F(3, 28) = 0.78$, $p = .514$, $\eta_p^2 = 0.077$. An explanation for this result is that the robot looked toward the participants for the majority of the time by tracking their faces including the *no gaze* condition.

Objective. The first hypothesis predicted that participants would have better recall of the story told by the robot when it displayed humanlike behaviors (i.e., properly timed between gaze and speech) than they would when the robot displayed alternative behaviors. The data confirmed this hypothesis. The number of correct answers out of ten questions in the recall test were on average 7.38 ($SD = 2.67$), 4.25 ($SD = 2.49$), 4.50 ($SD = 1.41$), and 4.75 ($SD = 1.39$) for humanlike, delayed, incongruent, and no gaze, respectively. The analysis of variance found a significant main effect of the robot's gaze behavior on recall accuracy, $F(3, 24) = 4.51$, $p = .012$, $\eta_p^2 = 0.360$. Pairwise comparisons using Tukey's HSD test revealed that the recall performance of the participants in the humanlike condition significantly outperformed those of the participants in delayed, $F(1, 24) = 10.45$, $p = .004$, $\eta_p^2 = 0.303$, incongruent, $F(1, 24) = 8.84$, $p = .007$, $\eta_p^2 = 0.269$, and no-gaze, $F(1, 24) = 7.37$, $p = .012$, $\eta_p^2 = 0.235$, conditions. I also found a main effect of gender on participants' recall accuracy; male participants had better recall performance than female participants had, $F(1, 24) = 5.22$, $p = .031$, $\eta_p^2 = 0.179$. These results are illustrated in Figure 3.10. Post-hoc tests showed that male participants' recall was significantly better than that of female participants in the humanlike condition, $F(1, 24) = 5.65$, $p = .026$, $\eta_p^2 = 0.191$.

The second hypothesis predicted that the participants would show better task performance—measured by the time that participants took to locate objects that

the robot referred to—in the collaborative work task when the robot displayed humanlike gaze behavior than they would when the robot showed alternative behaviors. This hypothesis was also supported by the data. When the end of the linguistic reference was represented by 0, the average time in milliseconds that the participants took to locate the object were -975.26 ($SD = 405.43$), 829.51 ($SD = 779.04$), 334.78 ($SD = 241.00$), and 457.05 ($SD = 292.51$) for humanlike, delayed, incongruent, and no-gaze conditions, respectively. The analysis of variance found a main effect of the robot's gaze behavior on locating time, $F(3, 24) = 23.22$, $p < .001$, $\eta_p^2 = 0.225$. Pairwise comparisons using Tukey's HSD test revealed that participants in the humanlike condition located the objects that the robot referred to in a significantly shorter time than participants in the delayed, $F(1, 24) = 61.12$, $p < .001$, $\eta_p^2 = 0.662$, incongruent, $F(1, 24) = 32.20$, $p < .001$, $\eta_p^2 = 0.578$, and no-gaze, $F(1, 24) = 38.50$, $p < .001$, $\eta_p^2 = 0.610$, conditions did (see Figure 3.10).

Subjective. The third hypothesis predicted that the participants would perceive the robot to be more natural, likable, and competent in the humanlike condition than they would in the other conditions. The data provided partial support for this hypothesis. Results from the subjective measures showed a main effect of the gaze manipulation on participants' perceptions of the robot's naturalness, $F(3, 24) = 4.05$, $p = .018$, $\eta_p^2 = 0.336$, and competence, $F(3, 24) = 7.79$, $p < .001$, $\eta_p^2 = 0.493$, while the effect of the manipulation on measures of likability was not significant, $F(3, 24) = 1.46$, $p = .249$, $\eta_p^2 = 0.155$. In particular, participants in the humanlike condition rated the robot to be more natural than they did in the incongruent, $F(1, 24) = 8.44$, $p = .009$, $\eta_p^2 = 0.260$, and no gaze, $F(1, 24) = 8.67$, $p = .007$, $\eta_p^2 = 0.265$, conditions. Male participants in the incongruent condition rated the robot to be less natural than female participants did, $F(1, 24) = 4.69$, $p = .041$, $\eta_p^2 = 0.163$. However, participants in both the humanlike and delayed conditions found the robot to be equally natural, $F(1, 24) = 1.58$, $p = .221$, $\eta_p^2 = 0.062$. The participants in the humanlike condition rated the robot to be more competent than they did in the delayed condition, $F(1, 24) = 10.81$, $p = .003$, $\eta_p^2 = 0.311$, incongruent, $F(1, 24) = 11.14$, $p = .003$, $\eta_p^2 = 0.317$, and no-gaze, $F(1, 24) = 21.37$, $p < .001$, $\eta_p^2 = 0.471$, conditions. These results are also illustrated in Figure 3.10.

3.3.4 Discussion

The results provide support for the majority of my hypotheses in measures of information recall, collaborative work, and perceptions of the robot. Participants had better recall of information and located objects that the robot referred to faster when it used humanlike gaze behavior generated by the toolkit (i.e., properly timed gaze and speech behaviors) than they did when the robot displayed alternative behaviors. Moreover, participants found the robot to be more natural and competent when it exhibited humanlike gaze behavior than they did in other baseline conditions. These results suggest that *initiating joint attention* using referential gaze was effective in evoking social, cognitive, and task outcomes in human-robot interaction.

The data indicates that the participants in the delayed, incongruent, and no-gaze conditions needed roughly 300 to 800 milliseconds to locate the object that the robot referred to after it completed the linguistic reference to the object (see Figure 3.10). This result is consistent with findings in the gaze literature; in the absence of speaker gaze cues, partners look toward the object of reference approximately 200 to 300 milliseconds after they hear the reference (Altmann and Kamide, 2004) and approximately 500 to 1000 milliseconds after the onset of the spoken reference (Fischer, 1998). The result suggests that the participants in the baseline conditions (i.e., delayed, incongruent, and no-gaze) relied primarily on the robot's speech to locate the object of reference, while those in the humanlike gaze condition used gaze information to locate the object, completing the task even before the robot ended the linguistic reference.

The experimental evaluation also showed that a small number of behavioral specifications was sufficient to generate robot behaviors that achieve significant social, cognitive, and task improvements in human-robot interaction. While this study used a small number of behavioral specifications for gaze behavior for experimental purposes, the Toolkit offers the potential to realize complex humanlike behaviors by combining a large number of specifications for multiple channels of behavior, which is a significant challenge when hard-coding behavioral specifications into robots. The Toolkit also offers social scientists and HRI researchers the ability to

validate new behavioral specifications by realizing them in interactive human-robot interaction scenarios.

Finally, a particular limitation to note in the experimental evaluation was the use of simulated sensor data. In this study, I focused on generating robot behavior rather than recognizing human behavior. However, more investigation is needed to understand how the Toolkit might function in more realistic interactive settings in which the recognition of human activity and the environment might be incomplete due to unreliable sensor data.

3.4 Study 2: effects of referential gestures on information delivery

Study 1 demonstrated the effectiveness of referential gaze cues on directing participants' attention, leading to improvement in their retention of information delivered by the robot. In this study, I focus on another prominent modality—gesture—that has been documented in the literature in human communication its efficacy in manipulating people's attention to support communication. In particular, I sought to understand how various types of gestures affect the effectiveness of communication. Below, I first report how people use various gestures when communicating with others (Section 3.4.1). Drawing on the understanding of gesture use, I implemented a robot system that could emulate human-style gestures (Section 3.4.2). The implemented system was evaluated in a human-robot interaction study to assess how a robot could use different gestures to elicit different interaction outcomes (Section 3.4.3).

3.4.1 User modeling

Data collection & processing

Task & setup — To study how people use gestures, I developed a narration scenario in which a narrator described "the process of making paper" with the aid of a

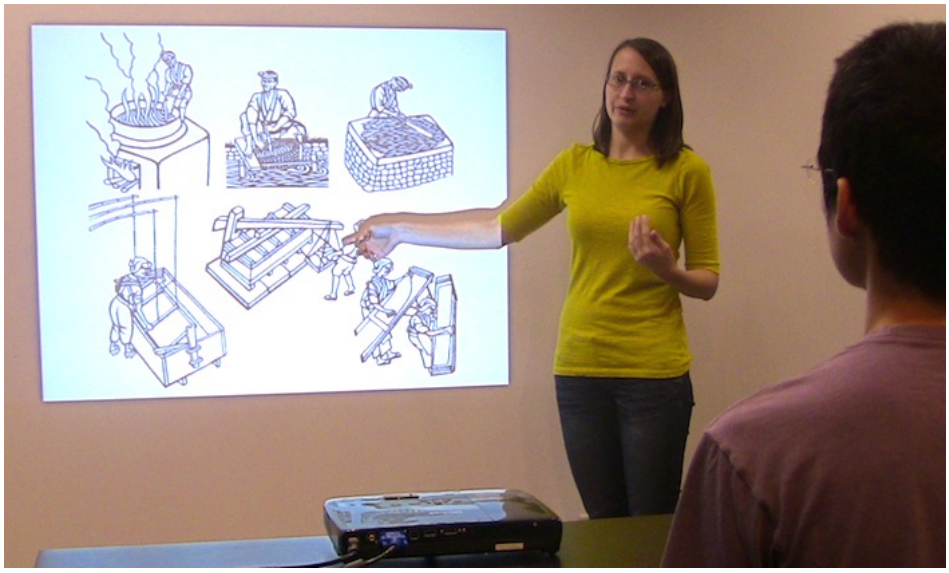


Figure 3.11: The data collection study in Study 2. Participants were asked to teach other participants about a topic they were trained on. The goal was to understand how participants used various gestures during their teaching sessions.

projected figure that depicted the process to a recipient, as shown in Figure 3.11. I chose narration as the setting for the study because narration naturally elicits the use of a rich set of gestures (McNeill, 1992) and fully engages the observer with the narrator's behaviors. Narration also serves as an appropriate scenario for human-robot joint action, as robots are envisioned to provide their users with information in joint activities. For example, robots are envisioned to serve as museum tour guides, shopping assistants, or receptionists (Burgard et al., 1999; Kanda et al., 2009; Lee et al., 2010). Finally, narration provides us with a rich set of metrics to measure outcomes such as story recall and perceptions of the narrator.

Participants — To model how human narrators employ gestures, I recruited four dyads of participants, aged 21.6 years on average (SD $= 1.77$), and matched them to represent all gender combinations (i.e., MM, MF, FM, and FF). For each dyad, one participant acted as the narrator, while the second acted as the recipient. The narrator was given text, pictures, and videos on the topic of the process of making paper and asked to review them for approximately 25 minutes. After the

Gesture	Categories	#	%	Example
Deictics *54 items*	Concrete references	30 items	55.6 %	*"the stamper"*
	Abstract references	18 items	33.3 %	*"the cooking process"*
	Pronouns	10 items	18.5 %	*"this person here"*
Iconics *105 items*	Action verbs	45 items	42.9 %	*"peel it off"*
	Nouns	44 items	41.9 %	*"a big basket"*
	Descriptors	8 items	7.6 %	*"the thickness"*
Metaphorics *51 items*	Actions	12 items	23.5 %	*"the order that we went in"*
	Relative quantities	8 items	15.7 %	*"more weight"*
	Time	7 items	13.7 %	*"the next day"*

Figure 3.12: Top lexical affiliate categories for each type of representative gesture. Percentages represent the amount by which categories of lexical affiliates co-occur with each gesture type. For example, 42.9% of the lexical affiliates for iconic gestures are "action verbs." Note that categories might overlap.

preparation phase, the narrator presented the process to the recipient. The average time for narrations across trials was 4.49 minutes (SD $=$ 1.13). The trials were videotaped for behavioral coding and analysis. A primary rater coded all of the data, and a secondary rater coded 10% of the data to assess inter-rater reliability. The reliability analysis showed perfect agreement for gesture type (Cohen's $\kappa = .845$) and gaze target (Cohen's $\kappa = .916$) based on guidelines suggested by Landis and Koch (Landis and Koch, 1977).

Developing a gesture model

Gesture and speech are co-expressive channels in human communication (McNeill, 1992; Kendon, 1994). In this work, I modeled two particular aspects of gesture with respect to speech: *gesture points*, the points in speech where the speaker displays gestures, and *gesture timing*, the times when a gesture begins and ends. I also modeled *gesture-contingent gaze cues*, gaze cues displayed during gesturing, based on research that has identified interdependencies between gestures, speech, and gaze (Sidnell, 2006; Streeck, 1988).

Gesture Points. Utterances involve *lexical affiliates*—words and phrases that co-express meaning with representative gestures, including deictic, iconic, and

metaphoric gestures (Schegloff, 1984)—that inform us on when a robot might need to gesture and what type of gesture it might perform. To capture this relationship, I identified lexical affiliates associated with representational gestures (i.e., deictic, iconic, and metaphoric gestures) and used affinity diagramming to group them into categories of gesture points for each type of gesture (Figure 3.12). These categorizations showed that deictic gestures were frequently used to describe references, including visual representations of steps and objects involved in the process of making paper that appeared on the projected figure, and in conjunction with pronouns. Iconic gestures were mostly observed during descriptions of actions and concrete objects, while metaphoric gestures were generally observed when the speaker described abstract concepts involving actions, relative quantities, or time. Because beat gestures connect an utterance not at the semantic level but rather at the structural level (McNeill, 1992), I did not empirically identify lexical affiliates for beat gestures. Instead, discontinuities in speech, such as introducing a new concept, served as gesture points for these gestures (McNeill, 1992).

Gesture Timing. Gestures and utterances are closely related in the temporal domain (McNeill, 1992). The timing of a gesture might affect how people perceive and interpret it. While research has suggested that the *stroke* of a gesture ends before the end of its lexical affiliate (McNeill, 1992), when a complete gestural phrase should begin and end relative to its lexical affiliate has not been specified. In this work, I empirically obtained these temporal parameters for representative

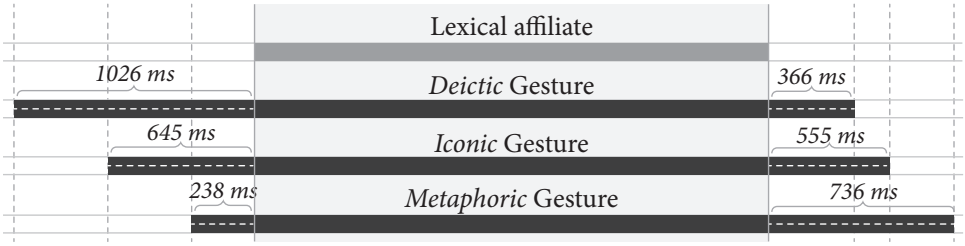


Figure 3.13: A schematic of temporal alignment between gestures and lexical affiliates (not to scale). For instance, iconic gestures began on average 645 ms before and ended 555 ms after their corresponding lexical affiliates.

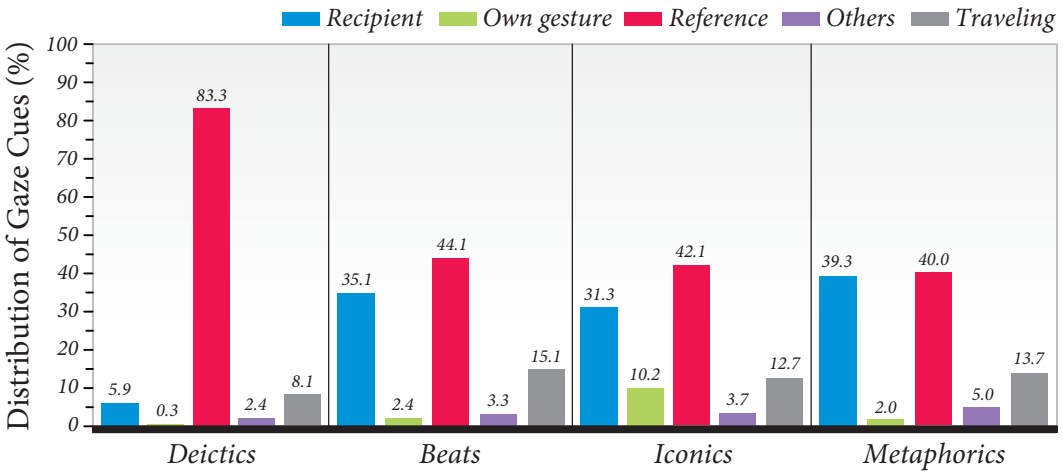


Figure 3.14: Distributions of targets for gesture-contingent gaze for each type of gesture. The human data showed four main gaze targets: the *recipient*, the narrator's *own gesture*, the *reference*, and *other*, non-task-relevant targets. *Traveling* represents transitions between these targets. Narrators gazed most toward references while displaying deictic gestures but split their gaze evenly between the recipient and references while displaying other types of gestures.

gestures, as summarized in Figure 3.13, which confirmed that gesture initiation typically precedes the onset of lexical affiliates (Schegloff, 1984; McNeill, 1992).

Gesture-contingent gaze cues. The distribution of gaze cues during each type of gesture was modeled in order to coordinate the production of gaze, gesture, and speech. I categorized gesture-contingent gaze cues into four gaze targets—*recipient*, narrator's *own gesture*, *reference*, and all *other* places. An additional category for *traveling*, the time spent transitioning between the four categories, was also created (Figure 3.14).

People tend to look toward shared references during conversation (Argyle and Cook, 1976). The data showed a similar trend, as narrators looked toward references most of the time while gesturing. This behavior was observed during deictic, beat, and iconic gestures and was particularly notable for deictic gestures, during which speakers looked toward the reference approximately 83% of the time. During metaphoric gestures, narrators looked toward references and recipients at equal rates. Interestingly, approximately 10% of narrator gaze during iconic gestures was directed toward the gestures, which narrators might display to direct the recipient's

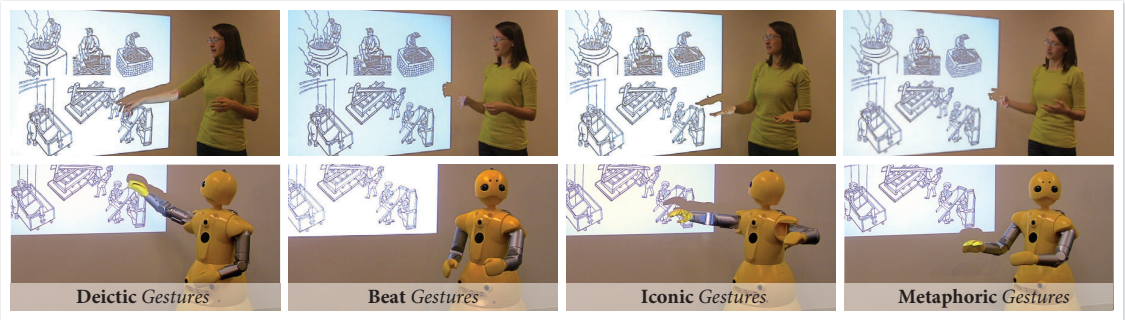


Figure 3.15: Examples of the four common types of gestures—*deictic*, *beat*, *iconic*, and *metaphoric* gestures—observed in human storytellers (top) and implemented into the robot (bottom). The narrator uses deictic gestures to point toward an object of reference, beat gestures before introducing a new concept, iconic gestures to depict a concrete object such as "a flat wooden surface," and metaphoric gestures to visualize abstract concepts such as "about six hours."

attention to the gesture and to signal that the current gesture is relevant to the ongoing utterance (Streeck, 1993). Narrators did not display this behavior during metaphoric gestures, potentially because the abstract lexical affiliates with which they are associated might require speakers to look toward the recipient to affirm mutual understanding of their speech.

3.4.2 System implementation

The robot's gestures were designed based on the observations of the human narrators' gestures in the modeling study (See Figure 3.15 for examples of the four typical gestures). While different narrators varied slightly in how they performed gestures at a given gesture point, they displayed semantically common elements. For example, when describing "beating (paper) with a stick," participants displayed one-handed or two-handed up and down movements at different speeds and with different degrees of tilt. For each unique gesture point, I created one robot gesture that captured the common elements that I observed from the human narrators displayed at that gesture point. Robot gestures were created through puppeteering, which involved manually moving the robot's arms while recording key frames of the gestural trajectories.

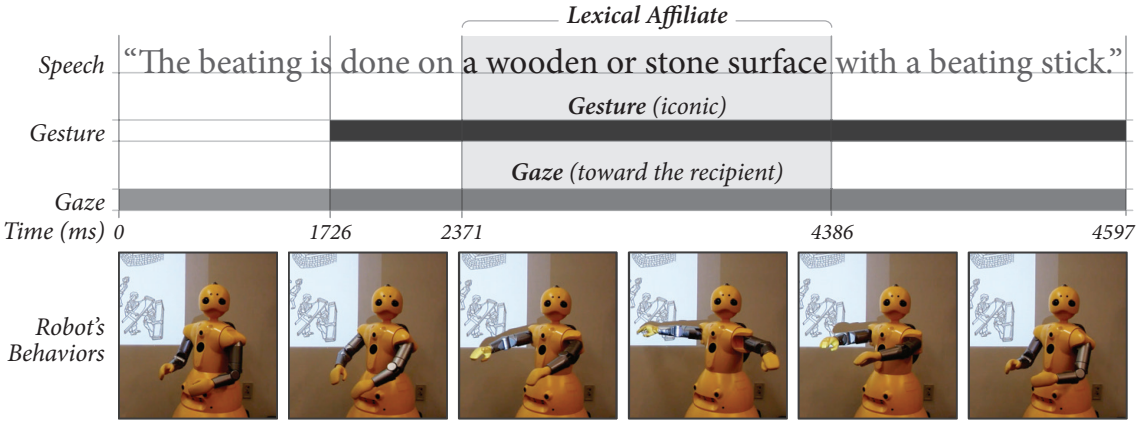


Figure 3.16: An example utterance from the robot's narration, including the lexical affiliate "a wooden or stone surface," its corresponding iconic gesture, and the gesture-contingent gaze behavior of looking toward the recipient.

For implementation and evaluation, I used the narrative script that the participants used for preparation in the user modeling study. I manually marked all possible gesture points in the script based on the lexical affiliate categories shown in Figure 3.12, identifying 30, 30, and 25 points for deictic, iconic, and metaphoric gestures, respectively. Twenty-four points for beat gestures were marked based on heuristic principles (McNeill, 1992). Gestures created for the robot were manually assigned to these gesture points. The gesture-contingent gaze cues were produced by mirroring the distributions shown in Figure 3.14.

I implemented the gesture model and its contingent gaze cues on a Wakamaru humanlike robot using the Robot Behavior Toolkit developed in Study 1. Synchronization among gestures, gaze, and the robot's speech was realized using Algorithm 1. Figure 3.16 illustrates an example synchronization of speech, gesture, and gaze cues.

3.4.3 Experimental evaluation

The goal of this evaluation was to understand how the four types of gestures may have different influences on interaction outcomes. To this end, I proposed the *mul-*

Algorithm 1 Synchronization Among Speech, Gaze, & Gestures

Require: Speech utterances, timestamp[onset,end]-ID pairs for lexical affiliates and beat points

```
for each utterance do
    new nonverbalBehavior
    for each timestamp[onset,end]-ID pair do
        gesture.selectGestureFromLibrary(ID)
        gesture.setGestureDuration(duration(timestamp[onset,end]))
        gaze.setGazeTarget(sampleFromGazeDistribution(gesture.getType()))
        gaze.setGazeDuration(gesture.getDuration())
        nonverbalBehavior.append(gesture,gaze)
    end for
    nonverbalBehavior.executeBehavior()
end for
```

tivariate evaluation to explore the complex relationships between the manipulated use of various gestures and measured outcomes in human-robot interaction. Below, I provide details on the use of the multivariate evaluation and experimental results.

Experimental design

Following a system-level evaluation paradigm (Peltason et al., 2012), I manipulated the amount by which the robot displayed each type of gesture and measured how this variability affected interaction outcomes. For each gesture type, I marked all possible gesture points at which human narrators displayed gestures of that type. To manipulate the *amount* by which the robot would display gestures of that type, a random number between 0 and the number of possible points (i.e., 30, 30, 25, and 24 for deictic, iconic, metaphoric, and beat gestures, respectively) was drawn from a uniform distribution, which served as the amount by which the robot would display gestures of that type during its narration. To manipulate the gesture *points* at which the robot would display gestures, a subset of gesture points that matched this percentage amount was randomly selected from the set of all possible gesture points. For each participant, this process was repeated for all gesture types,

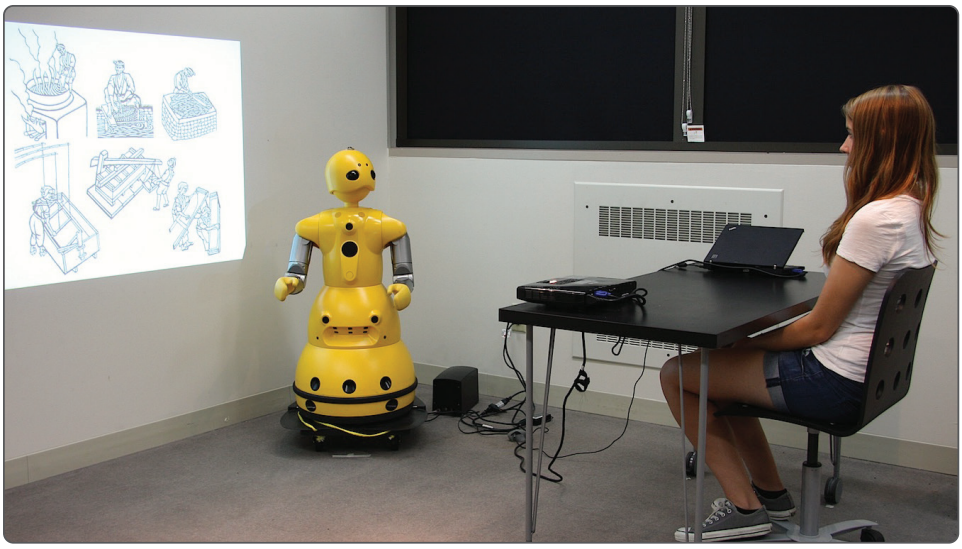


Figure 3.17: An experimenter demonstrating the setup of the evaluation study.

producing the necessary variability in the robot's gestures to investigate how well different gestures predicted interaction outcomes.

Procedure

The evaluation study started when the experimenter obtained the participant's consent to taking part in the study. Following this step, the experimenter asked the participant to be seated across from the humanlike robot and to listen to the robot tell a story on the process of making paper, as shown in the right image in Figure 3.17. The robot's narration lasted approximately six minutes. After the story, the participant was asked to complete a five-minute distractor task followed by a quiz on the topic of the narration. The participant was then asked to retell the process of making paper in the same experimental setting, which was videotaped for later analysis. The experiment concluded with a post-experiment questionnaire for evaluating the participant's perceptions of the robot. For each participant, the experimental protocol took approximately 30 minutes. Each participant received $5 for taking part in the study.

Measures

I developed objective, subjective, and behavioral measures to understand how the robot's use of different types of gestures affected the participants' information recall, their perceptions of the robot, and their ability to retell the robot's story.

The primary objective measure was how accurately the participants recalled the information presented by the robot, measured by a quiz consisting of 11 multiple-choice or true-or-false questions. The subjective measures sought to capture the participants's perceptions of the robot in terms of naturalness of behavior (5 items; Cronbach's $\alpha = .78$), competence (8 items; Cronbach's $\alpha = .82$), and effective use of gestures (2 items; Cronbach's $\alpha = .81$). We also measured their evaluations of their engagement (8 items; Cronbach's $\alpha = .82$) and rapport with the robot (6 items; Cronbach's $\alpha = .83$). Participants responded to these measures using seven-point rating scales.

In addition to the objective and subjective measures, I measured how the robot's gestures affected the participants' ability to retell the robot's story, particularly the participants' articulation of the process of making paper, as indicated by story length, and use of gestures during narration.

Analysis method

I used a backward stepwise multivariate linear regression to analyze the data. The linear regression method is used to model a response variable y by a linear combination of predictor variables x_i, represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e$$

In the model, β_0 is a constant, and β_i is a regression coefficient for corresponding predictor variable x_i, indicating to what extent the variable predicts the response variable. e denotes the residual error of the model. The backward stepwise regression process starts with all predictors included in the model. The process iteratively removes the predictor with the highest p-value as a relatively less powerful predictor of the response variable. The iterations continue until certain stop criteria are

met. I used a stop criterion of $p < .10$. The remaining predictors construct the final model for the response variable.

For this study, I constructed linear models for nine interaction outcomes. Each model involved eight predictor variables—the four types of gestures and the four gesture-contingent gaze targets—and a response variable that represented an interaction outcome. Each predictor variable represented the percentage amount by which the robot displayed that behavior during its narration, which varied by a small margin from the amount randomly generated for the manipulation due to overlaps and conflicts in gesture production. Only the amounts by which the robot displayed gestures were manipulated. Gesture-contingent gaze cues were drawn from the distributions in the modeling study and were included in the analysis as covariates. I constructed separate models for females and males based on findings from research in human-robot interaction that show strong differences in how females and males perceive and respond to robot behaviors (e.g., Mutlu et al. (2006)), producing a total of 18 models for nine interaction outcomes.

All predictor variables were standardized. A $\log$ transformation was applied to predictor and response variables to obtain linearity for each constructed model. A small number, 0.000001, was added to all predictor variables to avoid the singularity of taking $\log$ of values of zero.

Participants

A total of 32 participants, including 16 females and 16 males, were recruited from the University of Wisconsin–Madison campus community. The average age of the participants was 24.34 ($SD = 8.64$), ranging from 18 to 55. Based on seven-point rating scales, the participants reported that their familiarity with robots ($M = 2.47$, $SD = 1.34$) and their familiarity with the task ($M = 1.78$, $SD = 1.21$) were fairly low.

Results

The regression analysis yielded 15 significant models, categorized into four groups: *task performance*, *perceived performance*, *social and affective evaluation*, and *narration*

	Measure (y)	Gender	Function ($\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$)	R_A^2	Predictor	β	T-test	Significance
Task Performance	Information Recall	**F**	$-.45 + .123 \times Deictic + .202$	.232	*Deictic*	*.123*	$t(14)=2.35$	$p = .034^{*}$
		M	$-.739 + .623 \times Deictic + .335 \times Metaphoric - .549 \times Gaze_{Reference} - .314 \times Gaze_{Gesture} - .195 \times Gaze_{Other} + .393$	.364	*Deictic* *Metaphoric*	*.623* *.335*	$t(10)=3.08$ $t(10)=2.36$	$p = .012^{*}$ $p = .040^{*}$
Perceived Performance	Gesture Effectiveness	**F**	$1.193 + .452 \times Deictic + .252 \times Beat + .464 \times Metaphoric + .224 \times Gaze_{Gesture} - .455 \times Gaze_{Reference} + .373$	.575	*Deictic* *Beat* *Metaphoric*	*.452* *.252* *.464*	$t(10)=3.28$ $t(10)=2.10$ $t(10)=3.53$	$p = .008^{**}$ $p = .062^{\dagger}$ $p = .006^{**}$
		M	$1.215 + .414 \times Beat - .957 \times Gaze_{Listener} - .371 \times Gaze_{Gesture} - 1.05 \times Gaze_{Reference} - .386 \times Gaze_{Other} + .356$	.640	*Beat*	*.414*	$t(10)=4.02$	$p = .002^{**}$
	Competence	**F**	$1.600 + .122 \times Gaze_{Other} + .132$	.440	*No gesture predictor*			
		M	$1.698 + .127 \times Iconic + .117$	.527	*Iconic*	*.127*	$t(14)=4.21$	$p < .001^{***}$
	Naturalness	**F**	$1.454 + .198 \times Metaphoric - .453 \times Gaze_{Listener} - .580 \times Gaze_{Reference} + .183$	.554	*Metaphoric*	*.198*	$t(12)=3.02$	$p = .011^{*}$
		M	$1.529 + .146 \times Iconic - .117 \times Metaphoric + .190$	.333	*Iconic* *Metaphoric*	*.146* *–.117*	$t(13)=2.78$ $t(13)=-2.23$	$p = .016^{*}$ $p = .044^{*}$
Social & Affective Evaluation	Rapport	**F**	$1.212 - .673 \times Gaze_{Listener} - .198 \times Gaze_{Gesture} - .689 \times Gaze_{Reference} + .330$	.121	*No gesture predictor*			
		M	$1.179 + .281 \times Deictic - .608 \times Gaze_{Listener} - .246 \times Gaze_{Gesture} - .718 \times Gaze_{Reference} + .250$	.345	*Deictic*	*.281*	$t(11)=2.54$	$p = .028^{*}$
	Engagement	**F**	$1.575 - .127 \times Metaphoric - .418 \times Gaze_{Listener} - .347 \times Gaze_{Reference} + .174$	.279	*Metaphoric*	*–.127*	$t(12)=-2.04$	$p = .064^{\dagger}$
		M	$1.589 - .120 \times Metaphoric + .242$	.152	*Metaphoric*	*–.120*	$t(14)=-1.92$	$p = .076^{\dagger}$
	Gesture Use	**F**	$-1.229 + .830 \times Metaphoric + .621 \times Gaze_{Listener} + .698 \times Gaze_{Gesture} + .742$	.482	*Metaphoric*	*.830*	$t(12)=3.02$	$p = .011^{*}$
		M	$-.800 + .289 \times Iconic + 1.300 \times Gaze_{Listener} + 1.209 \times Gaze_{Reference} + .244 \times Gaze_{Others} + .468$	.409	*Iconic*	*.289*	$t(11)=2.11$	$p = .059^{\dagger}$
Narration Behavior	Narration Duration	**F**	*No significant model*	*N/A*	*N/A*			
		M	$11.740 + .511 \times Deictic + .204 \times Beat + .331 \times Metaphoric - .305 \times Gaze_{Listener} - .346 \times Gaze_{Gesture} - .819 \times Gaze_{Reference} - .109 \times Gaze_{Others} + .117$	.840	*Deictic* *Beat* *Metaphoric*	*.511* *.204* *.331*	$t(8)=7.50$ $t(8)=5.38$ $t(8)=7.06$	$p < .001^{***}$ $p < .001^{***}$ $p < .001^{***}$

Figure 3.18: Summary of significant models. For each interaction outcome, models for both genders, R_A^2 values, and details of statistical tests for significant gesture predictors are presented. There was no significant model for narration duration for females. (†), (*), (**), and (***) denote $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively. Significance was assessed using two-tailed t-tests. β coefficients are standardized and therefore comparable across gestures.

behavior. These models are shown in Figure 3.18 and illustrated in Figure 3.19. The statistical tests for the models are provided in Figure 3.18, and the paragraphs below provide only a textual description of the results for readability. In each model, the adjusted R-squared score, R^2_A, shows the degree to which the predictor variables account for the variance in the response variable. The standardized β coefficient for each predictor represents the individual effect of the predictor on this variance, and the t-test and p-value summarize the significance of this effect. In this study, $p < .05$ and $p < .10$ were considered as significant and marginal effects, respectively. Note that, because I cannot directly manipulate gesture-contingent gaze cues, I cannot draw conclusions on how they affect interaction outcomes. Hereafter, I only highlight results on gestures.

Task performance. The robot's use of *deictic* gestures significantly predicted information recall for both female and male participants, suggesting that it is important for the robot to point toward references to help the participants ground their understanding of the narration in the references, which in this particular scenario included illustrations of the steps involved in making paper. Additionally, *metaphoric* gestures significantly predicted male participants' recall performance, indicating that they might have leveraged the robot's visualization of abstract concepts such as actions and processes involved in making paper to reinforce their understanding of these concepts.

Perceived performance. Perceived performance involved three aspects: effectiveness of the robot's gestures, perception of competence, and naturalness of the robot's behavior. Females' ratings of the effectiveness of the robot's gestures were significantly predicted by the robot's use of *deictic* and *metaphoric* gestures and marginally predicted by its use of *beat* gestures, while males' ratings were significantly predicted only by *beat* gestures. The robot's use of *iconic* gestures significantly predicted males' perceptions of the robot's competence, while no gestures predicted female perceptions. The robot's use of *metaphoric* gestures positively predicted female participants' perceptions of the naturalness of the robot's behaviors while negatively predicting those of males. On the other hand, *iconic* gestures positively predicted male participants' perceptions of the naturalness of the robot's behaviors.

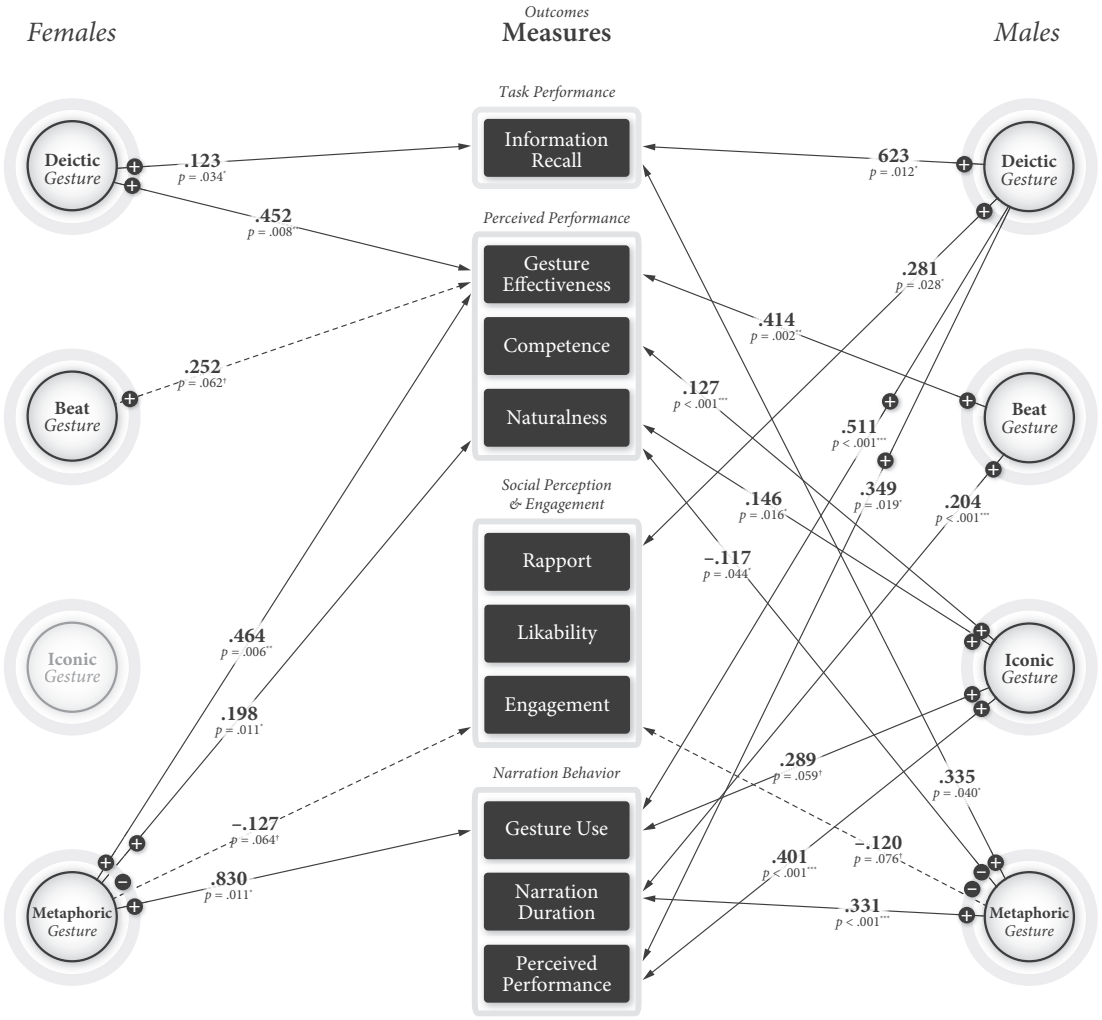


Figure 3.19: A visual summary of the results, highlighting the predictive relationships between gestures and outcomes for females and males. Solid and dashed lines represent significant and marginal effects, respectively. Numbers on lines represent standardized β coefficients and p-values for predictors.

These results suggest that a rich use of different types of gestures might positively shape the participants' overall perceptions of the robot's performance and that employing different gesture strategies might be beneficial to maximize perceived performance outcomes for females and males.

Social & affective evaluation. The robot's use of *deictic* gestures positively predicted male participants' rapport with the robot, while no particular type of gesture predicted this outcome for females. Surprisingly, *metaphoric* gestures marginally but negatively predicted how engaged with the robot males and females reported themselves. These findings indicate that the robot's gestures were less influential on the participants' evaluations of the social and affective characteristics of the interaction, which might be due to the limited interaction afforded by the narrative performance scenario.

Narration behavior. The robot's use of *deictic*, *beat*, and *metaphoric* gestures significantly predicted the length of male participants' retelling of the robot's story. Moreover, male participants who retold the story longer also performed significantly better in the recall test, $\beta = 0.272$, $t(14) = 2.47$, $p = .027$, suggesting that they might have had better recall and thus provided more detail. However, the analysis did not show these relationships for female participants. The robot's use of *metaphoric* gestures significantly predicted how much female participants used gestures during their retelling of the robot's story, while the robot's use of *iconic* gestures marginally predicted males' use of gestures during their retelling. This finding is consistent with the differential effects of metaphoric and iconic gestures on females' and males' perceptions of the robot's performance, suggesting that participants might have employed gestures that they thought were effective.

3.4.4 Discussion

Understanding the relationship between robot gestures and interaction outcomes, particularly how robots might selectively use different types of gesture to improve specific interaction outcomes, promises significant implications for designing robotic systems that not only communicate effectively with their users to maximize

certain outcomes, but also go beyond human capabilities in effective communication. This study serves as a first attempt toward building such an understanding in a narration scenario. To this end, I studied the gestures of human narrators, developed a model for controlling the gestures of narrative robots, and followed a system-level evaluation paradigm to investigate how the robot's use of different types of gestures affected participants' information recall, perceptions of the robot, and ability to retell the robot's story.

The results showed that the robot's use of *deictic* gestures helped improve information recall for both female and male participants. This is particular interesting because deictic gestures are typical gestures that people use to direct others' attention. Therefore, one speculation of this finding was that the robot was able to use deictic gestures to direct participants' attention to key information that it conveyed, thus resulting in a better information retention in participants.

Additionally, the robot's use of *deictic* gestures helped improve perceived effectiveness of the robot's gestures for females, and ratings of rapport with the robot for males. *Beat* gestures positively contributed to the perceived effectiveness of the robot's gestures for both female and male participants, potentially due to their role in signaling key structural information on the discourse. *Iconic* gestures predicted male participants' perceptions of the robot's competence and the naturalness of the robot's behavior and gesture use during their retelling of the robot's story, while not predicting any outcomes for females. *Metaphoric* gestures predicted information recall for males and perceptions of the naturalness of the robot's behavior and the effectiveness of its gestures for females. Interestingly, metaphoric gestures negatively predicted the participants' engagement with the robot, indicating that more gestures do not necessarily mean better outcomes. We speculate that the abstract content and the large number of arm motions involved in this type of gesture might have been a distraction for the participants. On the other hand, these gestures contributed to the participants' ability to retell the robot's story, positively affecting narration length for males and gesture use for females.

Design implications

These findings highlight the importance of robot gestures in shaping key outcomes in human-robot interaction, such as the ability to recall and retell information and perceptions of the performance of the robot and the social and affective characteristics of the interaction. Interaction designers and roboticists must leverage the design space for gestures to develop applications that maximize desired outcomes, such as increasing the use of deictic gestures by an instructional robot to improve student learning. Designers might also adapt the robot's use of the different types of gestures to the specific goals of the interaction and to user gender. For instance, the use of metaphoric gestures might be decreased if the application seeks to increase user engagement with the robot and increased if user ability to recall and retell information is important. Similarly, the robot might employ a different balance of metaphoric and iconic gestures in its interactions with females and males.

Limitations

The work presented here has two main limitations. First, the findings, such as the relative effects of different types of gestures on the measured outcomes, might have limited generalizability beyond the specific context of the study, requiring further work to establish the extent to which they generalize to other forms of interaction, cultural settings, and individuals with varying abilities to perceive and interpret non-verbal social cues. I expect the research approach presented here to provide the methodological basis for such future work and other research into understanding complex behavior-outcome relationships. Second, while the robot platform used in this work offered the expressivity needed to achieve research goals, hardware platforms that afford articulate hands and higher degrees of freedom would enable richer and more finely controlled gestures and thus a more thorough understanding of how robot gestures shape human-robot interaction.

3.5 Study 3: learning multimodal behaviors for effective communication

Study 1 showed that people could leverage the robot's properly timed gaze cues to augment their abilities in locating relevant task objects and remembering delivered information. Moreover, Study 2 revealed that the robot's gestures, particularly deictic gestures, aided in people's retention of information. Together, the results from the previous two studies provided evidence showing that gaze and gestures are effective ways to manipulate people's attention and to support verbal communication.

In this study, I focused on studying how *gaze* and *gestures* can be meshed together as coherent behaviors to support effective communication. In particular, I explored a learning-based approach to model the coupling relationships among speech, gaze, and gestures in human communication (Section 3.5.1). The learned model was implemented for a robot to perform a narration task (Section 3.5.2) and evaluated in a laboratory human-robot interaction experiment (Section 3.5.3). Figure 3.20 shows a visual summary of the investigation process of this study.

3.5.1 User modeling

The same data as collected in Study 2 (Section 3.4.1) was used in this study to model how speech, gaze, and gestures are coupled in time. In the following paragraphs, I provide details on the modeling technique used in this study and the resulting model of multimodal behavior.

Modeling multimodal behavior

I utilized a dynamic Bayesian network (DBN) to model the temporal relationships among behaviors. An DBN is a type of probabilistic graphical model (PGM) that provides compact representations of conditional independence among random variables (Kollar and Friedman, 2009). DBNs generalize hidden Markov models (HMMs) to represent the hidden state and the observation as a set of random

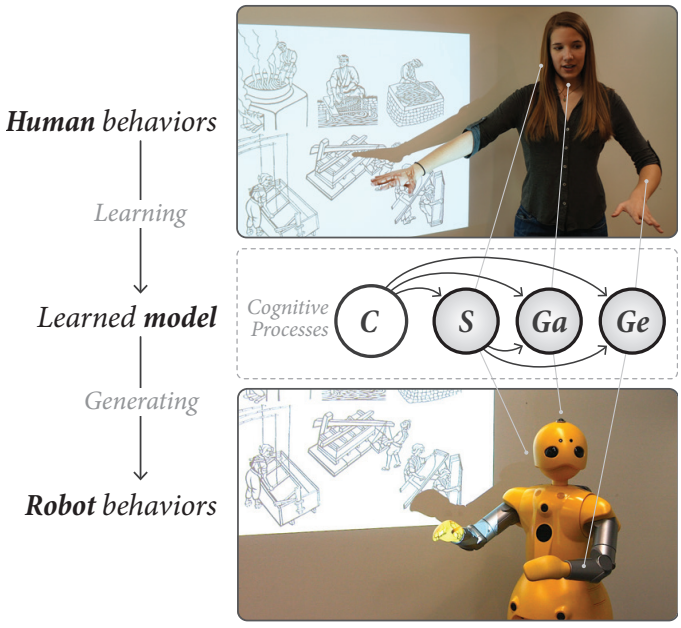


Figure 3.20: We used a learning-based approach to *model* how humans employ multimodal behaviors involving speech, gaze, and gestures during narration and *generate* multimodal behaviors for a humanlike robot to perform the same narration task.

variables. In their basic form, DBNs are directed acyclic graphs in which nodes represent random variables and edges represent conditional dependencies (e.g., Figure 3.22). Semantically, an edge from a parent node A to a child node B means that node A has influence over node B. A dynamic Bayesian network extends static Bayesian nets (BNs) to incorporate the temporal dependencies among variables. These characteristics for dealing with *uncertain* and *temporal* relations among random variables make dynamic Bayesian networks particularly useful in modeling the dynamics of multimodal behaviors. Murphy (2002) provides an extensive introduction to representation, learning, and inference in DBNs.

Model representation. Informed by literature in human communication (Henderson et al., 1966; Jaffe et al., 1964; McNeill, 1992), I proposed a network structure, shown in Figure 3.22, to represent the relationships among speech, gaze, and gestures. In developing this network, I included a *hidden random variable* denoting a

Gesture Type	Speech Features		
Deictic gestures	Concrete reference *"a big pot"*	Abstract reference *"the first step"*	Pronoun *"this person"*
Iconic gestures	Concrete object *"two boards"*	Descriptive verb *"peel it off"*	Non-descriptive action *"make it"*
Metaphoric gestures	Abstract concept *"for six hours"*	Abstract process *"how paper is made"*	Abstract object *"the water soluble elements"*
Beat gestures	Important information *"at least ten times of water"*	New information *"for example"*	Connector *"so that"*

Figure 3.21: Speech features for gestures. Features were identified through affinity diagramming of human data and by using the guideline suggested by McNeill (1992). An example of each type of feature is also provided.

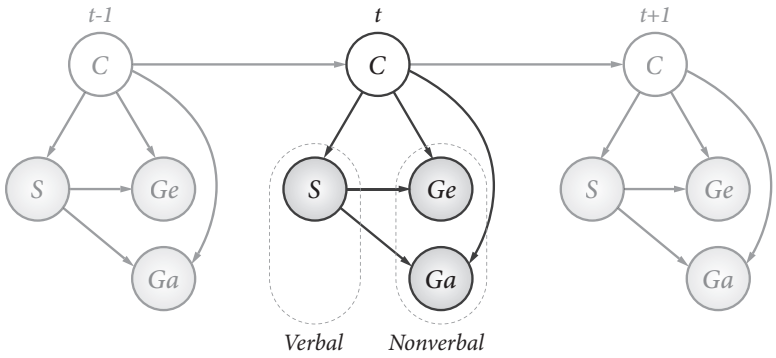


Figure 3.22: The proposed dynamic Bayesian network for modeling and generating multimodal behaviors. C denotes a latent cognitive process that directs verbal, involving speech (S), and nonverbal, involving gaze (Ga) and gestures (Ge), processes.

cognitive process (C) that directs how humans coordinate speech (S), gaze (Ga), and gestures (Ge), which were considered *observations*. I assumed the latent cognitive process was a discrete-time Markov process. This assumption is consistent with psycholinguistic models of speech production (Henderson et al., 1966; Jaffe et al., 1964). Additionally, I assumed that speech influences gaze and gestures, as research suggests that nonverbal behaviors might be contingent on verbal utterances (McNeill, 1992). Based on my exploratory tests, I empirically determined there were three hidden states and that the discrete time window of the Markov process was 500 ms.

Learning & inference. To learn the parameters of each conditional probability distribution (CPD) in the DBN (Figure 3.22), the expectation-maximization (EM) algorithm (Dempster et al., 1977) was used. The eight coded episodes of interaction were used as the training data.

To control the robot's behaviors using the learned DBN, I assumed that the robot's speech features would be given and the most probable gesture type and gaze target would be inferred at any given time t. To this end, I used a junction tree algorithm to perform offline smoothing (Murphy, 2002) and compute the most probable explanation (MPE)—the maximal posterior probability of a set of variables given observations of another set of variables—of gaze and gestures. I used the Bayes Net Toolbox (Murphy et al., 2001) for learning and inference.

The latent cognitive process contained three states ($C = c_i, i = 1, 2, 3$). Gaze and gestures contained four ($Ga = a_i, i = 1, ..., 4$) and five ($Ge = e_i, i = 1, ..., 5$) values, respectively. Speech (S) was represented by 12 boolean variables, each of which corresponded to a speech feature (Figure 3.21). As a result, the model is characterized by a vector of 15 discrete values at each time step. Given the speech features, the most probable latent cognitive state (C_t), gesture type (Ge_t), and gaze target (Ga_t) at any given time t over the duration of the speech ($S_{1:T}$) can therefore be computed using Equation 3.1. Here, X represents C, Ga, and Ge, and x represents c, a, and e.

$$\arg\max_i \; p(X_t = x_i | S_{1:T}) \qquad \forall t \in T \tag{3.1}$$

The features of the robot's speech were manually marked, and the annotated speech features were discretized into feature sequences at the rate of 500 ms per feature. During inference, the sequences of speech features were used as partial observations, and gaze and gesture were treated as missing values. The consecutive states inferred for each behavior were combined into a sequence of behavior that was considered to be one continuous instance of behavior. For example, six consecutive states of iconic gestures were combined into one iconic gesture lasting 3,000 ms.

Model evaluation

An evaluation of the model sought to determine to what extent the DBN model, as illustrated in Figure 3.22, accurately predicted gaze targets and gesture types when given speech features by comparing gaze and gestures predicted by Equation 3.1 against the human data. The evaluation was conducted using eight-fold leave-one-out cross validation. I used gaze and gesture data from seven dyads to train the DBN and evaluated the performance of the trained model in predicting gaze targets and gesture types, given speech features from the eighth dyad. While the behaviors are continuous in time, I chose to discretize the data using a window 500 ms and to compare the predicted behavior to the behavior in the test dataset at each discrete window. The accuracy of prediction was on average 54.57% for gaze targets and 62.77% for gesture type. While prediction accuracy is significantly better than chance (25% for gaze and 20% for gestures), it might be further improved by considering more features. Moreover, the highest and lowest accuracies in the cross validation for gaze were 68.23% and 46.48%, respectively, and for gestures were 69.76% and 44.25%, respectively. These results suggested a large variation in how people employed behaviors during narration. Collecting more data to train the model might also improve predictive accuracy.

3.5.2 System implementation

I conceptualized the problem of generating multimodal robot behaviors into two levels—the *feature* level and the *domain* level. Figure 3.23 illustrates this conceptualization for the process of generating speech, gaze, and gestures. The feature

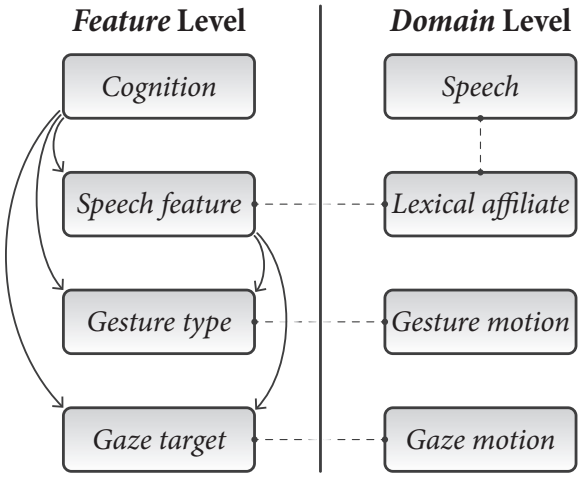


Figure 3.23: Our conceptualization of the process of generating speech, gaze, and gestures. Learning and inference for high-level features are performed at the *feature* level, while specific motions are defined at the *domain* level.

level represents high-level behavioral features of the target channel, such as "iconic gesture" for gesture type and "listener" for gaze target. At the domain level, behavioral features are associated with specific motions, such as specific arm motions for an iconic gesture and gaze shifts toward the listener. This separation modularized the problem space and allowed for development of and improvement in different components in isolation.

The learned model of multimodal behavior was operated at the feature level to inform what gestures and gaze cues to use given the robot speech. I then employed simple mechanisms to bridge the feature and domain levels to generate actual robot behaviors. At the *domain* level, the robot's gestures were designed based on my observations of human narrators' gestures. The robot gestures were the same as those in Study 2 (e.g., created though puppeteering gestural movements).

To generate robot behaviors, simple mechanisms were used to link components at the feature and domain levels. Lexical affiliates in the robot's speech were manually annotated and tagged with gesture types. The mechanism for linking inferred gesture types to actual robot gestures functioned as follows. For each

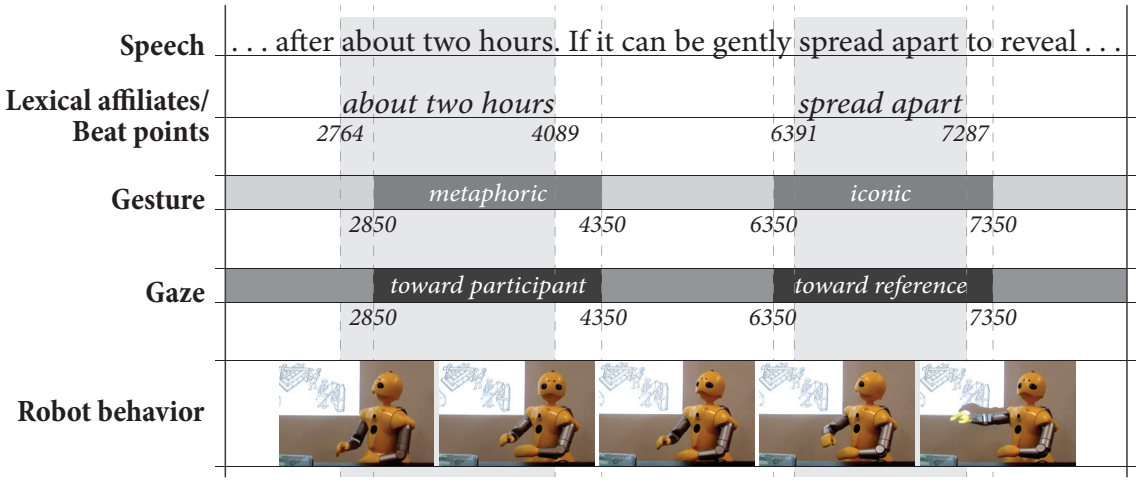


Figure 3.24: An example of the robot displaying speech, gaze, and gesture behaviors generated by the proposed learning-based approach.

inferred gesture, I first checked whether the robot's speech included lexical affiliates that temporally overlapped with the gesture within a window of 2,000 ms—1,000 ms before and 1,000 ms after the beginning of the gesture. No lexical affiliates or a lexical affiliate for a gesture type that did not match with the inferred gesture type within this window prompted the robot to randomly select and perform a gesture from the gesture library of the inferred gesture type. If the gesture type of the found lexical affiliate was consistent with the inferred gesture type, the lexical affiliate was linked to an actual gesture using an association mechanism.

Speech and inferred gaze and gesture motions were synchronized using the Robot Behavior Toolkit (Section 3.3.2). Figure 3.24 illustrates a sample sequence of speech, gaze, and gesture behaviors that were controlled by the proposed learning-based approach.

3.5.3 Experimental evaluation

In a laboratory experiment, I evaluated how robot behaviors produced by the learning-based approach shaped outcomes of human-robot interaction. The evalua-

tion sought to test the central hypothesis that robot behaviors produced by learned parameters would improve interaction outcomes, specifically improved perceptions of the robot in terms of immediacy, naturalness, effectiveness, likability, and credibility and improved ability of the participants to retell the robot's story, over baseline behaviors such as behaviors produced by randomly generated parameters or no behaviors, while resulting in outcomes that are comparable to those elicited by conventional modeling approaches (as used in Study 2).

Experimental design, task, & conditions

The evaluation study used the same narration task as the one used in Study 2. I manipulated the method used to control the robot's gaze and gesture behaviors while keeping the robot's speech the same across conditions. I designed a between-participants study, in which each participant was randomly assigned to one of the following conditions:

1. In the *learning-based* condition, the learned DBN described in Section 3.5.1 directed the robot's gaze and gestures.
2. In the *unimodal* condition, the robot only used the speech channel to verbally present the narration topic without gaze or gesture behaviors. This condition served as the experimental control.
3. In the *random* condition, the same network structure as described in the previous section was used to generate behaviors. However, the model instead used randomly generated parameters, introducing temporal and spatial randomness in the produced behavior.
4. The *conventional* condition involved directing the robot's gaze and gestures based on designer-specified parameters for aligning behaviors. In particular, the parameters specified how different types of gestures were aligned with speech features (i.e., lexical affiliates) according to the literature on human communication (McNeill, 1992). The parameters also specified gaze-gesture relationships by extracting distributions representing where a speaker looked while performing different types of gestures. Every time its gesture state

changed, the robot determined its new gaze target based on these distributions. These parameters were used in Study 2 to direct the robot's behavior.

Procedure

After obtaining informed consent, the experimenter directed participants to a controlled laboratory environment, where each participant listened to the robot narrate the process of making paper (Figure 3.17). The robot's narration lasted approximately six minutes. Following the narration, participants completed a distractor task that lasted approximately five minutes and then took a quiz on the presented information. Participants then retold the narration and filled out a post-experiment questionnaire that evaluated their perceptions of the robot. Participants received $5 for their participation. The entire experiment took approximately 30 minutes per participant.

Participants

A total of 29 participants (16 males, 13 females), whose ages ranged 18–38 ($M = 22.62, SD = 4.35$), were recruited from the University of Wisconsin–Madison campus. There were six, eight, seven, and eight participants in the learning-based, unimodal, random, and conventional conditions, respectively. Participants reported relatively low familiarity with robots ($M = 2.48, SD = 1.53$) and with the process of making paper ($M = 1.69, SD = 1.11$) in seven-point rating scales.

Measures

I used a post-experiment questionnaire to evaluate the participants' perceptions of the robot's behavior. Four measurement scales were developed using seven-point questionnaire items. *Immediacy*, defined as psychological distance between individuals (Mehrabian, 1971), assessed how close participants felt to the robot and how engaging they thought the robot was (3 items, Cronbach's $\alpha = .79$). *Naturalness* gauged how natural the robot's motions were (5 items, Cronbach's $\alpha = .84$). *Effectiveness* measured how participants perceived the robot's effectiveness as a presenter (4 items, Cronbach's $\alpha = .87$). *Likability* evaluated how likable the

robot was (8 items, Cronbach's $\alpha = .88$). An additional item measured the robot's *credibility* using a question about whether or not the robot provided sufficient information for the participant to answer the quiz questions. Moreover, I asked participants to choose from a list of 20 adjectives to describe the robot's overall behavior. The list of 20 adjectives consisted of 10 positive and 10 negative adjectives. Two manipulation-check items were included to ensure that manipulations in the trained model to create the unimodal and random conditions were successful.

In addition to questionnaire evaluation, I evaluated participants' performance in retelling the information that the robot presented, calculating measures of *familiarity* with the narration topic, their use of *body language*, and an *overall evaluation* of presenter effectiveness. Three raters who were blind to the experimental conditions rated video recordings of the participants' retelling performance on these measures using seven-point scales. Inter-rater reliability analysis using intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979) as a measure revealed high correlations among the three raters on measures of familiarity of content (ICC(3,3)=.884), use of body language (ICC(3,3)=.897), and overall evaluation (ICC(3,3)=.771).

I also included a measure that involved a quiz consisting of 18 questions about the content of the robot's story. This measure explored whether or not the manipulations in the robot's behaviors affected participants' recall of the narrated information.

Results

One-way fixed-effects analysis of variance (ANOVA) tests, using the manipulation in the robot's behaviors as the fixed factor, were conducted to analyze the manipulation checks, questionnaire measures, quiz data, and retelling evaluation. Planned many-to-one multiple comparisons used the Dunnett's method, considering the learning-based approach as the comparison baseline, to assess how the unimodal, random, and conventional conditions compared against the learning-based condition. Additionally, I performed Dunnett's tests, considering the unimodal and random conditions as baselines, using data on manipulation checks to verify whether the learned model was successfully manipulated to create

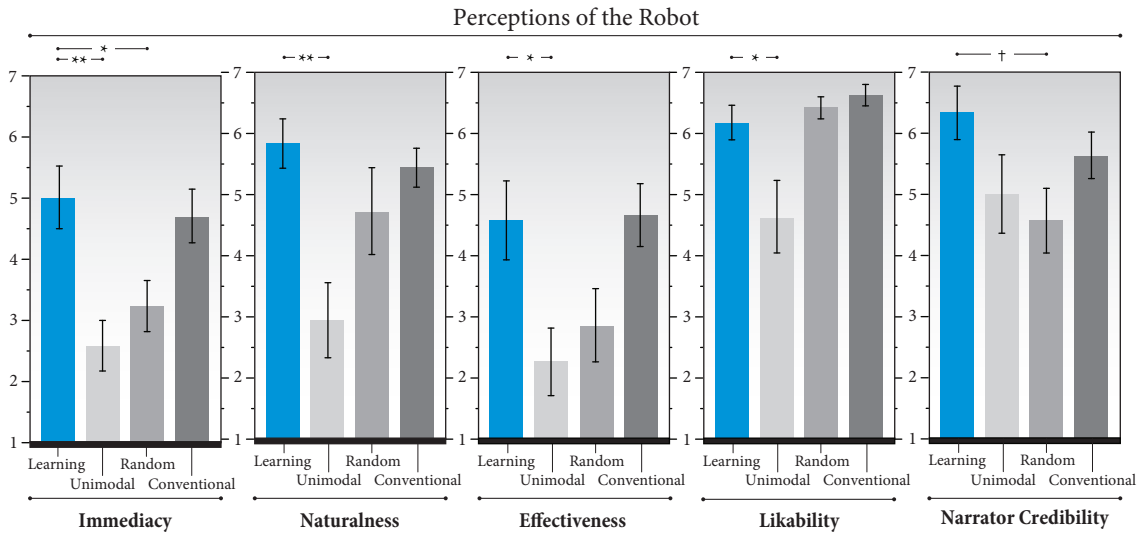


Figure 3.25: Results on perceptions of the robot and retelling performance. Only significant results are marked. (NS), (†), (*), (**), and (***) denote $p > .10$, $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively.

the unimodal and random conditions. To determine whether behaviors produced by the learning-based and conventional approaches are comparable, we followed guidelines suggested by Walker and Nowacki (2011) and Julnes and Mohr (1989), applying a conservative equivalence margin of 0.50 (i.e., $p > .50$) to the comparisons between these two conditions.

Manipulation checks. To check whether our manipulations to create the unimodal behavior condition were successful, I asked participants whether or not they perceived the robot to be motionless. The analysis of variance found a significant effect of our manipulation on this measure, $F(3, 25) = 22.88, p < .001$. Comparisons using the Dunnett's test showed that participants in the unimodal condition perceived the robot to be more motionless than those in the learning-based, $p < .001$, random, $p < .001$, and conventional, $p < .001$, conditions. I also asked participants whether the robot's motions appeared random to verify that we successfully created the random condition. The analysis found a significant effect of our manipulation on this measure, $F(3, 25) = 6.12, p = .003$. Comparisons showed that participants in

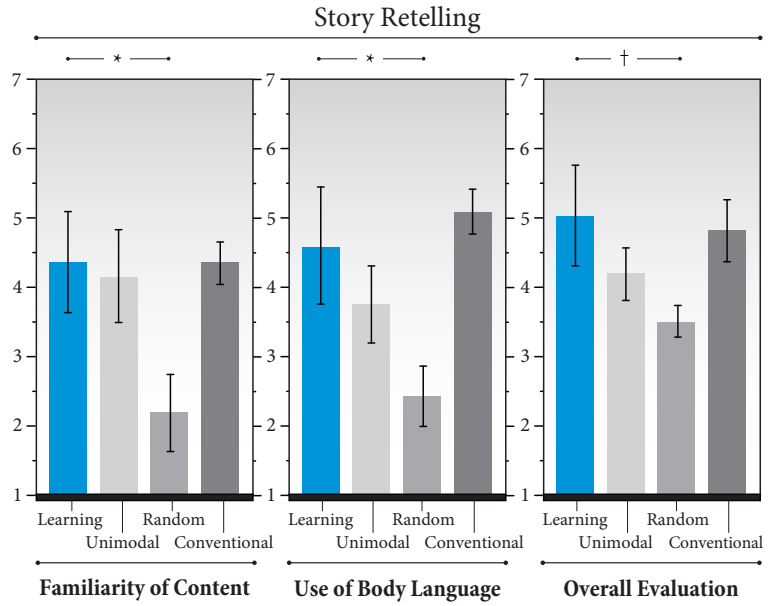
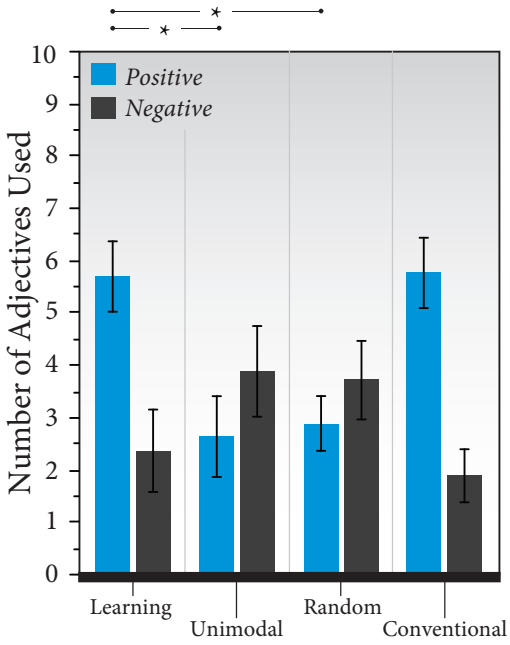


Figure 3.26: Results on perceptions of the robot and retelling performance. Only significant results are marked. (NS), (†), (*), (**), and (***) denote $p > .10$, $p < .10$, $p < .050$, $p < .010$, and, $p < .001$, respectively.

the random condition perceived the robot's motions to be more random than those in the learning-based, $p = .005$, unimodal, $p = .017$, and conventional, $p = .002$, conditions did.

Perceptions of the robot. I found that our manipulation had a significant effect on the perceived immediacy, $F(3, 25) = 6.91, p = .002$, naturalness, $F(3, 25) = 5.77, p = .004$, effectiveness, $F(3, 25) = 4.59, p = .011$, and likability, $F(3, 25) = 6.44, p = .002$, of the robot, but not on participants' perceptions of the robot's credibility, $F(3, 25) = 2.07, p = .130$. Comparisons further revealed that the learning-based and conventional approaches showed equivalence in measures of perceived immediacy, $p = .923$, naturalness, $p = .917$, effectiveness, $p = .999$, likability, $p = .722$, and credibility, $p = .645$.

Comparisons also showed that participants in the learning-based condition rated the robot to have higher immediacy, $p = .002$, and to be more natural, $p = .003$, more



	Positive	*Negative*
Learning-based (6) *Coherent (5)* *Scripted (5)* *Communicative (5)* *Confident (5)*	**70.83%**	**29.17%**
Unimodal (8) *Rigid (6)* *Precise (5)* *Robotic (6)* *Scripted (5)* *Artificial (5)*	**40.38%**	**59.62%**
Random (7) *Artificial (5)* *Confident (5)* *Robotic (5)*	**43.48%**	**56.52%**
Conventional (8) *Coherent (7)* *Lively (6)* *Confident (7)* *Scripted (6)* *Communicative (6)*	**75.41%**	**24.59%**

Figure 3.27: Results on participants' use of positive and negative adjectives in describing the robot's behavior. Top three choices of adjectives and percentages of used positive and negative adjectives are listed. Values in parentheses indicate how many participants used the adjective to describe the robot. Only significant results are marked. (*) denote $p < .050$.

effective, $p = .027$, and more likable, $p = .023$, than those in the unimodal condition. However, no significant differences were found between the learning-based and unimodal conditions in perceived credibility, $p = .191$. While the comparisons showed that participants in the learning-based condition perceived the robot to have higher immediacy, $p = .032$, and marginally more credibility than in the random condition, $p = .072$, no differences were found between the learning-based and random conditions in measures of naturalness, $p = .378$, effectiveness, $p = .131$, and likability, $p = .931$. Figure 3.25 summarizes these results.

The analysis also found a significant effect of our manipulation on the number of positive adjectives that participants used to describe the robot's behaviors, $F(3, 25) = 6.44, p = .002$. Comparisons showed that participants in the learning-based condition used more positive adjectives than those in the unimodal, $p = .013$,

and random, $p = .027$, conditions did. The learning-based and conventional conditions demonstrated equivalence in the use of positive adjectives, $p = .999$ (Figure 3.27) . However, the manipulation did not have a significant effect on the number of negative adjectives used, $F(3, 25) = 1.86, p = .162$. Comparisons showed that participants in the unimodal, $p = .354$, random, $p = .460$, and conventional, $p = .947$, conditions used a similar number of negative adjectives in describing the robot's behavior to those in the learning-based condition did. Figure 3.27 shows the top three adjectives used to describe the robot's behaviors.

Retelling performance. I found a significant effect of the manipulation on participants' perceived familiarity with the narration topic, $F(3, 25) = 3.37, p = .034$, and effective use of body language, $F(3, 25) = 4.67, p = .010$, but not on the overall evaluation of the participant as an effective presenter, $F(3, 25) = 2.16, p = .118$. Comparisons revealed equivalence between the learning-based and conventional approaches with respect to familiarity of the topic, $p = 1.000$, effective use of body language, $p = .848$, and overall evaluation of presenter effectiveness, $p = .977$, as shown in Figure 3.26.

Comparisons also showed that participants in the learning-based condition were rated to be more familiar with the topic, $p = .042$, and to more effectively use body language, $p = .033$, than those in the random condition. They were also rated marginally higher in overall evaluation than those in the random condition, $p = .084$. However, comparisons did not find significant differences in participants' familiarity with the topic, $p = .986$, effective use of body language, $p = .569$, and overall performance as an effective presenter, $p = .447$, between the learning-based and unimodal conditions (Figure 3.26).

Cognitive assessment. The manipulation in the robot's behaviors did not have a significant effect on participants' information recall in the quiz, $F(3, 25) = 0.98, p = .418$. No significant differences were found between the learning-based condition and the unimodal, $p = .621$, random, $p = .866$, and conventional, $p = .986$, conditions in comparisons.

3.5.4 Discussion

Overall, the results showed that participants in the learning-based condition rated the immediacy, naturalness, effectiveness, and likability of the robot higher than those in the unimodal condition did. Participants in the learning-based condition were also rated higher in their retelling of the narration topic compared with those in the random condition. Additionally, the learning-based and conventional conditions showed equivalence in participants' perceptions of the robot and retelling performance.

While behaviors in the unimodal condition negatively affected how participants perceived the robot in terms of immediacy, naturalness, effectiveness, and likability, it did not affect participants' perceived credibility of the robot, retelling performance, or information recall. The following excerpts from post-experiment interviews provide some insight into these results:

> *"The robot didn't particularly move, which I didn't think was realistic, and also it didn't make eye contact with me."*

> *"When it switched steps, it would be more engaging if it would [like] point to the new step..."*

> *"I was more focused on what the robot was saying and the screen..."*

While participants in the unimodal condition noticed that the robot's behaviors were not natural, they still paid attention to the verbal information and the projected illustration to understand the presented information, demonstrating comparable results in perceived credibility, retelling of the topic, and learning.

Participant comments also provided insight into why the random manipulation did not affect their perceptions of the robot but affected their retelling performance, as illustrated in the excerpts below:

"The motion was kind of distracting ... sometimes maybe it used hand gestures in a way that I wouldn't necessary to use that ... [but] I feel like people do that sometimes too..."

"It was very distracting... what are you [the robot] trying to do, I don't know what you [the robot] are trying to do. It was just odd."

Most participants in the random condition found the robot's behaviors to be distracting, which suggests that participants may have found it difficult to focus on the information presented by the robot, potentially leading to poor retelling performance. The discrepancy between quiz results and topic familiarity in retelling may reflect the effects distraction may have had on their learning.

However, while some of the participants in the random condition perceived that the robot's behaviors deviated from social norms, others found these behaviors to be acceptable. These results are consistent with previous work on robots' use of gestures to support speech, which found no differences in participants' ratings of a robot displaying gestures that were semantically incongruent with its speech and one that displayed congruent behaviors in measures of how lively, active, engaged, communicative, and fun-loving they perceived the robot to be (Salem et al., 2012).

The participants perceived the robot in the learning-based condition as showing higher immediacy—displaying more engagement and psychological closeness—than they did in the unimodal and random conditions. This result is consistent with research in education, which shows that instructors with high immediacy use more gestures and greater eye contact with students (Mehrabian, 1971).

Finally, the results suggested that the behaviors driven by the learning-based and conventional approaches demonstrated similar effectiveness in every measure. Participant comments also highlight similar limitations both approaches have in generating natural, humanlike behaviors, as suggested by the following comments on the robot's gaze behavior, the first by a participant in the learning-based condition and the other two from participants in the conventional condition:

"It seemed not to sometimes make eye contact when I expected it to. There were a few periods where it was talking and talking straight at the screen and not making eye contact."

"What it [the robot] could do more is to look as if it is an actual human were presenting. It could look more at the screen [be]cause that's how I feel a lot people do."

"There were a few times I noticed it pointed to the screen without looking at it."

These comments also highlight the variability in people's expectations of the behaviors of an effective presenter and suggest that gaze behaviors controlled by the learning-based and conventional approaches will require further improvement to meet the expectations of a broader population of users.

3.6 Summary

Joint attention—the process brings interaction partners to a shard attentional focus—serves as the basic mechanism in joint action to establish perceptual common ground. This chapter presents a series of studies to investigate how a robot can use multimodal behaviors, including gaze cues and gestures, to manipulate its human partner's attention to achieve effective communication. In particular, Study 1 showed that a robot using referential gaze cues that were temporally coupled with speech references based on findings in human communication can effectively reorientate its partner's attention to locate task-relevant objects in a shorter time and to remember details of the delivered information (Section 3.3). Study 2 revealed that a robot using deictic gestures pointing toward a shared visual reference more with its partner improved the partner's retention of the information delivered by the robot. Additionally, the robot's uses of various gestures shaped the partners' perceptions of the robot and their interaction experiences, showing that gestures are not only communicative but also affective. Study 3 provided insights into how

a robot might integrate gaze cues and gestures coherently to produce engaging behaviors while facilitating users in acquiring information intended for them.

In addition to the exploration of how different behavioral cues might serve devices for establishing joint attention, the three studies produced methodological innovations for modeling, generating, and evaluating social behaviors for robots. Specifically, Study 1 introduced the *Robot Behavior Toolkit*, an implementation of a repertoire of robot behaviors, to allow for generating coherent social behaviors for robots. Study 2 proposed a *multivariate evaluation* approach to systematically explore how directly manipulated behaviors in a robot might contribute to outcomes of human-robot interaction. Study 3 demonstrated a data-driven approach using a dynamic Bayesian network to jointly model the temporal dynamics among multiple social behaviors. These methodological innovations serve as design tools to facilitate the process of "expression" (i.e., helping users to understand the robot) in joint action (Figure 2.1).

Finally, the three studies presented in this chapter together showed the importance of joint attention via the use of multimodal behaviors in joint action. This basic mechanism for manipulating attention to establish perceptual common ground allows for more advanced mechanisms of joint action such as action observation (Chapter 4) and task-sharing & action coordination (Chapter 5).

4 ACTION OBSERVATION: FROM OBSERVATION TO PREDICTION TO REACTION

Action observation allows for the monitoring of a partner's actions in preparation of one's own actions. In this chapter, I demonstrate how the mechanism of action observation might be realized for robots to facilitate joint action with humans. Particularly, I focus on using users' gaze patterns to infer their intentions and how a robot can leverage the inference of user intentions to proactively prepare and execute actions in a joint task.

4.1 Introduction

To achieve successful joint action, people observe each other's actions and task progress, predict each other's intentions, and adjust their own actions accordingly (Sebanz and Knoblich, 2009). Such action observation and intention prediction are integral to the establishment of common ground between parties engaged in joint action. As a result, parties consciously and subconsciously exhibit behavioral cues, such as eye gaze and gestures, to manifest intentions for others to read (i.e., the process of "expression" in Figure 2.1) while interpreting others' behavioral cues to understand their intentions (i.e., the process of "prediction" in Figure 2.1). These behavioral cues are a gateway to understanding a person's mental states, including attention, intentions, and goals. Increasing evidence from neuroscience and developmental psychology has shown that action observation allows people to use their behavior repertoire and motor system to predict and understand others' actions and intentions (Blakemore and Decety, 2001; Buccino et al., 2001; Rizzolatti and Craighero, 2004).

Among other behaviors, gaze cues are particularly informative in the manifestation of mental states. As presented in the previous chapter, deictic gaze toward an object may signal the person's interest in the object and has been found to be temporally coupled with the corresponding speech reference to the object (Griffin,

2001; Meyer et al., 1998). Moreover, people use gaze cues to draw others' attention toward an intended object in the environment in order to establish perceptual common ground (Study 1). The ability to understand and monitor such cues is critical for sharing mental states (Butterworth, 1991) as well as completing joint tasks (Tomasello, 2009). Furthermore, gaze cues may also signal planned actions; empirical evidence has shown that gaze cues indicate action intent and lead motor actions that follow (Johansson et al., 2001; Land et al., 1999).

This present study aimed to develop a model quantifying how patterns of gaze cues may characterize and even predict intentions and to investigate how a robot may proactively prepare its actions in response to the predicted user intentions. To this end, I first built a support vector machine (SVM) to model a potential predictive relationship between gaze patterns and users' task intent in a dyadic joint task[1]. I then developed an autonomous interactive system in which a robotic manipulator could anticipate its human partner's intentions using a SVM-based predictive model and responded to the partner proactively. Finally, the developed system was evaluated in a human-robot interaction study to assess system performance and user experience[2].

4.2 Related work

Successful recognition of their partner's intent enables each person to adapt their behavior to accommodate their partner's intentions in performing joint action. In this section, I review prior research relevant to this present study, particularly focusing on human intent, communication of intention, and enabling anticipatory robot actions.

[1]Study results presented in this chapter were also published in Huang et al. (2015a).

[2]A scientific report of this human-robot interaction study has been submitted to the 11th international conference on human-robot interaction (HRI'16).

4.2.1 Human intent

The concept of intentionality is defined as the commitment of a person to executing a particular action (Malle and Knobe, 1997). The formulation of an intent is often driven by the individual's desire to achieve a particular goal (Astington, 1993). From an early age, children begin to attribute intent to the actions of others. For example, children at 15 months of age are capable of understanding the intentions of others in physical tasks, even when the goal is not achieved (Meltzoff, 1995). Prior work suggests that, after developing a capacity for understanding intent, humans also develop *Theory of Mind* (ToM)—the ability to attribute mental states to others (Leslie, 1987). ToM then shapes the way people interact with one another in a way that is most easily observable in physical tasks, such as moving a table together or navigating through a crowd. In these scenarios, humans rely on ToM abilities to attribute intent to other participants and to adapt their own behaviors to accommodate the intent of others, resulting in seamless interactions.

Robots that seek to engage in seamless joint action with people must also read and predict their human partners' intentions and adjust their actions accordingly.

4.2.2 Behavioral communication of intent via gaze

One approach people subconsciously use to infer the intent of others is by observing their behavioral cues (Blakemore and Decety, 2001). Humans employ a number of behavioral cues, such as gaze and gestures, when working with others on a task (Shibata et al., 1995; Bangerter, 2004; Clark and Brennan, 1991; Morris and Desebrock, 1977; White, 1989; Meltzoff and Brooks, 2001; Baron-Cohen et al., 2001). These cues aid in their partner's understanding of and fluency in the task, enabling their partner to adjust their behavior accordingly to accommodate intended actions (Blakemore and Decety, 2001). While a number of behavioral channels can be used to understand intent, gaze is considered preeminent among them due to the clarity with which it can indicate attention; for instance, partners would assume that an area being gazed toward will be the next space to be acted upon (Meltzoff and Brooks, 2001; Baron-Cohen et al., 2001).

Gaze behavior is crucial to human communication of intent throughout the development of social behavior. During infancy, children can follow the gaze cues of adults, which serve as the basis of joint attention (Butler et al., 2000), and use their own gaze to communicate an object of interest (Morales et al., 1998). The use and understanding of gaze become more complex and nuanced with age, allowing humans to better identify targets of joint attention (Heal, 2005). This development of gaze understanding mirrors the development of understanding of intent and ToM discussed above, allowing humans to gradually develop a more complex intuition of others and their intentions.

During a joint task, awareness of a partner's gaze behavior helps enable effective task coordination between participants (Tomasello, 1995). Prior work by Brennan et al. (2008) used head-mounted eye trackers to examine gaze patterns during a joint search task. Awareness of a partner's gaze behavior was not only sufficient for completing the task, but it also resulted in significantly faster search times than verbal coordination did. Additionally, participants who were aware of their partner's gaze behavior offered more precise help during the task when it was necessary.

Ordering of gaze fixations has been used to infer the type of visual task a person is performing, such as memorizing a picture versus counting the number of people photographed in a picture (Haji-Abolhassani and Clark, 2014). Prior work used eye gaze and its associated head movements as input for a sparse Bayesian learning model (McCall et al., 2007) to predict a driver's future actions when operating a motor vehicle (Doshi and Trivedi, 2009). Additionally, work by Yi and Ballard (2009) built a dynamic Bayesian network from a user's gaze and hand movements to predict their task state in real time during a sandwich-making task.

While prior work has examined the connection between gaze and intent in a variety of situations, the current study aims to provide an empirical approach to modeling gaze behavior to predict task intent during collaboration. Specifically, it extends prior work in two ways. First, the current study investigates the relationship between gaze cues and task intent in a collaborative context, whereas prior work employed tasks that involved only one person completing them, e.g., making a

sandwich (Yi and Ballard, 2009) or driving a car (Doshi and Trivedi, 2009). Second, the prior predictive models utilized multiple sources of information, while this present study focuses on using gaze cues only.

4.2.3 Anticipatory robot actions

Increasing evidence from neuroscience has shown that people observe and predict others' actions and intentions (Frith and Frith, 2006; Walter et al., 2004). As a result, it has been argued that people leverage the processes of observation and prediction to proactively prepare their goal-directed actions (Pezzulo and Ognibene, 2012). Drawing on these scientific findings, prior work in human-robot interaction and robotics has explored how robots might anticipate people's actions and intentions to facilitate effective joint action with them. Towards safer and efficient human-robot collaboration, Mainprice and Berenson (2013) used observed human motion to predict workspace occupancy and plan robot motion accordingly to minimize interference with humans. Their approach involved both offline model building and online prediction of human motion. The approach was tested in simulation to predict which target humans were reaching to. With a similar goal to predict the intended target from human reaching motion, Pérez-D'Arpino and Shah (2015) proposed an online algorithm utilizing time series classification. Prior knowledge of the collaborative task was used to improve predictive accuracy. Based on the prediction of human reaching target, the robot would reach toward a different target.

Moreover, Koppula and Saxena (2013) demonstrated how anticipation using object affordance and motion trajectories could improve human activity recognition, which further allowed a robot to perform proper actions as responses to the anticipated human activity. In another investigation, anticipatory actions allowed a simulated robot to adapt to its user's workflow in an assembly scenario (Hoffman and Breazeal, 2007). Such anticipation yielded a greater concurrent activity, arguably a more fluid collaboration, within the human-robot team compared to that of a team with a non-anticipatory robot.

Prior research has also studied using predictions of human intention to direct a robot's focus (e.g., gaze fixation). For example, Ognibene and Demiris (2013) and Ognibene et al. (2013) utilized people's motions to predict their intentions and used these predictions to control the attention of a robotic observer. This behavior not only brought the robot's attention to a task-relevant place but also indicated to human partners that the robot shared the same interest with them.

Beside using behavioral observation, anticipation of human intentions and actions could be realized via task representations. Hawkins et al. (2014) proposed an AND-OR tree structure to represent finite human task actions. This task representation along with an inference method allowed probabilistically inference of what the human partner would do and when. Their implemented system using the proposed anticipatory approach could handle task and sensor uncertainty. Evaluations in simulation and physical HRI (N=6) showed that their system reduced human idle time, and therefore increased task efficiency, in a human-robot collaboration task.

Overall, benefits of anticipation in human-robot collaboration have been found to improve task efficiency, reduce human idle time (Hawkins et al., 2014; Hoffman and Breazeal, 2007), create safer collaboration (Mainprice and Berenson, 2013), and enhance activity recognition for both past occurrences and future events (Koppula and Saxena, 2013).

4.3 Study 4: predicting user intentions to support anticipatory actions

This section presents my investigation into enabling anticipatory actions for robots based on predicted user intentions. My investigation involved *modeling* the predictive relationship between patterns of gaze cues and user intentions (Section 4.3.1), *implementing* an autonomous robotic system that leveraged predictions of user intention in preparing anticipatory actions (Section 4.3.2), and *evaluating* the performance of and user experience with the robotic system in a human-robot interaction study (Section 4.3.3).

4.3.1 User modeling

Data collection

Task & setup — This data collection involved pairs of human participants engaged in a collaborative task. I used this study to collect data for the predictive model. During the data collection study, participants performed a sandwich-making task in which they sat across from each other at a table that contained 23 possible sandwich ingredients and two slices of bread. The initial layout of the ingredients was the same for each pair of participants (Figure 4.1). One participant was assigned the role of "customer," and the other was assigned the role of "worker." The customer used verbal instructions to communicate to the worker what ingredients he/she wanted on the sandwich. Upon hearing the request from the customer, the worker immediately picked up that ingredient and placed it on top of the bread.

Throughout the data collection study, both participants wore mobile eye-tracking glasses developed by SMI.[3] These eye-trackers perform binocular dark-pupil tracking with a sampling rate of 30 Hz and gaze position accuracy of 0.5 degrees. Each set of glasses contains a forward-facing high-definition (HD) camera that was used to record both audio and video at 24 fps. The gaze trackers were time-synchronized with each other so that the gaze data from both participants could be correlated.

Participants — Thirteen dyads of participants were recruited for the data collection study. All dyads were recruited from the University of Wisconsin–Madison campus and were previously unacquainted. Prior to the experiment, participants completed a written consent of participation. Each dyad carried out the sandwich-making task twice so that each participant acted as both customer and worker. The customer was instructed to request 15 ingredients for their sandwich. Participants kept their own count of the number of ingredients ordered, stopping when they had reached 15. The customer was further instructed to only request a single ingredient at a time and to refrain from directly pointing to or touching the ingredients. Upon completing the first sandwich, an experimenter entered the study room and reset

[3] http://www.smivision.com/en/gaze-and-eye-tracking-systems/home.html

Figure 4.1: Data collection of dyadic interactions in a sandwich-making task. Top Left: Two participants, wearing gaze trackers, working together to make a sandwich. Top Right: The participant's view of the task space from the gaze tracker. Orange circle indicates their current gaze target. Botoom: The layout of ingredients on the table. The ingredients, from top to bottom, left to right, are *lettuce1*, *pickle1*, *tomato2*, *turkey*, *roast beef*, *bacon2*, *mustard*, *cheddar cheese*, *onions*, *pickle2*, *ham*, *mayo*, *egg*, *salami*, *swiss cheese*, *bologna*, *bacon1*, *peanut butter*, *lettuce2*, *pickle3*, *tomato1*, *ketchup*, *jelly*.

the ingredients back to their original locations on the table, and the participants switched roles for the second sandwich.

Data processing

Following data collection, the proprietary BeGaze software created by SMI was used to automatically segment the gaze data into fixations—periods of time when the eyes were at rest on a single target—and saccades—periods of time when the eyes were engaged in rapid movement. Fixations were labeled with the name of the target fixated upon. Possible targets included the sandwich ingredients (Figure

4.1), the slices of bread, the conversational partner, and elsewhere in space. Speech was also transcribed for each participant. Customer requests for specific objects were tagged with the ID of the referenced object.

Charactering task intent

In this work, I considered the customers' intentions to be their chosen ingredients. Informed by the literature, I hypothesized that the customers' gaze patterns would signify their intent of which ingredients they wanted on their sandwich and aimed to develop a model to accurately predict intentions based on their gaze patterns. The data collection resulted in a total of 334 episodes of ingredient requests. I excluded episodes where more than 40% of the gaze data was missing before verbal requests, yielding 276 episodes for data analysis and modeling.

A naive, but plausible, strategy to predict a person's intent is solely based on his or her current gaze, which may indicate the person's current attention and interest (Frischen et al., 2007). To evaluate the efficacy of this strategy, I built an *attention-based* intention predictor that performed predictions according to which ingredient the customer most recently fixated on. An evaluation of the 276 episodes showed that the attention-based predictor achieved 65.22% accuracy in predicting the customers' choice of ingredient. This strategy outperformed random guesses of the ingredient, which were between 4.35% (i.e., 1/23) and 11.11% (i.e., 1/9), depending on how many potential ingredients were still available at that point in the interaction.

While the attention-based method was reasonably effective in predicting the intended ingredients, it only relied on the most recently glanced-at ingredient and omitted any prior gaze cues. However, the history of gaze cues may provide richer information for understanding and anticipating intent. In particular, I made two observations from the 276 episode analysis. First, participants seemed to glance at the intended ingredient longer than other ingredients. Second, participants glanced multiple times toward the intended ingredient before making the corresponding verbal request. These observations, along with significance of attention, informed my selection of characteristic features, as listed below, to represent patterns of

participant's gaze cues. Each of the four features was computed for all potential ingredients in every episode of an ingredient request.

Feature 1: Number of glances toward the ingredient before the verbal request (Integer)

Feature 2: Duration (in milliseconds) of the first glance toward the ingredient before the verbal request (Real value)

Feature 3: Total duration (in milliseconds) of all the glances toward the ingredient before the verbal request (Real value)

Feature 4: Whether or not the ingredient was most recently glanced at (Boolean value)

I applied a support vector machine (SVM) (Cortes and Vapnik, 1995)—a type of supervised machine learning approach that is widely used for classification problems—to classify the participants' gaze patterns into two categories, one for the intended ingredient (i.e., positive) and the other for non-intended, competing ingredients (i.e., negative). In this work, I used Radial Basis Function (RBF) Kernels and the implementation of LIBSVM (Chang and Lin, 2011) for the analysis and evaluation reported below.

To evaluate the effectiveness of the constructed model in classifying gaze patterns for user intentions, I conducted a 10-fold cross-validation using the 276 episodes of interaction. For each episode, I calculated a feature vector, including Features 1 to 4, for each ingredient that the customer looked toward before making a verbal request. To train the SVM, if an ingredient was the requested ingredient, the classification label was set to 1; otherwise, it was set to -1. In the test phase, the trained SVM determined the classification for each ingredient glanced at. On average, the SVMs achieved 89.00% accuracy in classifying labels of customer intention. Feature selection analyses (Chen and Lin, 2006) revealed that Feature 3 was the most indicative in classifying intentions, followed by Feature 4, Feature 1, and then Feature 2.

Evaluation of intention prediction

The SVM classifier was further modified to predict the customers' intentions. The input to the SVM predictor was a stream of gaze fixations. As the interaction unfolded, I maintained a list of candidate ingredients, their corresponding feature vectors, and the estimated probabilities of the ingredient being the intended request, calculated using the method based on Wu et al. (2004). When a new gaze fixation on an ingredient occurred, I first checked whether or not the ingredient was in the candidate list. If the ingredient was already in the list, I updated its feature vector and estimated probability; otherwise, I added a new entry for the ingredient to the list.

A traditional SVM was used to classify an ingredient to be the potential request if the estimated probability was greater than 0.5. If more than one ingredient was classified as a potential request, the traditional SVM predictor picked the ingredient with the highest probability as the final prediction. If, however, none of the ingredients were classified as potential requests, the predictor made no prediction. The effectiveness of such a traditional SVM predictor was assessed via a 10-fold cross-validation using our 276 episodes. For this evaluation, a prediction was considered to be correct only when the prediction matched the actual request. Note that this intention prediction was different from the classification of gaze patterns reported in the previous section. The accuracy of intention prediction was assessed by whether or not the predicted ingredients matched the requested ones, whereas the accuracy of intention classification was based on comparisons of classified labels, including both positive and negative, with actual labels. The traditional SVM predictor on average reached 61.52% accuracy in predicting which ingredients the customer would pick. Further analysis revealed that 28.99% of the time the SVM predictor made no predictions. However, when it made predictions (i.e., 71.01% of the time), the SVM provided predictions at 86.43% accuracy. This accuracy could be interpreted as the confidence of the traditional SVM predictor in predicting intention when it had a positive classification.

I defined an anticipation window as the time period starting with the last change in the prediction and ending with the onset of the speech utterance (see Figure 4.2

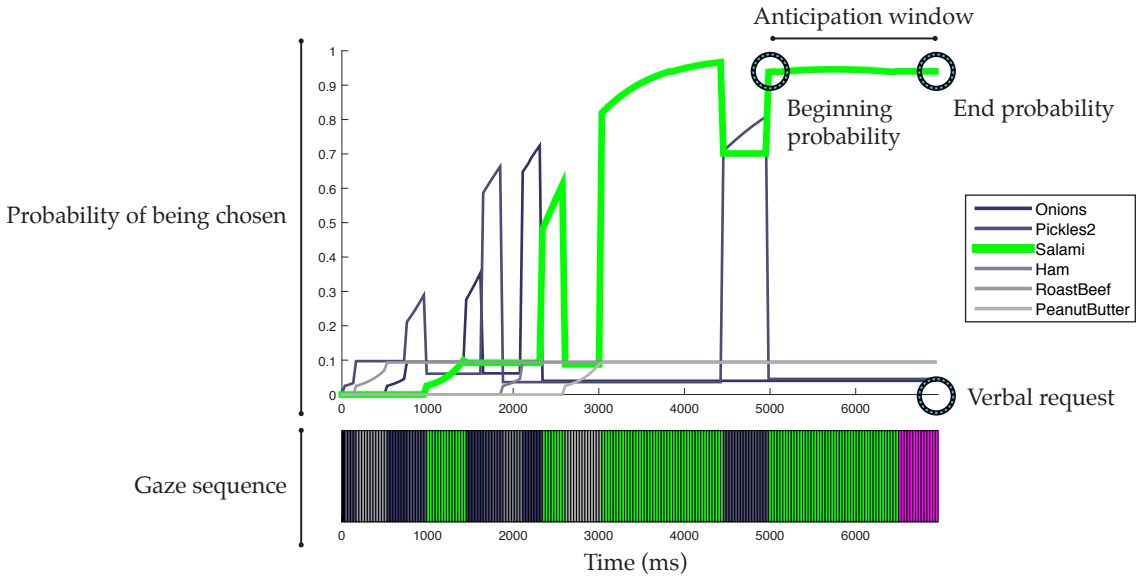


Figure 4.2: Illustration of episodic prediction analysis. Each illustrated episode ends at the start of the verbal request. The top plot shows probabilities of glanced ingredients that may be chosen by a customer. Note that the plotted probability was with respect to each ingredient. By calculating the normalized probability across all ingredients, we can determine the likelihood of which ingredient will be chosen. The bottom plot shows the customer's gaze sequence. Ingredients are color coded. Purple indicates gazing toward the bread. Black indicates missing gaze data. An anticipation window is defined as the time period starting with the last change in the prediction and ending with the onset of the speech utterance. The beginning and end probabilities are the probabilities of the predicted ingredient at the beginning and end of the anticipation window.

as an example). This anticipation window allowed us to understand how early the predictor could reach the correct predictions. For the traditional SVM predictor, the anticipation window for the correct predictions was on average 1420.57 milliseconds before the actual verbal request, meaning that the predictor could anticipate the intended ingredient about 1.4 seconds in advance. The interaction duration before the verbal request for the episodes with correct predictions was on average 3802.56 milliseconds (SD $=$ 1596.45).

The predictive accuracy of the traditional SVM predictor was largely impaired by the frequency with which it made no predictions. To address this issue, I ensured that the SVM-based predictor always made a prediction, choosing the ingredient

Table 4.1: Summary of the quantitative evaluation of the effectiveness of different intention prediction approaches.

Predition method	Predictive accuracy	Anticipation time
Chance	4.35% - 11.11%	N/A
Attention-based	65.22%	N/A
SVM-based	76.36%	1831 ms

with the highest probability. A 10-fold cross-validation using the 276 episodes showed that our SVM-based predictor on average reached 76.36% predictive accuracy and could make those correct predictions 1831.27 milliseconds ahead of their corresponding verbal requests (Interaction duration $M = 3802.56$, $SD = 1596.45$). Table 4.1 summarizes these results. Moreover, I analyzed the probabilities of the chosen ingredients that were at the beginning and end of the anticipation window (see Figure 4.2). On average, the beginning and end probabilities for the correct predictions were 0.36 and 0.75, respectively, whereas the beginning and end probabilities for the incorrect predictions were 0.28 and 0.43, respectively. These probability parameters indicate the confidence of the SVM-based predictor in making a correct prediction. For example, when the probability of an ingredient is over 0.43, the ingredient is likely to be the intended choice. I note that this threshold (0.43) is lower than the threshold used by the traditional SVM (0.50). Similarly, if the probability of an ingredient is lower than 0.36, the ingredient is less likely to be the intended choice. These parameters allow the construction of a real-time intention predictor that anticipates the customers' choices on the fly (Section 4.3.2).

In the next section, I provide examples and further analyses of when the SVM-based predictor made correct and incorrect predictions. These analyses revealed gaze patterns that may provide additional insight into understanding the customers' intentions.

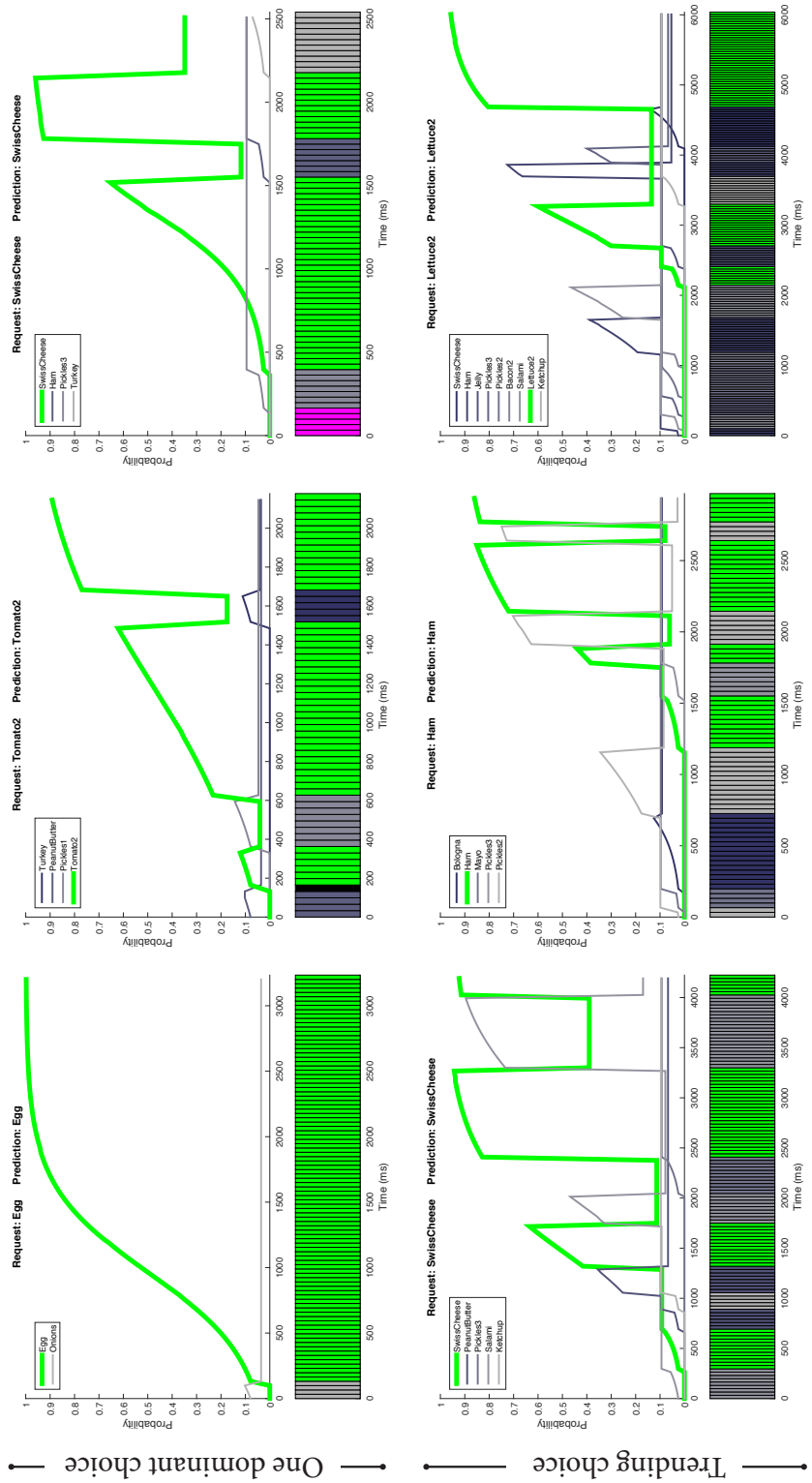


Figure 4.3: Two main categories of correct predictions: one dominant choice (top) and the trending choice (bottom). Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Purple indicates gazing toward the bread and yellow indicates gazing toward the worker. Black indicates missing gaze data.

Qualitative analysis

To further understand how the intention predictor made correct and incorrect predictions in the collected interaction episodes, I plotted the probability of each glanced-at ingredient over time, aligned with the corresponding gaze sequence received from the gaze tracker, for each interaction episode (see Figure 4.2 for an example). These plots facilitated a qualitative analyses of gaze patterns and further revealed patterns that were not captured in our designed features but may signify user intentions. In the following paragraphs, I present these analyses and discuss exemplary cases.

Correct predictions. Two categories—one dominant choice and the trending choice—emerged from the episodes with correct predictions (see examples in Figure 4.3).

One dominant choice — In this category, customers seemed to be focused toward one dominant ingredient, which was apparent in their gaze cues (Figure 4.3, Top). In particular, I found two types of gaze patterns. In the first, participants looked toward the intended ingredient for a prolonged time. In the second, they looked toward the intended ingredient multiple times in the course of their interaction. For both patterns, the intended ingredient received the majority of the gaze attention relative to other ingredients. This dominance allowed the predictor to give correct predictions.

Trending choice — In contrast to the previous category, there were situations in which customers did not seem to have a single ingredient in mind. In these situations, the customers exhibited a "shopping" behavior by looking toward multiple ingredients to decide which one to order. These situations usually involved the participants' visual attention being spread across multiple candidate ingredients. However, the customers generally looked toward the intended ingredient recurrently compared to other competing ingredients throughout the interaction. This recurrent pattern resulted in the intended ingredient becoming a trending choice, as illustrated in the bottom examples of Figure 4.3. The SVM-based predictor was observed to capture this pattern effectively.

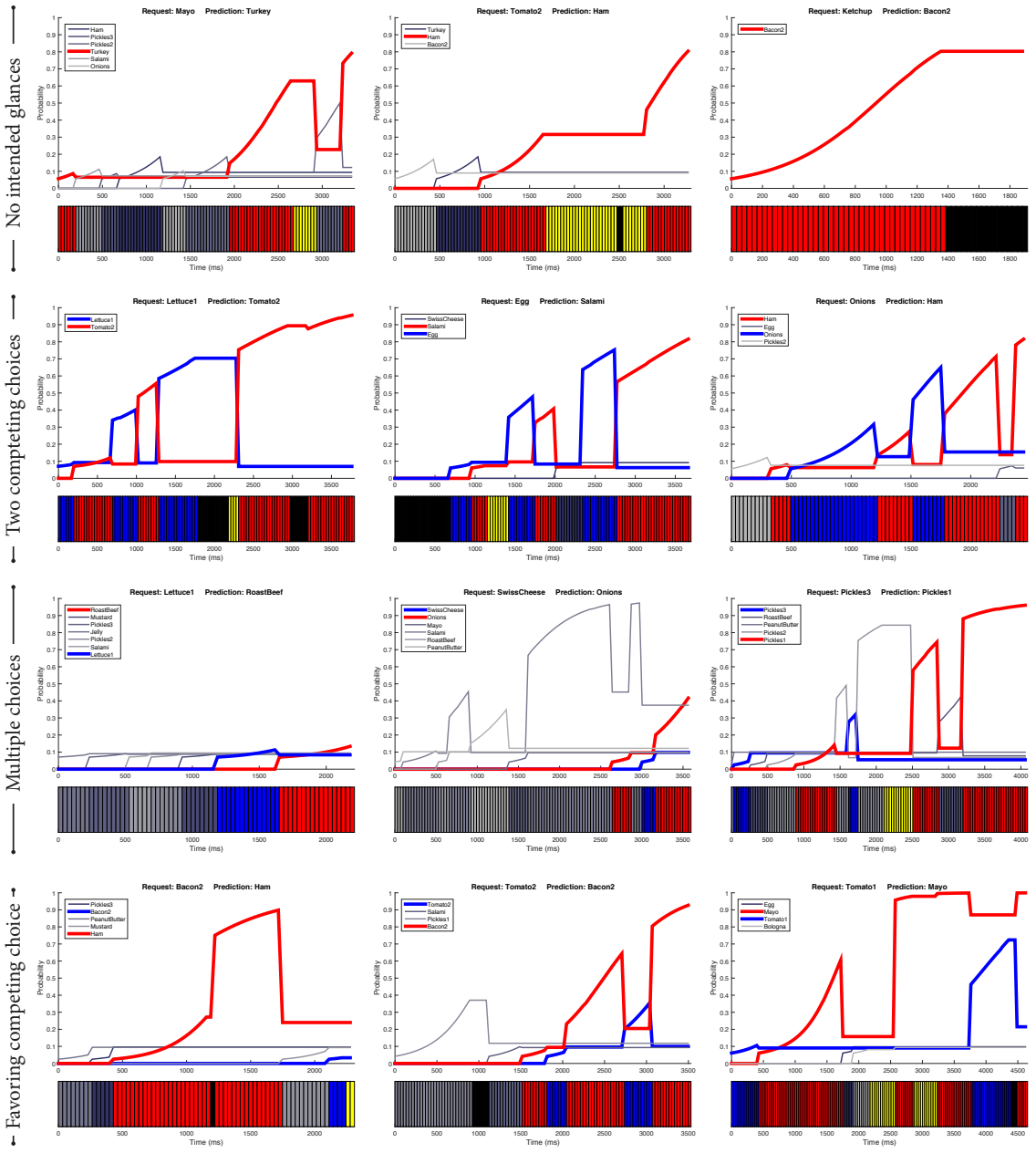


Figure 4.4: Examples of incorrect predictions. Red indicates the prediction made by the SVM-based predictor, whereas blue indicates the actual ingredient requested by the customers. Purple indicates gazing toward the bread whereas yellow indicates gazing toward the worker. Black indicates missing gaze data.

Incorrect predictions. From the 10-fold evaluation of the SVM-based predictor, there were a total of 62 episodes resulting in incorrect predictions. In the following paragraphs, I describe the characteristics of four identified categories of these incorrect predictions.

No intended glances — Among the incorrect predictions, there were 23 episodes (37.10%) during which the customers did not glance at the intended ingredients (Figure 4.4, First row). There are three reasons that might explain these cases. First, the customers had made their decisions in previous episodes. For example, when they were glancing around to pick an ingredient, they may have also decided which ingredient to order next. Second, their intentions were not explicitly manifested through their gaze cues. Third, the gaze tracker did not capture the gaze of the intended ingredient (i.e., missing data). In each of these cases, the predictor could not make correct predictions as it did not have the necessary information about the intended ingredients.

Two competing choices — Sometimes, customers seemed to have two ingredients they were deciding between (Figure 4.4, Second row). In this case, their gaze cues were similarly distributed between the competing ingredients. Therefore, gaze cues alone were not adequate to anticipate the customers' intent. I speculate that the determinant factors in these situations were subtle and not well-captured via gaze cues. Therefore, the predictor was likely to make incorrect predictions in these situations.

Multiple choices — Similar to the case of two competing choices, the customers sometimes decided among multiple candidate ingredients (Figure 4.4, Third row). As gaze cues were distributed across candidate ingredients, the predictor had difficulty in choosing the intended ingredient. Additional information, either from different behavioral modalities or new features of gaze cues, is necessary to distinguish the intended ingredient from the competing ones.

Favoring competing choices — In situations where the customers looked toward competing ingredients more frequently as compared to the intended ingredient, our predictor made incorrect predictions (see examples in Figure 4.4, Fourth row). One potential explanation for this type of gaze pattern is that the customers changed

their decision after quick glances at the intended ingredients. For instance, as shown in the bottom examples of Figure 4.4, while the customers looked longer and multiple times at the red ingredient, they requested the blue ingredient with smaller gaze attention. Our features failed to capture such quick decisions, likely resulting in incorrect predictions.

Special patterns. In analyzing the efficacy of the SVM-based intention predictor, I observed some special, potentially informative gaze patterns that were not explicitly captured in our derived features emerge. I discuss these patterns in the following paragraphs.

Initiating joint attention — Initiating joint attention is the process of using behavioral cues to direct the other's attention to a shared artifact. One such behavioral instantiation involves alternating gaze cues—looking toward the intended ingredient, looking toward the worker, and then looking back at the intended ingredient (Mundy and Newell, 2007). I found such patterns of initiating joint attention in the data, as shown in the first row of Figure 4.5. This pattern usually emerged toward the end of the episode, serving as a signal to the worker that the intended ingredient had been chosen.

Confirmatory request — The inverse pattern of initiating joint attention is that of the customer looking toward the worker, toward the intended ingredient, and then back toward the worker. Conceptually, we can characterize this pattern as a confirmatory request, meaning that the customer sought the worker's attention, directed their attention, and checked if the intention was understood. From our data, this pattern of confirmatory request seemed to signify intention. As illustrated in the second row of Figure 4.5, the single ingredient between fixations at the worker was the intended ingredient.

Goal referencing — Another pattern that emerged from the data was visual references to the goal, which in our context was the bread where ingredients were moved. This type of reference was found in a variety of combinations. It could be found before, after, or in between choosing the intended ingredient. Examples are provided in the third row of Figure 4.5. There may be different meanings to these combinations. For instance, the customers might have checked which

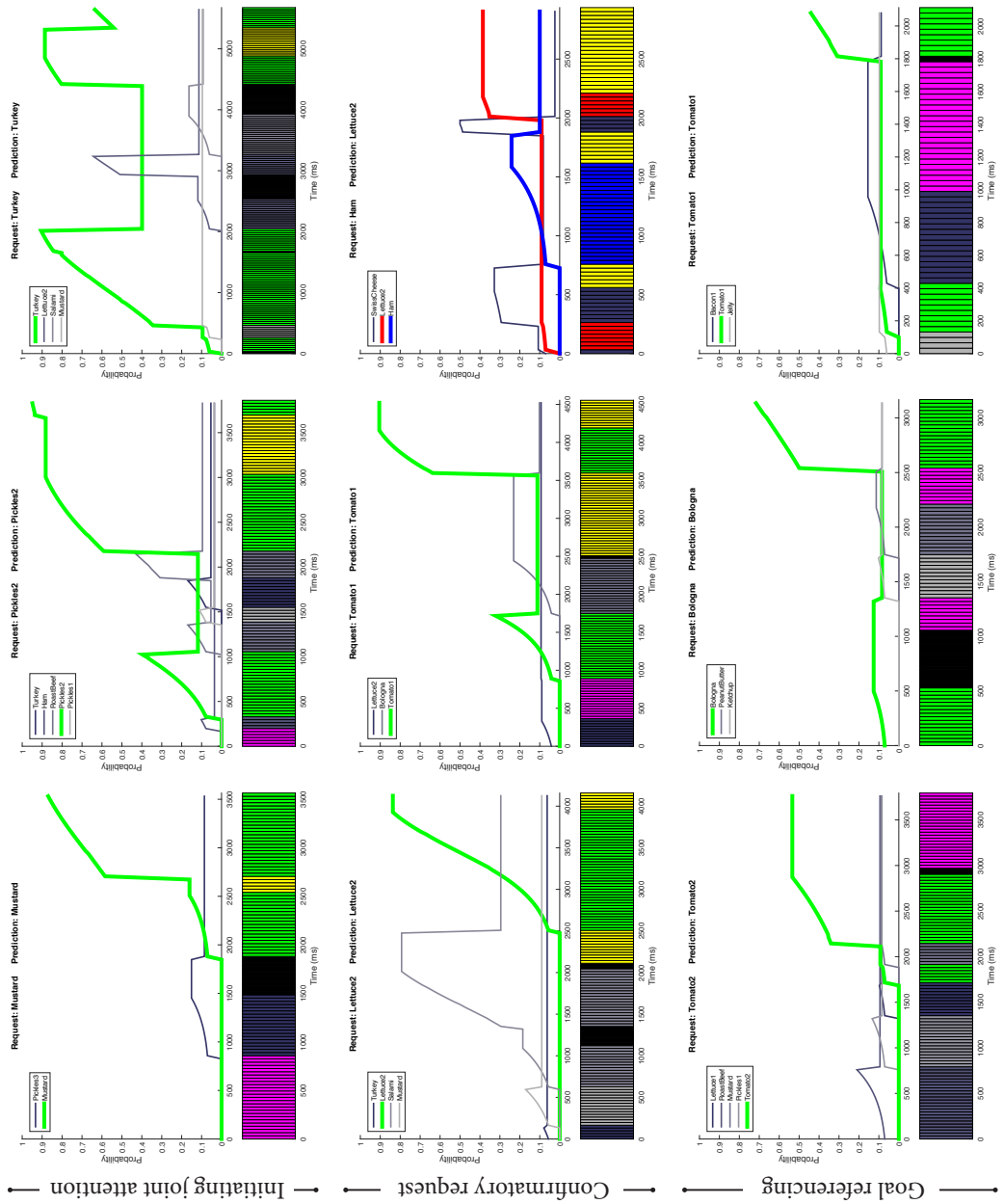


Figure 4.5: Examples of special gaze patterns. Green indicates the ingredients predicted by our SVM-based predictor that were the same as the actual ingredients requested by the customers. Blue lines indicate the ingredients that the customers picked. Red lines are our predictions. Purple indicates gazing toward the bread, whereas yellow indicates gazing toward the worker. Black indicates missing gaze data.

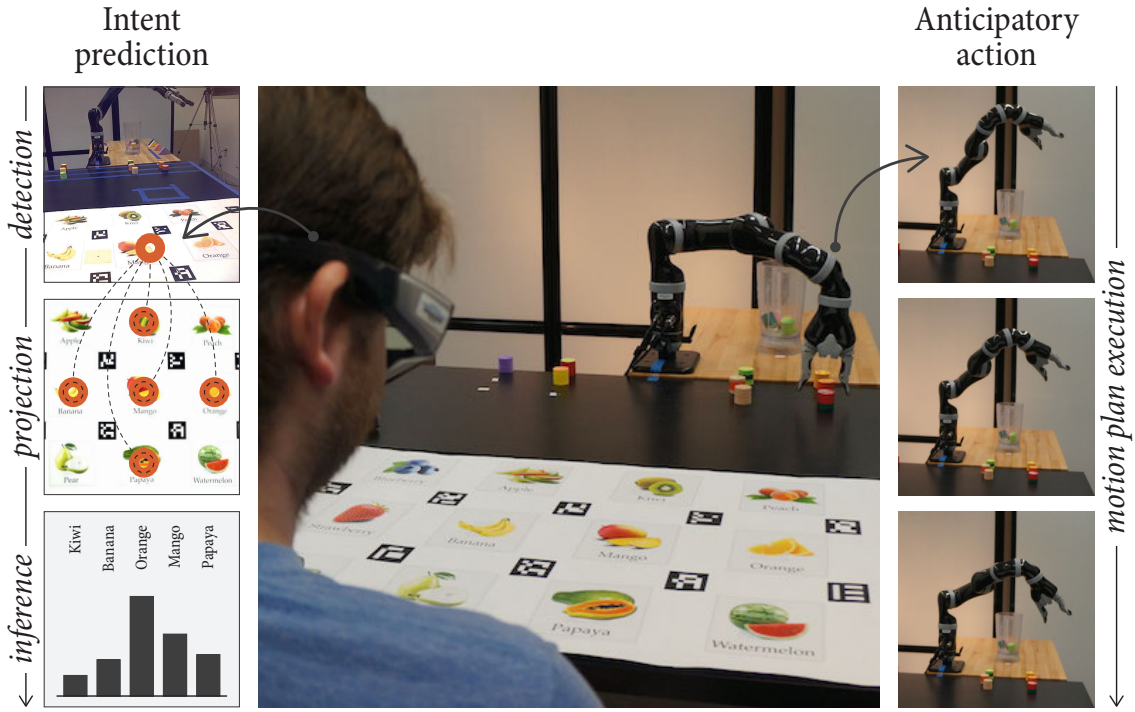


Figure 4.6: I propose an "anticipatory control" method that enable robots to proactively plan and execute actions based on an anticipation of a human partner's task intent as inferred from their gaze patterns.

ingredients had been added to the sandwich and used that information to decide which ingredient to pick next.

4.3.2 System implementation

The above modeling analyses revealed much about how gaze patterns might indicate user intentions. In this section, I propose an *anticipatory control* method that involves monitoring user actions, predicting user task intent, and proactively controlling robot actions according to predicted user intent as an alternative to *reactive control* methods that utilize direct, explicit user input. In this section, I present the implementation of this method as a real-time autonomous robot system following a sense-plan-act paradigm (Figure 4.6). To provide a context for the development

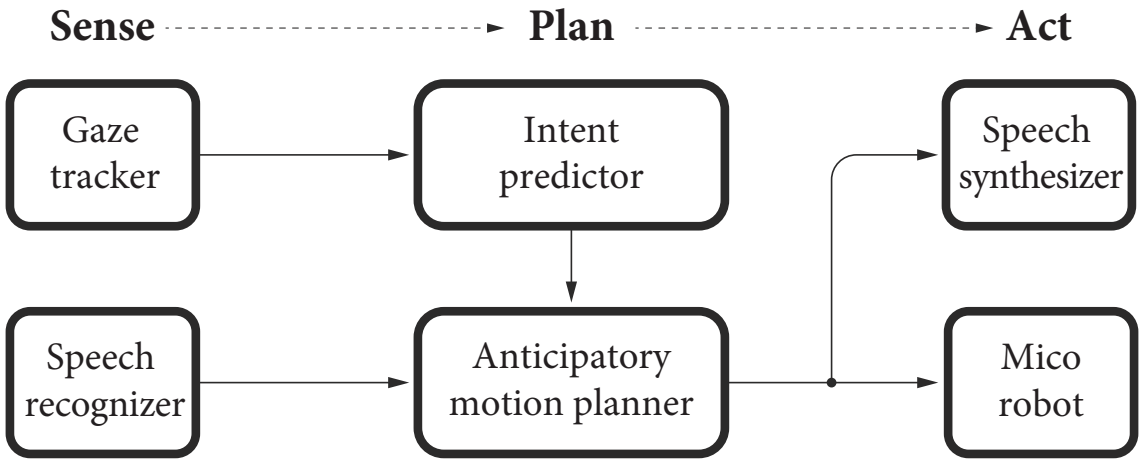


Figure 4.7: System diagram of the implemented system for anticipatory action preparation and execution.

and implementation of our proposed method, we devised a task in which a robot works as a "server" preparing smoothies for a human "customer" to represent interactions common in day-to-day collaborations.

The proposed method integrated six components: (1) gaze tracking, (2) speech recognition, (3) intent prediction, (4) anticipatory motion planning, (5) speech synthesis, and (6) robotic manipulation. Figure 4.7 illustrates how these components are integrated by the implemented system, and the sections below provide detail on their functioning and implementation.

Gaze tracking

The gaze-tracking component captured gaze fixations from a pair of SMI Eye-Tracking Glasses V.1[4] worn by the user. It then performed a *projective transformation* using the Jacobi method to map gaze fixations in the camera-view space to locations in the physical task space. These points were subsequently used to infer what task-relevant items were being looked toward. The mapping between the camera-view space and physical space and the association between locations in the physical

[4]http://www.eyetracking-glasses.com

space and environmental items were realized by locating a set of predefined Aruco markers[5].

Speech recognition & synthesis

Microsoft Speech API 5.4 was used to build a speech-recognition component to recognize user utterances and a speech-synthesis component to realize the robot's speech. In order to minimize speech recognition errors, the recognition grammar was designed to be flexible to accommodate different forms of verbal requests such as, "I would like to have mango," "Could I have papaya," or simply "Peach." The robot's speech included greetings, confirmations of user requests such as, "You ordered mango," task instructions such as, "Next one," and a "Thank you" remark uttered at the end of the interaction.

Robotic manipulation

A Kinova MICO robot arm was used as the manipulator to pick up the requested items and place them at a target location, which in the context of our task involved placing smoothie ingredients into a blender. The arm was controlled using the `MoveIt!` platform[6] and was given a clear representation of the environment for motion planning.

Intent prediction

In building the intent-prediction component, I extended the intent-prediction framework described in Section 4.3.1 to train a support vector machine (SVM) with data on partner gaze patterns and task intent in a collaborative sandwich-making scenario. In the scenario, one participant worked as a "server" to add ingredients requested by a "customer" participant to his or her sandwich.

In realizing the anticipatory control method presented here, I used the collected data to train an SVM classifier that predicted user task intent based on the history of what items the user has looked toward. In the context of our task, the input to the classifier was the ingredients (shown on a menu placed in front of them) users

[5]http://www.uco.es/investiga/grupos/ava/node/26
[6]http://moveit.ros.org

looked toward, and the outputs were a prediction of what item the user would request next and the confidence level of the model in its prediction.

Anticipatory motion planning

Using the `MoveIt!` platform, the anticipatory motion planner utilized the prediction and confidence that the intent-prediction component provided to proactively plan and execute motion toward the predicted item (Algorithm 2). If the confidence of the prediction was higher than `planThreshold`, set to 0.36, the motion planner planned a motion toward the predicted item. If the confidence was higher than `execThreshold`, set to 0.43, it executed part of the planned motion based on its current confidence (see the description of the `splitPlan` method below). Taken from the user modeling study (Section 4.3.1), these thresholds indicate that if the confidence of a prediction is higher than 0.36, the prediction could be correct, and that if it exceeds 0.43, the prediction was unlikely to be incorrect.

Algorithm 2 Anticipatory Robot Control

Require: currPred, currProb
1: **while** true **do**
2: predHistory ← updatePredHistory(currPred, currProb)
3: weightedPred, weightedProb ← getWeightedPred(predHistory)
4: **if** weightedProb ⩾ planThreshold **then**
5: motionPlan ← retrievePlan(weightedPred)
6: **if** (motionPlan = ∅) or (weightedPred ≠ currMotionTarget) **then**
7: makePlan(weightedPred)
8: **end if**
9: **end if**
10: **if** weightedProb ⩾ execThreshold **then**
11: motionPlan ← retrievePlan( )
12: subPlan1, subPlan2 ← splitPlan(motionPlan)
13: requestExec(subPlan1)
14: updatePlanLibrary(weightedPred, subPlan2)
15: **end if**
16: **end while**

Instead of using the current prediction and confidence, denoted as currPred and currProb, respectively, directly from the intent-prediction component, the anticipatory motion planner maintained a history of the 15 latest predictions, including the current prediction. The gaze-tracking component provided readings at approximately 30 Hz, and thus the length of the prediction history was chosen to be approximately 500 milliseconds. The prediction history was then used to calculate a weighted prediction that discounted past predictions using the exponential decay function defined in Equation 4.1.

$$p'_i = p_i \times (1 - \mathrm{decayRate})^i \tag{4.1}$$

In this function, p_i denotes the probability of the ith prediction in the history. The decayRate, set to 0.25, indicates the rate at which the weight of the prediction decayed, and the resulting prediction (*weightedPred*) is the prediction with the highest weight summed over the prediction history.

The anticipatory motion planner maintained a plan library that stored a set of candidate motion plans from which it chose when the robot had made a prediction of the user's request. The currMotionTarget variable denotes the motion target associated with the most recent plan. The makePlan function utilized the RRT-Connect algorithm (Kuffner and LaValle, 2000) (the `RRTConnectkConfigDefault` planner in `MoveIt!`) to create a motion plan toward the *weightedPred* item. The splitPlan function took a motion plan and split it into two sequential sub-plans proportionally based on the confidence of the prediction, denoted as *weightedProb*. Higher confidence values moved the robot closer and closer to the predicted item. Although this iterative planning could bring the robot to a position in which it could grasp the ingredient, we chose to delay the grasp until the user made a verbal request in order to more easily and fluidly recover from errors.

The implementation of the anticipatory-motion-planning component involved three threads: a *planning* thread that implemented Algorithm 2, an *execution* thread that executed motion plans, and a *speech* thread that processed user requests. The planning thread put a motion request into a plan queue using the requestExec

function. The execution thread regularly checked the queue of plans and executed them. When processing a verbal request, the speech thread checked if the robot's current motion target—if it had one—matched the user's request. If it did, the robot carried out the rest of the motion plan in order to complete the request. Otherwise, it stopped the current motion and made a new plan, directing motion toward the item requested by the user. We note that anticipatory control was used for determining and reaching toward requested items and not for transporting grasped items to the target location.

System limitations

The anticipatory robot system had three main sources of error: tracking, projection, and prediction. Tracking errors resulted directly from the eye-tracking system. Even with a state-of-the-art eye-tracking system that was calibrated for each user following the manufacturer-recommended calibration procedure, some amount of tracking error was unavoidable. A second source of error arose from the projection process of gaze fixations provided by the eye tracker to the workspace. Mismatched tracking rates between the eye tracker and the tracker used for Aruco markers led to incorrect inferences of which items were gaze targets. Finally, the intent-prediction component provided erroneous predictions partly due to errors that cascaded through the tracking and projection processes and partly due to the limitations of the trained model.

4.3.3 Experimental evaluation

In this section, I describe the design of and findings from a human-robot interaction experiment that evaluated the effectiveness of the proposed anticipatory control method in supporting team performance and user experience.

Hypothesis

Our central hypothesis is that *anticipatory* control, as implemented in the anticipatory robot system (Section 4.3.2), would enable the robot to more effectively respond

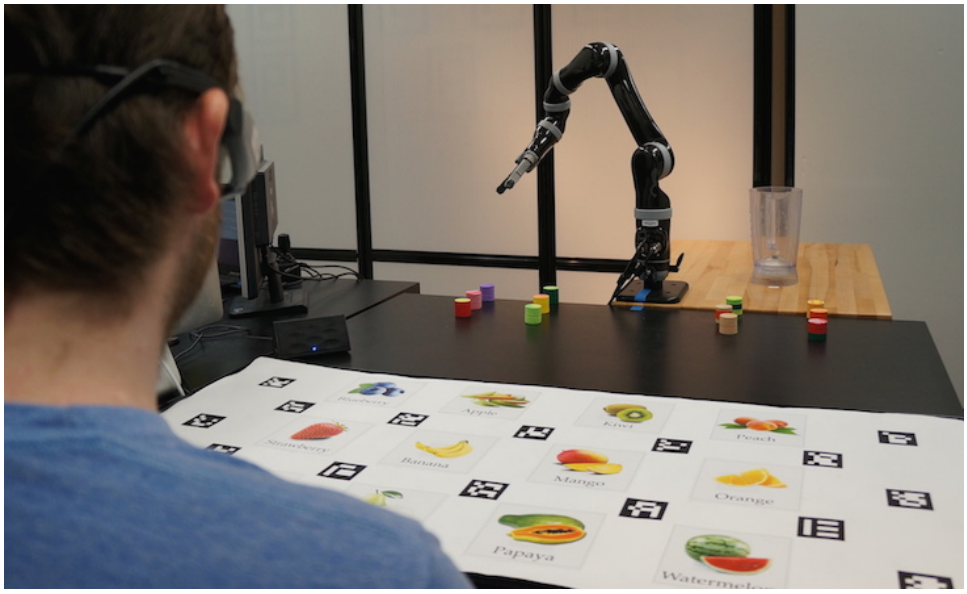


Figure 4.8: The task setup of the human-robot interaction experiment in Study 4. The robot served as a "worker" to prepare a smoothie ordered by a human "customer." Between the robot and the user were a menu for the user to look for his/her choice and a workspace for the robot to prepare the order.

to user requests, thus resulting in improvements in team performance and user perceptions of the robot, compared to other, more *reactive* forms of control.

Experimental task, design, & conditions

To test our hypothesis, we devised an experimental task in which human participants, acting as customers, ordered two fruit smoothies from a robot system that served as a café worker. During the task, participants sat across from the robot with a menu of 12 different fruit choices placed in front of them (Figure 4.8). Participants were asked to choose five fruits from the menu for each order and to request one fruit at a time using verbal requests.

Two experimental conditions—*anticipatory* and *reactive*—were implemented on the robot system for evaluation. In the *anticipatory* condition, the robot predicted the user's choices and proactively planned and executed its motions based on its

prediction, as described in Section 4.3.2. In the *reactive* condition, the robot simply responded to the user's verbal requests.

The experiment followed a within-participants design. The only independent variable was whether or not the robot anticipated user choices before acting on them. Each participant interacted with the robot in both conditions, and the order of conditions was counterbalanced. We designed the experimental task to involve practices people commonly follow in daily interactions that one would expect at a café in order to minimize learning effects and the need for extensive training.

Procedure

Upon receiving informed consent from the participant, the experimenter provided the participant with an explanation of the task and how they should interact with the robot. The participant was fitted with head-worn eye-tracking glasses developed by SMI. The experimenter then performed the calibration procedure for the eye-tracker and an additional procedure for gaze-projection verification. In this procedure, the experimenter asked the participant to look toward four different fruits on the menu one at a time. The participant was also asked to report verbally toward which fruit they were looking. This procedure assessed the accuracy of the gaze projection capability after the calibration of the eye tracker. The participant then followed the robot's instructions to complete a drink order and filled out a questionnaire to evaluate their experience with and perceptions of the robot. This procedure was then repeated for the other condition. After interacting with the robot in both conditions, the experimenter collected demographic information and interviewed the participants for additional comments on differences they may have observed in the robot's behaviors between the two conditions.

Measures

We expected the performance of the anticipatory robot system to be affected by the potential errors accumulated throughout the pipeline of tracking the participant's eyes, inferring gaze targets, and predicting participant intent. To gain a more

detailed understanding of the effects of these errors on team performance, we employed two system measures: *projection* accuracy and *prediction* accuracy.

Projection accuracy (%): The number of matches between gazed and reported items divided by the total number of items (i.e., four per participant), measured during the projection-verification procedure.

Prediction accuracy (%): The number of matches between system predictions and user requests divided by the total number of user requests (i.e., five per interaction episode), measured during the experimental task.

To assess the effectiveness of anticipatory and reactive control methods in supporting human-robot collaboration, we utilized a number of objective and subjective measures. Objective measures included *response time*, *time to grasp*, and *time to release*.

Response time (milliseconds): The duration between when the participant verbally placed a request and when the robot started moving toward the requested item. For the anticipatory system, this measure captured the time it took to initiate a planned motion if the robot's prediction matched the user's request. Otherwise, it additionally captured the time needed to stop the current motion toward an incorrect prediction and the time to plan and initiate motion toward the correct target. For the reactive system, the measure only captured the time needed to plan and initiate motion toward the requested item as soon as the request was recognized.

Time to grasp (seconds): The duration between when the participant verbally requested an item to when the robot grasped the requested item.

Time to release (seconds): The duration between when the participant verbally requested an item and when the robot released the requested item at the target location (i.e., the blender). This measure was also considered as *task time*, as it represented how much time was needed for the robot to complete user requests.

In addition to the objective measures described above, we used a questionnaire to assess participants' subjective perceptions of the robot's anticipatory behaviors, particularly its perceived *awareness* and *intentionality*. The awareness scale, consisting of four items (Cronbach's $\alpha = 0.74$), aimed to measure how aware participants

thought the robot was of their intended choices. The intentionality scale, consisting of four items (Cronbach's $\alpha = 0.83$), aimed to capture participant perceptions of how mindful, conscious, intentional, and intelligent the robot appeared.

Finally, a single item, "The robot only moved to pick up an item after I verbally issued a request," served as a manipulation check, examining whether or not users were able to discern the difference between the anticipatory and reactive systems.

Participants

Twenty-six participants were recruited from the local community. Two participants were excluded from the data analysis due to failures of eye tracking and the operation of the online motion planner. The resulting 24 participants (16 females, 8 males) were aged between 18 and 32 ($M = 22.21$, $SD = 4.15$). Four participants reported to have interacted with a similar robot arm prior to their participation in the current study. The study took 30 minutes, and participants were paid $5 USD.

Results

The paragraphs below report on results from our system, objective, and subjective measures. We describe findings from the system measures first in order to provide context for the objective and subjective measures, as they were affected by the potential errors accumulated through the eye-tracking, gaze-projection, and intent-prediction phases.

System measures — The overall *projection accuracy* for our anticipatory system was 81.25%. Incorrectly inferred items were usually immediate neighbors (i.e., above, below, to the left, and to the right) of the intended targets. This accuracy rose to 91.67% if neighbors were considered as correct.

Out of 120 predictions, the anticipatory system made 53 incorrect predictions, yielding 55.83% *prediction accuracy*. However, eight of these incorrect predictions were due to not being able to make any prediction, because the users did not look toward any items on the menu prior to making requests. Additionally, in another 18 trials, participants did not look toward the requested item but rather looked toward other items, resulting in incorrect predictions. Possible explanations for

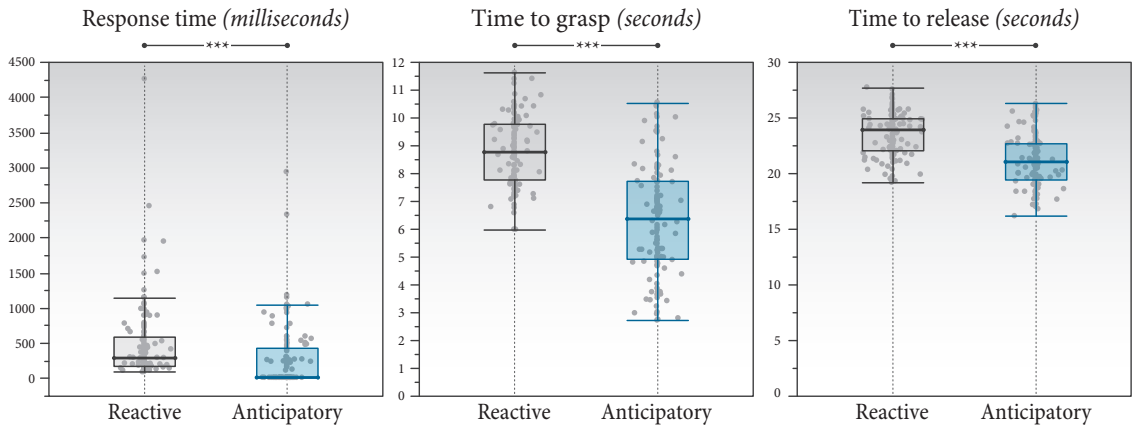


Figure 4.9: Tukey box plots of data from the objective measures. The extents of the box represent the the first and third quartiles. The line inside the box represents the second quartile (the median). The difference between the first and third quartiles is the interquartile range (IQR). The ends of the whiskers represent the first quartile minus 1.5 times IQR and the third quartile plus 1.5 times IQR. (***) denotes $p < .001$.

these behaviors are that participants decided on their next ingredient during the previous request, that the eye tracker failed to accurately capture gaze direction, or that gaze projection was erroneous. Our system reached 59.82% prediction accuracy in cases where a prediction was made and 77.5% accuracy if the user had glanced at the intended item. Baseline accuracy (chance) varied between 8.33% (1/12) and 12.5% (1/8).

For the analysis of the objective and subjective measures, we used one-way repeated-measures analysis of variance (ANOVA) following a linear mixed-models procedure in which *control method*, either anticipatory or reactive, was set as an independent variable, and *participant ID* and the interaction between the two were set as random variables. Tables 4.2 & 4.3 and Figure 4.9 & 4.10 provide results from this analysis.

Manipulation check — We found a significant difference in users' perceptions of when the robot moved toward the requested item (Table 4.2), indicating that participants were able to discern the differences due to our experimental manipulation.

Table 4.2: Statistical test results for objective measures, correlations between prediction accuracy and objective measures, and test results for subjective measures.

Objective measures

Control method	Response time *(ms)*	Time to grasp *(s)*	Time to release *(s)*
Reactive	482.71 (*SD*=551.33)	8.80 (*SD*=1.26)	23.44 (*SD*=1.91)
Anticipatory	256.41 (*SD*=443.31)	6.29 (*SD*=1.99)	21.10 (*SD*=2.31)
	$F(1,23)=10.44$ $p=.004$	$F(1,23)=120.20$ $p<.001$	$F(1,23)=66.71$ $p<.001$
Prediction accuracy	$r(118)=-0.52$ $p<.001$	$r(118)=-0.50$ $p<.001$	$r(118)=-0.41$ $p<.001$

Subjective measures

Control method	Manipulation check	Awareness	Intentionality
Reactive	6.79 (*SD*=0.41)	3.91 (*SD*=1.56)	4.54 (*SD*=1.73)
Anticipatory	4.13 (*SD*=2.31)	5.09 (*SD*=1.29)	4.66 (*SD*=1.58)
	$F(1,23)=33.45$ $p<.001$	$F(1,23)=31.57$ $p<.001$	$F(1,23)=0.50$ $p=.487$

Objective measures — Tables 4.2 & 4.3 and Figure 4.9 provide results from our objective measures, including *response time*, *time to grasp*, and *time to release*. We found that anticipatory control enabled the robot to more efficiently respond to and complete participants' requests than did reactive control. The average duration for finding and initializing a valid motion plan toward the target item (i.e., the response time of the reactive system) was 482.71 ms, which we argue to be reasonably responsive in the context of our task. Anticipatory control based on predicted participant intent reduced the response time by 226.3 ms. If the predictions were correct, the response time on average for the anticipatory system was 51.03 ms.

Table 4.3: Descriptive statistics of objective measures broken down into correct (A-Correct) and incorrect (A-Incorrect) predictions as well as correct (A-N-Correct) and incorrect (A-N-Incorrect) neighboring predictions.

Control method	Response time *(ms)*	Time to grasp *(s)*	Time to release *(s)*
Reactive	482.71(*SD*=551.33)	8.80 (*SD*=1.26)	23.44 (*SD*=1.91)
A–Correct	51.03 (*SD*=195.64)	5.40 (*SD*=1.72)	20.26 (*SD*=2.00)
A–Incorrect	516.04 (*SD*=527.35)	7.41 (*SD*=1.74)	22.17 (*SD*=2.45)
A–All	256.41 (*SD*=443.31)	6.29 (*SD*=1.99)	21.10 (*SD*=2.31)
A–N–Correct	164.70 (*SD*=300.38)	5.80 (*SD*=1.85)	20.64 (*SD*=2.20)
A–N–Incorrect	587.97 (*SD*=673.68)	8.06 (*SD*=1.40)	22.76 (*SD*=1.95)

Moreover, the anticipatory system proactively moved toward the predicted item of choice based on its confidence in the prediction. This proactive execution reduced time to grasp by 2.51 seconds. When predictions were correct (55.83% of the time), the anticipatory system would have partially completed its movement toward the requested item by the time it received the participant's verbal request, resulting in a 3.4-second advantage. When predictions were incorrect but involved items neighboring the requested item (78.33% of the time), anticipatory control still benefited time to grasp (3-second advantage), as the system would have moved toward the vicinity of the correct item, providing it with a time advantage in moving toward the correct item.

Additionally, anticipatory control provided a 2.34-second advantage in completing the participants' requests, as indicated by the *time to release* measure. This difference was similar to the difference in *time to grasp* between the two systems because system behavior for transporting the grasped item to the blender did not differ between the two systems.

We also found that the ability to correctly predict user intent was strongly associated with improvements in the three objective measures that resulted from the use of anticipatory control. Correlation analyses using Pearson's product-moment method showed that prediction accuracy was strongly correlated with response

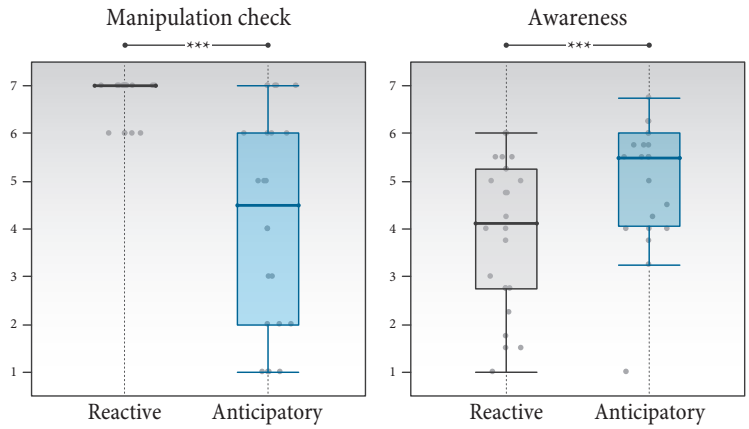


Figure 4.10: Tukey box plots of data from the manipulation check and subjective measures. The extents of the box represent the the first and third quartiles. The line inside the box represents the second quartile (the median). The difference between the first and third quartiles is the interquartile range (IQR). The ends of the whiskers represent the first quartile minus 1.5 times IQR and the third quartile plus 1.5 times IQR. (***) denotes $p < .001$.

time, time to grasp, and time to release (Table 4.2). This interdependence between prediction accuracy and objective measures suggests that correctly predicting user intent is important in achieving efficient human-robot collaboration.

Subjective measures — Table 4.2 and Figure 4.10 summarize the results from our subjective measures, particularly the perceived *awareness* and *intentionality* of the robot. Participants rated the anticipatory system to be significantly more aware of their intended choices than the reactive system. However, no significant differences were found in how intentional participants found the two robot systems to be.

Post-experiment interview — In the post-experiment interview, we asked participants open-ended questions about their perceptions of how the two systems behaved in preparing their orders. Several participants described the proactive behavior of the anticipatory robot as being efficient, which was in line with the findings from our response time, time to grasp, and time to release measures, as illustrated in the excerpts below:

P3: *"[The anticipatory robot] seemed like it's moving toward what I was going to order, so I thought it knew... I guess that would be more time efficient if it already knew."*

P4: *"[The anticipatory robot] just moved the arm closer to the fruit before I said something and so it was faster... it was preparatory... it was being more efficient."*

P5: *"[The anticipatory robot] was going for, I guess, what my eyes were looking towards before I even made a decision."*

Eight participants explicitly mentioned that they preferred the anticipatory system over the reactive one because of the perceived efficiency and proactivity of the robot. However, two participants expressed preference toward the reactive system, one participant describing the robot's anticipatory actions as "freaky" and reporting feeling "unnerved" and "bothered:"

P1: *"I could tell [the anticipatory robot] was watching my gaze or aware of my gaze... It has awareness... and that almost felt kind of freaky... that it almost could guess what I wanted... I didn't like it as much."*

The other participant who preferred the reactive system over the anticipatory one cited an instance of the anticipatory robot making a wrong prediction and moving toward the opposite direction as the primary basis of this preference:

P8: *"[The anticipatory robot] shouldn't move before I said what I wanted... so I guess that's [its] fault..."*

4.3.4 Discussion

In this study, I modeled the predictive relationship between exhibited gaze cues and intentions, implemented an autonomous system that utilized the predictive relationship in preparing anticipatory robot actions, evaluated the implemented system

in a laboratory human-robot interaction study. Below, I discuss the results from the user modeling study and the HRI evaluation experiment, potential applications of the anticipatory system, and limitations of the present study.

How gaze patterns reveal intention

The qualitative analyses (Section 4.3.1) provided not only insight into how the SVM-based predictor made correct and incorrect predictions, but they also revealed special patterns that may *signal* intentions via visual references to the other person and the goal. Signaling is an intentional strategy that people use to manifest actions and intentions in a way that is more predictable and comprehensible to interaction partners (Pezzulo et al., 2013). For example, parents exaggerate intonation in infant-directed speech (Kuhl et al., 1997). The use of signaling strategies facilitates the formation of common ground. The special patterns *initiating joint attention* and *confirmatory request* involved interleaving gaze cues between the partner and the intended ingredient. These displays of interleaving gaze may serve as an intentional signaling strategy, highlighting the relevance of the intended ingredient. Similarly, the visual references to the goal, which is the bread in the scenario, may be signaling the intentional link between the bread and the intended ingredient, as shown in the pattern *goal referencing*.

The four features of gaze cues explored in this work were based on statistical measures of the customers' gaze sequences. While these features seemed to capture how the distribution of gaze cues may indicate intentions, they did not explicitly encode sequential structures from gaze sequences. However, sequential structures—such as gaze toward the target, then partner, and then the target again—may encapsulate particular semantic meanings, such as directing the partner's attention toward the target. The capability to recognize these sequential structures as those of *initiating joint attention*, *confirmatory request*, and *goal referencing*, could reveal the underlying meanings of gaze sequence and potentially improve the efficacy of the SVM-based predictor. For example, the last plot of the examples of *confirmatory request* showed that the intention predictor could benefit from recognizing the sequential human-target-human pattern. One way to recognize such sequential

structures is through template matching, which has been explored to recognize communicative backchannels (Morency et al., 2010).

However, the special patterns, identified in Section 4.3.1, should be used with caution when predicting intentions. The last plot in Figure 4.4 illustrated a contradictory example; even though there was a clear pattern of *confirmatory request*, it did not signify the intended ingredient. Further research is necessary to investigate how the incorporation of sequential structures into the predictive model may enhance predictive performance.

Performance of intention prediction in HHI & HRI

How comparable were the results of intention prediction from the human-human and human-robot studies? Recall that in the user modeling study I showed that the SVM-based model reached 76.36% of accuracy in predicting users' choices of ingredient with manually annotated data. However, careful annotation of which item the user looked at was not available during the online interaction. Instead, online projection of gaze fixations to real-world items was applied, and it was showed that the projection accuracy was 81.25%. One way to compare the prediction accuracy in the offline and online analyses is to discount the offline performance with the online projection accuracy (i.e., $76.36\% * 81.25\%$), yielding the predictive performance to be 62.04%. In comparison, the online prediction accuracy was 59.82%.

Moreover, the imperfection in gaze projection mostly resulted in associating the gaze point to one of the neighbor items—the immediate up, down, left, and right items—as opposed to the intended item. The projection accuracy would be improved by 10% if considering neighbors were correct associations. The online prediction accuracy would be 78.33% if neighbors were considered to be the intended items. In comparison, the offline prediction accuracy was 76.36%. Together, these analyses asserted that the predictive performance of the anticipatory system was subject to the quality of projection of gaze fixations. Moreover, the performance of intention prediction in human-human and human-robot interactions was comparable if considering the limitation of gaze projection.

In the offline analysis of intention prediction, I further showed that the SVM-based predictor could anticipate the intended user choice approximately 1.8 seconds (i.e., *anticipation window*) before the corresponding user request. In comparison, the HRI study revealed that the anticipatory robot had about 2.5 seconds advantage in reaching toward to the correct item as well as completing the task. Presumably, this 2.5-seconds advantage was enabled by the anticipation capacity that the robot was equipped with.

Taking together these comparisons of predictive performance, I would argue that the implemented anticipatory capacity for the robot was similar to that in people in the user modeling study.

Applications

The capability to interpret others' intentions and anticipate actions is critical in performing joint actions (Huber et al., 2013; Sebanz and Knoblich, 2009). Prior research has explored how reading intention and performing anticipatory actions might benefit robots in providing assistance to their users, highlighting the importance of intention prediction in joint actions between humans and robots (Hoffman and Breazeal, 2007; Sakita et al., 2004). Building on prior research, this study provides empirical results showing the relationship between gaze cues and human intentions. It also presents an implementation of an intention predictor using SVMs. With the advancement of computing and sensing technologies, such as gaze tracking systems, I anticipate that an even more reliable intention predictor could be realized in the foreseeable future. Computer systems such as assistive robots and ubiquitous devices could utilize intention predictors to augment human capabilities in many applications. For example, robot co-workers could predict human workers' intentions by monitoring their gaze cues, enabling the robots to choose complementary tasks to increase productivity in manufacturing applications. Similarly, assistive robots could provide necessary assistance to people by interpreting their gaze patterns that signal intended help. Smart driving vehicles might also benefit from appropriately monitoring and predicting human gaze. For instance, if vehicles detect that drivers are losing attention to the road, they can prompt timely

to re-gain the drivers' attention. In addition to applications involving physical interactions, recommendation systems could provide better recommendations to users by utilizing their gaze patterns. For instance, an online shopping website could dynamically recommend products to customers by tracking and interpreting their gaze patterns.

Limitations

The present study has limitations that provide directions for future investigations. First, I employed SVMs for data analysis and modeling to quantify the potential relationship between gaze cues and intentions. Alternative approaches, such as decision trees and hidden Markov models (HMMs), may also be used to investigate such relationships and interaction dynamics. However, similar to most machine learning approaches that are sensitive to the data source, the results were subject to the interaction context and the collected data. Yet, in this study, I demonstrated that characteristics of gaze cues, especially duration and frequency, are a rich source for understanding human intentions.

Second, I formulated the problem of intention prediction as forecasting users' choices of item based on their gaze patterns. However, intention is a complex construct that may not be simply represented as the choice of item. While this study focused solely on using gaze cues to predict customer intent, predictions of intention might be benefited from incorporating additional features, including facial expressions and other cues from the customer, and other forms of contextual information, such as preferences expressed previously toward particular toppings or knowledge of what toppings might "go together." Disentangling the contributions of different features to observer performance in these predictions would significantly enrich our understanding of the process people follow to predict intent. Nevertheless, the findings were in line with literature indicating that gaze cues manifest attention and lead intended actions (Butterworth, 1991; Johansson et al., 2001; Land et al., 1999). Future work may also compare the performance of human observers and the types of errors they make to those of the presented machine

learning model. Such a comparison may inform design choices for feature selection and learning algorithms in building systems that recognize user intent.

Third, as discussed in Section 4.3.2, there were potential errors accumulated over the processes of tracking eye gaze, projecting gaze fixations, and predicting intended items. While the results of the HRI evaluation should be comprehended with this understanding of accumulated errors, this study presented a proof-of-concept system that was able to read users' gaze fixations, interpret what items they looked at, predicted their intended choices, and proactively prepare robot actions to respond accordingly. With future advances in sensory technologies and machine learning, anticipatory robots will emerge to serve people in daily activities.

4.4 Summary

Eye gaze is a rich source for anticipating a person's intentions. Robots can utilize users' eye gaze to anticipate their intentions and subsequently act proactively to achieve efficient joint action. In this chapter, I report my investigation into enabling such efficient human-robot joint action. My investigation started with developing a SVM-based approach to quantify how gaze cues may signify a person's intention. Using the data collected from a dyadic sandwich-making task, I demonstrated the effectiveness of the SVM-based predictor in making correct prediction of intention (76%) compared to the attention-based approach (65%) that only relied on the most recently glanced-at item. Moreover, the SVM-based approach provided correct predictions approximately 1.8 seconds before the requests, whereas the attention-based approach did not afford such intention anticipation. Qualitative analyses of the episodic interactions further revealed gaze patterns that suggested semantic meanings and that contributed to correct and incorrect predictions. These patterns informed the design of gaze features that offer a more complete picture of human intentions.

The predictive model linking the relationships between eye gaze and intention was incorporated into an anticipatory robotic system. The system involved tracking a user's gaze fixation in real-time, associating the fixations with real-world objects,

predicting the user's intentions based on the knowledge of a sequence of gazed objects, and triggering the robot's actions in response to the predicted user intentions to complete a joint task. A laboratory human-robot interaction experiment was conducted to evaluate the anticipatory system. The evaluation results showed that the anticipatory system led to a more efficient joint action with humans compared to the reactive system that did not anticipate user intentions. Moreover, participants rated the anticipatory system to have higher awareness of what they were intended to choose and attributed the anticipatory characteristics to perceived efficiency of the system.

In brief, this study demonstrated that how *action observation* might be realized for human-robot joint action. It provides insight into linking human intentions and gaze cues and offers implications for designing anticipatory robotic systems that aim to achieve efficient joint action with humans. The capacity of anticipation explored in this study was enabled solely by behavioral signals without any knowledge of the task structure. In the next chapter, I present a study that explored how *task-sharing*, in addition to action observation, might be realized to support *action coordination* during human-robot joint action.

5 TASK-SHARING & ACTION COORDINATION: ADAPTING ACTIONS TO ACHIEVE FLUID COLLABORATION

People not only monitor their partner's task actions as enabled by *joint attention* and *action observation* but also monitor their partner's task progress (*task-sharing*). Based on these observations, people adapt their actions to those of their partner (*action coordination*) to achieve seamless joint action. Such awareness and adaptivity are critical to team performance (Sebanz et al., 2006) and the psychological consequences of the interaction (Marsh et al., 2009).

5.1 Introduction

In this chapter, I focus on the mechanisms of *task-sharing* and *action coordination* in joint action and how to realize these mechanisms for robots to engage in joint action with humans[1]. My investigation was focused on the joint action of *handover*—the transfer of an object from a giver to a receiver—and contextualized in a household scenario, where a giver and a receiver work together to unload dishes from a drying rack. Handover is fundamental to a variety of joint actions that involves physical collaborations, such as passing tools among workers in manufacturing applications and handing medicine to patients in healthcare applications. While prior research has extensively studied the joint action of handover in human-human and human-robot interaction, I seek to better understand adaptation strategies that enable humans to seamlessly coordinate their actions and to explore how robots may leverage such strategies to more successfully engage in physical interactions with their users.

Next, I review prior work on handovers in human-human and human-robot interactions. I then present a study investigating how people coordinate their actions during a handover task and how a robot might utilize human-inspired

[1]Presented study results were previously published in Huang et al. (2015b).

coordination strategies to adapt to users in a similar handover task. This chapter concludes with a discussion and a summary of the study.

5.2 Related work

5.2.1 Designing adaptive human-robot teams

Hoffman and Breazeal (2007) developed a computational model that generated anticipatory actions for an assistive agent to enable the agent to adapt to its user's workflow in a simulated assembly scenario. They showed that an agent using anticipatory actions, compared to a reactive agent, formed a more fluid team with its users, resulting in greater concurrent activity with them. Inspired by coordination behaviors of human teams, Shah et al. (2011) developed a plan execution system called *Chaski* that adapts to human partners and seeks to minimize its partners' idle time.

5.2.2 Designing handovers for human-robot teams

Handovers involve the transfer of objects from a giver to a receiver and serve as a fundamental skill for complex collaboration and interaction. Prior work has investigated various aspects of *human-human handovers*, including velocity, force, and style, in order to inform the design of human-robot handovers. For instance, Becchio et al. (Becchio et al., 2010) found that people display different velocity profiles when they hand over an object compared to when they place the object on a surface. Their characterization of handover motion profiles was consistent with earlier observations, which showed that people display minimum-jerk motions during handovers (Huber et al., 2008; Shibata et al., 1997). Other work modeled grip force patterns during handovers to design handover controllers (Chan et al., 2012). To design a robot that more effectively delivers flyers, Shi et al. (Shi et al., 2013) studied different styles of handovers that shopkeepers used when distributing flyers to passersby in a mall. While these studies provide useful insights into the different

facets of handovers, how people coordinate handover actions in situations where the receiver is distracted or delayed by a secondary task—a common occurrence in everyday situations—is unknown.

Human-robot handovers offer a rich design space with a large number of parameters. Research to date has explored the role of gaze (Moon et al., 2014), approach angle and saliency (Unhelkar et al., 2014), contrast between start and end points of handover motion (Cakmak et al., 2011b), and anthropomorphism (Sivakumar et al., 2013). Moreover, prior work developed methods to choose handover parameters such as pose and trajectory that consider user preferences (Cakmak et al., 2011a), user comfort (Aleotti et al., 2012), object affordances Aleotti et al. (2014); Chan et al. (2014), and user mobility constraints (Mainprice et al., 2012) that facilitate human-robot handovers. Although most work focused on a robot handing objects to humans, how robots may take objects from humans has also been investigated (Aleotti et al., 2012; Edsinger and Kemp, 2007). Previous research has also explored how robots may utilize handovers across a number of application scenarios, including manufacturing (Koene et al., 2014; Unhelkar et al., 2014), household service (Edsinger and Kemp, 2007; Mainprice et al., 2012), or serving drinks (Cakmak et al., 2011b; Grigore et al., 2013). Our work contributes to the exploration of the rich design space for handovers by investigating the effects of task demand on handover actions and developing new methods to enable robots to adapt to their users' changing task demands.

5.3 Study 5: designing adaptive strategies for object handovers

This study investigated how a robot might adapt to users' task demands to achieve fluid collaboration with them by utilizing the mechanisms of task-sharing and action coordination. This study involved modeling how people adapt to each other in the joint action of handover, developing human-inspired adaptive strategies for a

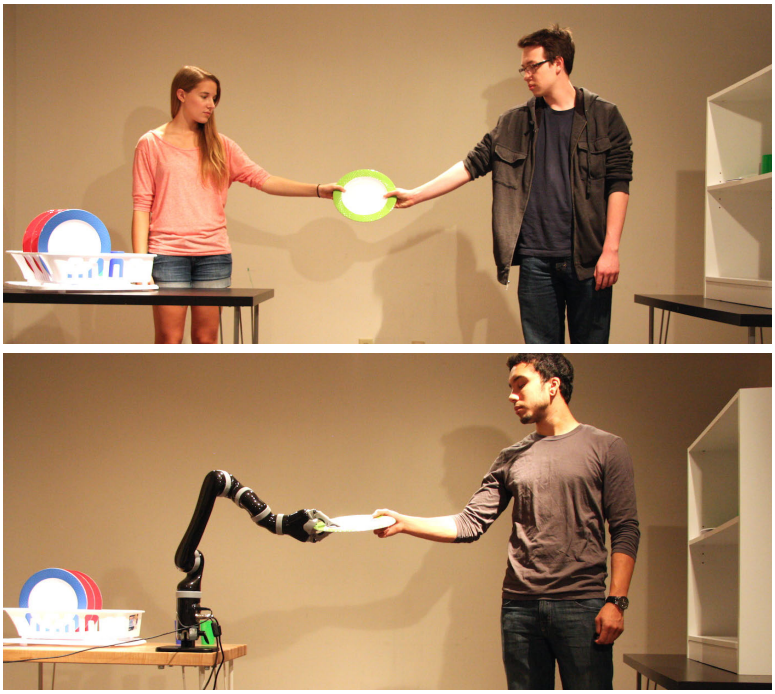


Figure 5.1: I studied human-human handovers (top) in a household scenario, identified strategies that humans used for coordination, implemented them on an robotic manipulator, and evaluated their effectiveness in supporting coordination in human-robot handovers (bottom).

robotic manipulator, and evaluating the developed system with users in the context of robot-human handover.

5.3.1 User modeling

In this section, I describe a data collection study of human-human handovers that informed the design of adaptive strategies to facilitate human-robot handovers.

Data collection

Task & setup — To better understand how people adapt handover actions to task demands, I observed pairs of human participants collaboratively unload a dish rack (Figure 5.1, top). The task required one participant, the *giver*, to pick up plates

and cups in a drying rack and hand them to the other participant, the *receiver*, who then placed these items on a nearby shelf. Participants performed two variations of this task. In the first variation, both the giver and receiver were engaged only in unloading dishes, denoted by *regular unloading* hereafter. In the second variation, in addition to unloading dishes, the receiver was given a secondary task of matching patterns on the items to specified target locations, denoted by *tasked unloading* hereafter. I introduced the secondary task in order to add to the receiver's cognitive load (Stroop, 1935), thereby increasing task demand, such that the giver would need to adapt to the receiver's workload. Each dyad performed both task variations once. The roles of giver and receiver within a dyad were maintained throughout the interaction. Positions of participants' body joints were recorded using Microsoft Kinect version 2. The interactions were also video-recorded. Participants received $5 USD for their participation in the study.

Participants — Eight dyads, two for each gender combination, were recruited. Participant ages ranged 20–34 ($M = 23.94, SD = 4.58$). All dyads but one included participants who did not know each other prior to this study.

Data processing

Overall, from the collected data I observed that givers (1) monitored the receivers' task progress, especially in the *tasked unloading* task, and (2) adapted to the receiver's pace by pausing and/or slowing down their actions. These observations informed the following analysis for further understanding coordination strategies in handovers.

Annotating user state — I categorized the handover activity into six different states for the giver and the receiver. The giver's states include (1) *reach*: moving hand from a resting position or the center of body to grasp an object; (2) *retrieve*: grasping and moving the object to the center of body; (3) *give*: moving the object from the center of body to the handover position; (4) *handover*: both the giver and receiver touching the object and the giver releasing the object; (5) *retract*: moving the hand back to the center of body; and (6) *idle*: all other actions. Similarly, the receiver's states include (1) *take*: moving hand from the resting position or the

center of body to the handover position; (2) *handover*: both the giver and receiver touching the object and the giver releasing the object; (3) *retrieve*: the giver releasing the object and the receiver moving it to the center of body; (4) *place*: moving the hand from the body center to hand off of the object; (5) *retract*: returning hand to the body center or a resting position; and (6) *idle*: all other actions.

The video data was annotated with these states for a detailed analysis of handover actions and for training an algorithm for online prediction of user states during interaction with a robotic manipulator (see Section 5.3.2). The data included 8389 and 12793 joint readings collected in the regular and tasked unloading conditions, respectively. A primary rater coded all of the data, and a secondary rater coded 10% of the data. Inter-rater reliability analysis showed substantial agreement between the raters (Cohen's $\kappa = .74$) (Landis and Koch, 1977).

Smoothing sensor data — I applied an Exponentially Weighted Moving Average (EWMA), a common noise reduction technique for time-series data defined in Equation 5.1, to reduce noise in the raw joint position data from the Kinect sensor.

$$\mathbf{x}_t = \alpha\, \mathbf{y}_{t-1} + (1-\alpha)\, \mathbf{x}_{t-1} \qquad \text{for } t > 0,\ \mathbf{x}_0 = \mathbf{y}_0 \tag{5.1}$$

where $\mathbf{y}_t$ is the raw sensor measurement of the joint positions in Cartesian space at time t, and $\mathbf{x}_t$ is the filtered measurement at time t. The weighting parameter α controls how much to discount prior data, which was determined to be 0.2 for best performance during testing. The results reported below were based on analyses using smoothed data.

Adaptive strategies

To understand how people coordinate their actions, we first need to know *when* coordination strategies are needed. An inspection of the average durations of receiver actions in each state in the regular and tasked conditions, as illustrated in Figure 5.2, showed that receivers stayed in the *retrieve* and *place* states longer in the tasked condition than in the regular condition. These differences suggest that these states were likely to be when givers had to adapt to the availability of the receivers.

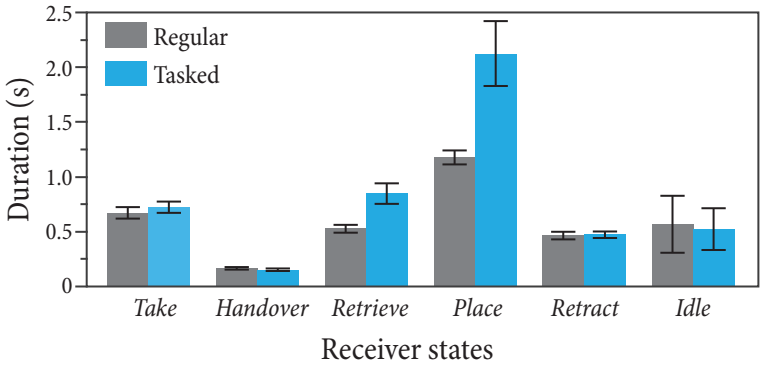


Figure 5.2: Average durations of the receiver's states in the regular and tasked conditions. Receivers took longer in the *retrieve* and *place* states, suggesting that these were the states in which the giver had to adapt their actions to the receiver's availability. Error bars indicate 95% confidence intervals.

The analysis next focused on giver behaviors in order to better understand the adaptations they displayed (Figure 5.3), identifying two adaptive strategies: *waiting* and *slowing down*.

Waiting strategy — In the tasked condition, the giver adapted his/her actions by *waiting* for the receiver to complete the secondary task before resuming unloading dishes. The giver was found in this idle state a total of 79 times (across 96 tasked trials). Waiting was most commonly observed, a total of 35 times, after the giver retrieved the item but before passing it to the receiver (e.g., keeping the item in front of body and ready for handover). The giver also waited after retracting from the handover pose but before reaching for the next object (24 times), and after reaching toward but before retrieving the object (17 times). In rare occasions, the giver waited in the handover pose (3 times). This waiting strategy is consistent with prior work that reported the giver pausing his/her action until the receiver was ready for handover (Lee et al., 2011).

Slowing-down strategy — In addition to the waiting strategy, I observed that the giver adapted to the receiver's availability by *slowing down* his/her actions while the receiver was occupied by the secondary task. I further inspected the velocity of the giver's hand across the two task conditions (Figure 5.4) and found it to be

Figure 5.3: Velocity profiles and example snapshots from adaptive coordination strategies—slowing down (top) and waiting (bottom)—displayed by givers in human-human handovers.

slower during the *retrieve*, *give*, and *retract* states in the tasked condition than in the regular condition. These observations confirmed that the giver slowed down actions to adapt to the receiver's availability and highlighted the states in which the giver slowed down.

I note that people used these two strategies in a combined and interchangeable way, as shown in Figure 5.3, and that slowing down usually occurred prior to waiting, indicating that givers slowed down their action and then paused if necessary.

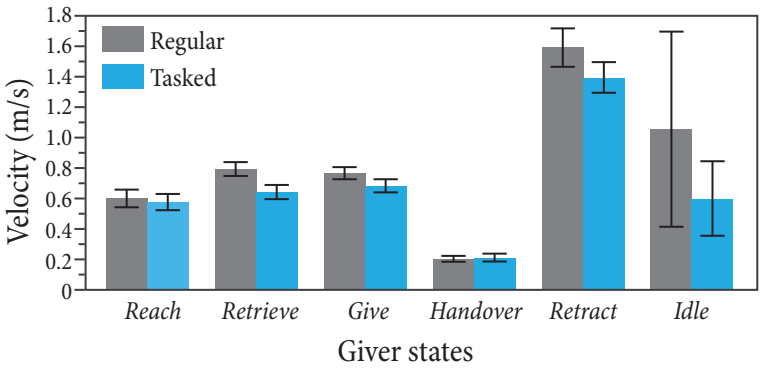


Figure 5.4: Average velocities of the giver's hand across different states in the regular and tasked conditions. Givers slowed down their actions during *retrieve*, *give*, and *retract* states. Error bars indicate 95% confidence intervals.

5.3.2 System implementation

In this section, I describe the design of coordination strategies to enable adaptive human-robot handovers based on the strategies identified in the human-human interaction study. The goal is to develop a human-aware controller that adapts to the availability of the user in planning robot actions for handover. To that end, I developed an autonomous system that uses human joint positions to predict the state that the user is currently in and determines what action the robot should take accordingly.

Predicting user state

Coordination in joint action requires knowledge of the ongoing actions and the current states of interaction partners (Sebanz et al., 2006). To inform the robot of user actions and states, I employed a Microsoft Kinect version 2 camera to track the user's body joints, extracted features from the body joints that represented characteristics of the user's current action, and predicted the user's current state using a K-Nearest Neighbor (KNN) algorithm. I provide details of this process below.

Temporal and Spatial Features — Prior to extracting features, I applied an EWMA filter (Equation 5.1) to the raw joint data in order to smooth out sensor noise. Using the filtered data, I derived the following features to represent the spatial and temporal state of the user.

- *Hand velocity* (Δx_t^{hand}): This feature captures the velocity of the user's active hand. Note that the user might use different hands in different states. For instance, the receiver may use the right hand to take the object and the left hand to place it. This feature was calculated using equation:

$$\Delta x_t^{hand} = \frac{x_t^{hand} - x_{t-1}^{hand}}{\Delta t} \tag{5.2}$$

  where $x_t^{hand} \in \mathbb{R}^3$ is the Cartesian position of `hand_left` or `hand_right` joints in the Kinect joint structure, and Δt is the duration between sensor readings (about 30ms).

- *Extension* (x_t^{ext}): This feature represents the extension of the user's arm as a vector from the origin of the body, denoted as x_t^{origin}, to the active hand.

$$x_t^{ext} = x_t^{hand} - x_t^{origin} \tag{5.3}$$

  The position of the `spine_mid` joint is used as x_t^{origin}.

- *Approach* (x_t^{app}): This feature characterizes the extent to which the user's arm approaches the other agent as a vector between the active hand and the midpoint of the body centers of the two parties. The basis of this feature was the observation that handovers happen approximately at the midpoint between two parties (Basili et al., 2009).

$$x_t^{app} = x_t^{hand} - x_t^{midpoint} \tag{5.4}$$

$$x_t^{midpoint} = \frac{x_t^{origin} + x_t^{origin\_other}}{2} \tag{5.5}$$

The state of the user is represented with a feature vector consisting of these three features, $f = (\Delta x_t^{hand}, x_t^{ext}, x_t^{app})$.

State prediction — I used a KNN classifier to predict the user's current state based on the features described above. When a new observation arrives in the form of a feature vector, the algorithm finds the K most similar instances in the training dataset (the annotated data from the human-human interaction study) according to the distance measure in Equation 5.6.

$$d(f, \bar{f}) = \|\Delta x_t^{hand} - \overline{\Delta x_t^{hand}}\| + \|x_t^{ext} - \overline{x_t^{ext}}\| + \|x_t^{app} - \overline{x_t^{app}}\| \tag{5.6}$$

where f denotes the new observation to be classified; $\bar{f}$ denotes a sample in the training dataset; and the distance between individual feature pairs are the L_2 norms of the difference vector. The weights of the features in f were chosen to be equal, indicating that each feature contributes to the prediction equally, based on our preliminary testing. I set K to be 35 for this application, i.e., the 35 most similar instances in the annotated data formed a group of candidate predictions. The prediction of the user's state was based on the majority vote of the 35 candidates. The algorithm's confidence for each prediction was calculated using Equation 5.7.

$$Confidence = \frac{number\ of\ majority\ classifications}{K} \tag{5.7}$$

I calculated average confidence scores for correct and incorrect predictions, shown in Table 5.1, and used these scores as thresholds to filter out predictions with low confidence.

Parameter selection & evaluation — The parameters, including α for the EWMA filter and K for the KNN algorithm, were tuned based on eight-fold leave-one-out cross-validation, using seven dyads for training and the remaining dyad for testing. Since the two datasets—*regular* and *tasked* unloading—yielded similar results (0.4% difference in accuracy), I chose to use the best parameters learned from the regular

Table 5.1: Confidence scores for KNN-based prediction of receiver states.

		Take	Handover	Retrieve	Place	Retract	Idle	Dropping %
Regular	*Correct*	92.03	68.07	89.77	94.39	90.27	83.99	38.44%
	Incorrect	72.96	64.64	70.11	75.52	71.13	75.94	17.80%
Tasked	*Correct*	89.93	61.74	88.02	92.46	86.84	92.95	40.64%
	Incorrect	68.29	56.25	65.29	74.16	69.22	62.97	16.62%

dataset. To assess the effectiveness of the KNN model in predicting the receiver's state, I conducted another eight-fold leave-one-out cross-validation using the tasked dataset with the chosen parameters (Figure 5.5). The results of this test showed that our KNN model more accurately predicted receiver states than two baselines—*chance* and *most common guess* (Admoni and Scassellati, 2014)—did. The *chance* baseline involves random guesses of user state, and the *most common guess* baseline always predicts the state that most commonly appeared in the training data.

In addition to the ordinary use of KNN, I explored confidence thresholding using the thresholds of incorrect predictions. This method resulted in a tradeoff

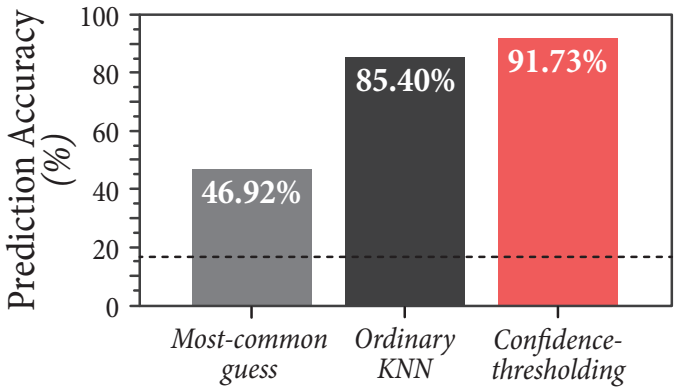


Figure 5.5: Cross-validation results of our KNN model in predicting receiver states using the *tasked* dataset. The dashed line indicates baseline accuracy (16.67%).

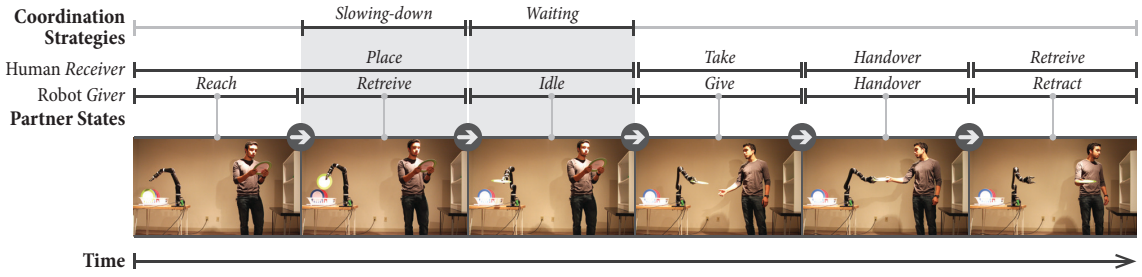


Figure 5.6: An example human-robot handover using our method for adaptive coordination that employed the *waiting* and *slowing-down* strategies.

between improved accuracy and the number of dropped predictions; the results showed that confidence thresholding improved accuracy by approximately 6%, although approximately 17% of the predictions were dropped. The dropped predictions usually occurred during transitions between states. For the human-robot interaction study, reported in Section 5.3.3, the algorithm dropped 15.35% of the predictions. The average duration of dropped predictions was 198.68 ms, while the algorithm made a prediction every 46.4 ms. I found the rate and duration of dropped predictions to be acceptable in our application, as the algorithm provided predictions at a high frequency.

Generating robot actions

The adaptive strategies that were observed in human-human interactions informed the development of a model to emulate human-style adaptive coordination, as outlined in Algorithm 3. Here, `userState` is provided by the KNN model described above. The `delayThreshold` parameter was set to 1.36 seconds, which was the average duration for people to finish retrieving and placing an object when no secondary task was present. In Line 1 of Algorithm 3, the robot's waiting position is determined based on a probability distribution obtained from our data on human-human interactions (Section 5.3.1). In Line 3, `currentRobotState` is specified by a deterministic finite state machine that represents the sequence of giver states. This model was implemented in ROS (Quigley et al., 2009) to control a Kinova MICO

robotic arm, shown in Figure 5.1. I used force-sensor information from the robot's joints to determine whether or not items in the robot's gripper were grasped by the user and to plan for their release.

5.3.3 Experimental evaluation

In this section, I report on a user study of the effects of adaptive coordination on objective and subjective outcomes of human-robot handovers. In particular, I evaluated the effectiveness of the adaptive strategies described above in improving task performance and user perceptions in handover interactions where a robotic manipulator passed objects to human participants in a common household scenario.

Hypothesis

This evaluation tested the central hypothesis stated below regarding the effects of a robot's use of human-inspired adaptive strategies on team performance and perceptions of the robot under different levels of task demand.

Algorithm 3 Adaptive Coordination Strategy for Handover

Require: userState, delayThreshold
probabilistically select a robot waiting position
while isCurrentHandoverTrialActive( ) **do**
    currentRobotState ← getRobotState( )
    robotAction ← getRobotAction(currentRobotState)
    if userState = *retrieve* or *place* **then**
        if ElapsedStateTime( ) ⩾ delayThreshold **then**
            if isRobotAtWaitingPosition( ) **then**
                wait( )
            else
                slowExecution(robotAction)
            end if
        end if
    end if
    regularExecution(robotAction)
end while

Hypothesis — When users are under high levels of task demand, the robot employing adaptive strategies that enable it to adapt its handover actions to its user's task will improve team performance and user experience with and perceptions of the robot, while employing these strategies will not offer similar benefits when users are under low levels of task demand.

Experimental design, task, & conditions

To test the above hypothesis, I conducted a 3×2 within-participants study in which I manipulated the *coordination method* that the robot employed and the level of *task demand* under which the participants worked. The paragraphs below describe the three coordination methods considered in the study.

Proactive coordination — Following this method of coordination, the robot did not take the user's task demand into account when planning its actions. After handing an object to the user, the robot *proactively* fetched the next object and presented it to the user, even if the user was not ready to take the next object. Using this method, the robot aimed to minimize the user's idle time and to maximize concurrent activity with the user, measures proposed by prior work as objective indicators of fluid teamwork (Hoffman, 2013; Nikolaidis et al., 2013; Shah and Breazeal, 2010; Shah et al., 2011).

Reactive coordination — In contrast to the proactive coordination, the robot following this method waited for its user to fully complete the task, thus *reacting* to the user's availability. When the user returned to an idle state, the robot fetched the next object. This method enabled a turn-based interaction in which interaction partners asynchronously contributed to the task. Prior work has characterized this method of coordination as a form of joint action (Mutlu et al., 2013).

Adaptive coordination — The robot achieved adaptive coordination by following Algorithm 3, which enabled the robot to utilize the *waiting* and *slowing-down* strategies to adapt to its user's task demand.

Moreover, I introduced a manipulation to the experimental task to create two levels of task demand as described below.

Low task demand — In this setting, participants were asked to engage in the task of unloading dishes, where the robot passed four plates and two cups to its user, and the user placed the objects on a shelf, as shown in the bottom of Figure 5.1. Because people can take and place an object faster than it takes the robot to fetch and deliver the object, this setting creates a lower level of task demand for the user.

High task demand — In this setting, in addition to the task of unloading dishes, participants were engaged in a secondary task, which involved solving a math problem attached to each object and placing the object in a location on the shelf that corresponds to the answer. Each problem consisted of nine single-digit numbers and involved all four arithmetic operations (i.e., addition, subtraction, multiplication, and division) twice (e.g., $7 \times 2 - 2 + 6 \div 3 - 9 + 2 \times 8 \div 4$). These problems were designed to ensure that each problem had a similar level of difficulty. The same set of 18 unique math problems was used for each participant. The problems were randomly assigned to objects and to coordination methods in order to prevent any systematic bias due to differences in problem difficulty. In each round of interaction, nine potential answers, six of which were correct for the six handover actions, were provided along with the option "not able to find the answer." The design of this task aimed to create a higher level of task demand for the user.

The three coordination methods and two task-demand levels comprised six experimental conditions. The robot followed the same pre-programmed motion trajectories across all conditions.

Measures

I used objective and subjective measures to assess the effectiveness of the three coordination methods. Objectively, I measured *team performance* as informed by prior work (Hoffman, 2013; Nikolaidis et al., 2013; Shah et al., 2011). In particular, I measured task completion time (the time between the robot's first move to pick up the first object and the user placing the last object), concurrent activity (the proportion of the total time that both the user and the robot were in action to the total task completion time), user idle time (the proportion of task completion time to total time of inaction by the user or the robot), and robot idle time.

In addition to objective measures, I developed four scales—*fluency*, *intelligence*, *awareness*, and *patience*—to measure participants' experience with and perceptions of the robot. The fluency scale extended a previously proposed measure of subjective fluency (Hoffman, 2013) and consisted of five items (Cronbach's $\alpha = 0.91$). The scales of intelligence, awareness, and patience consisted of four (Cronbach's $\alpha = 0.88$), two (Cronbach's $\alpha = 0.83$), and four (Cronbach's $\alpha = 0.85$) items, respectively. All items were on a seven-point rating scale (1 = Strongly Disagree, 7 = Strongly Agree).

Procedure

Following informed consent, participants spent a minute to review the order of operations and the multiplication table prior to beginning the task. The study involved six rounds of interaction—one round for each condition. The order of coordination methods and levels of task demand were counterbalanced. After each round of interaction, participants filled out a questionnaire regarding their experience with and perceptions of the robot. Finally, the experimenter conducted a post-experiment interview. The study took approximately 50 minutes. All participants received $10 USD as compensation.

Participants

A total of 26 participants were recruited from the local community. However, two participants were excluded from the data analysis, as one of the them did not finish the task, and the other one failed to follow instructions. The resulting 24 participants (13 females, 10 males, and one unspecified) were aged between 18 and 35 years ($M = 20.54, SD = 3.72$) and reported little familiarity with robots ($M = 2.79, SD = 1.47$ on a seven-point scale). None of the participants in this study had taken part in the human-human data collection study.

Results

The analysis of the data from each measure involved a two-way repeated-measures analysis of variance (ANOVA), using coordination method, task demand, and their interaction as independent variables and participant ID as a random vari-

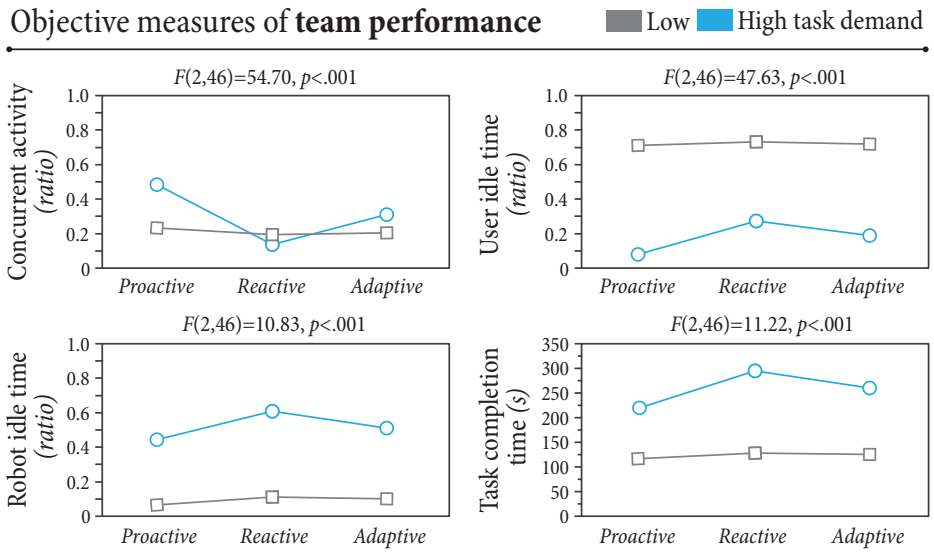


Figure 5.7: Interaction plots and ANOVA test details for objective measures of team performance.

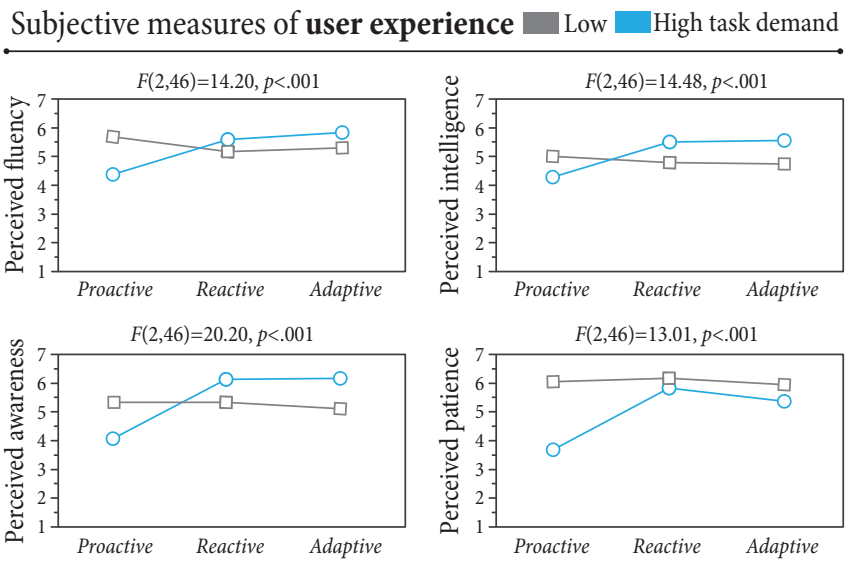


Figure 5.8: Interaction plots and ANOVA test details for subjective measures of user experience.

able. Guided by the hypothesis, the analyses focused the interaction between coordination method and task demand. Six *a priori* pairwise comparisons, using a Bonferroni-adjusted α level of .008 (.05/6) for significance, were carried out to identify differences across conditions for each objective and subjective measure. For readability, details of statistical tests are omitted from the text, and test details for interaction effects and pairwise comparisons are provided in Figure 5.7–5.10.

Objective measures. Across all objective measures, the analyses revealed significant interaction effects between coordination method and task demand. At lower levels of task demand, pairwise comparisons showed no differences in any measure of team performance across coordination methods. In contrast, at high levels of task demand, the coordination method that the robot used had a significant effect on the task performance of the human-robot team in measures of task completion time, concurrent activity, user idle time, and robot idle time. In particular, the results showed that, among the three methods, proactive coordination yielded the greatest outcomes in our task performance measures, followed by adaptive coordination, while reactive coordination resulted in the poorest outcomes.

Subjective measures. Similar to the objective measures, the analyses revealed significant interaction effects between coordination method and task demand for all subjective measures. Comparisons showed that, across different coordination methods, there were no significant differences in participants' perceptions of team fluency and of the robot in terms of intelligence, awareness, and patience when they were under high levels of task demand. However, differences in subjective measures emerged when participants were under high levels of task demand during their interactions with the robot.

Interestingly, while objective outcomes of team performance in measures of concurrent activity, user idle time, and robot idle time indicated a better performance when the robot used the proactive coordination method, the subjective measure of fluency indicated otherwise. Participants perceived their interactions with the robot using the proactive coordination method to be less fluid compared to the other two coordination methods. Both reactive and adaptive coordination yielded a similar degree of perception of team fluency. The analyses of the data from measures of

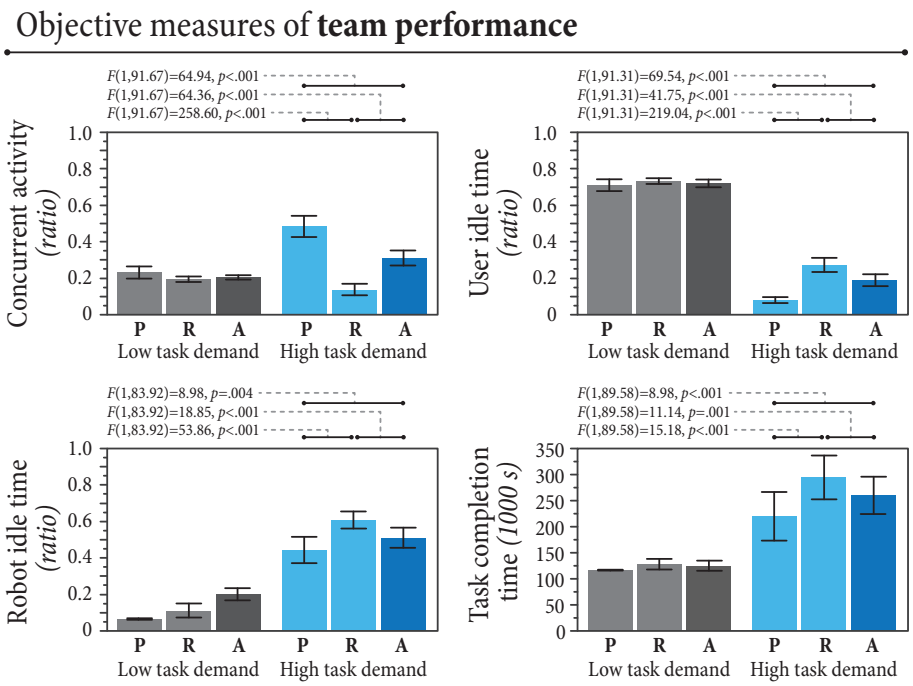


Figure 5.9: Data from measures of team performance. P, R, and A represent the *proactive*, *reactive*, and *Adaptive* coordination methods, respectively. Pairwise comparisons use a Bonferroni-adjusted α level of .008 for significance. Error bars indicate 95% confidence intervals.

perceived intelligence, awareness, and patience of the robot were consistent with these findings.

5.3.4 Discussion

I identified two adaptive strategies, *waiting* and *slowing down*, from data on human-human handovers in a household application scenario and developed an adaptive coordination method involving these two strategies. I implemented a robot system that autonomously performed handovers with users in a similar scenario and evaluated the effectiveness of the adaptive coordination method against two alternative methods in facilitating human-robot handovers. The results showed a tradeoff

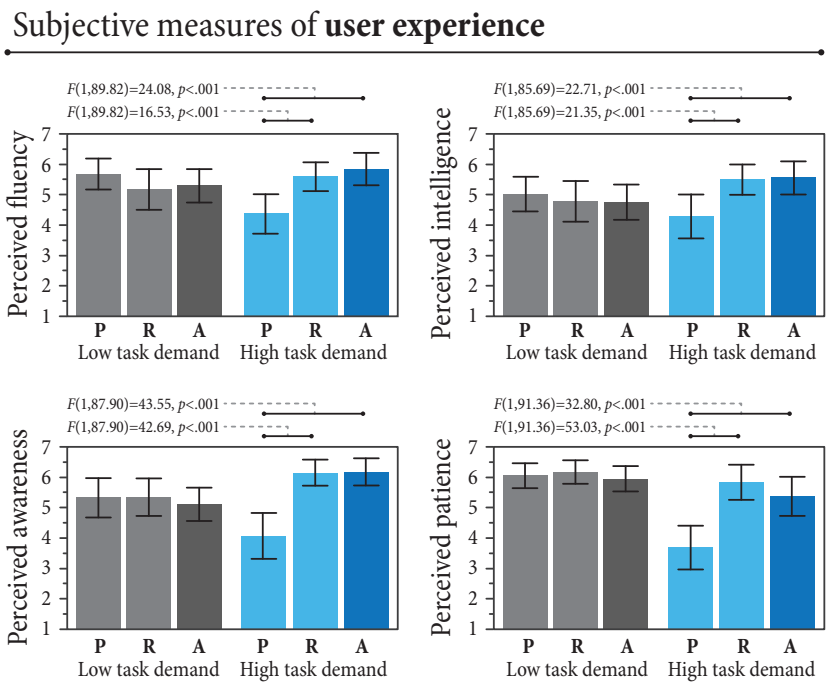


Figure 5.10: Data from measures of user experience. P, R, and A represent the *proactive*, *reactive*, and *Adaptive* coordination methods, respectively. Pairwise comparisons use a Bonferroni-adjusted α level of .008 for significance. Error bars indicate 95% confidence intervals.

between team performance and user experience. Below, I discuss the results and their implications for enabling human-robot collaboration.

Design implications

The results showed that the effects of coordination methods were differentiable only when participants' task demand was greater than that of the robot, suggesting that a robotic assistant should monitor the task progress of its user and employ coordination methods as appropriate and when necessary. The results further revealed a tradeoff between task performance and user experience based on the coordination method employed. While proactive coordination significantly improved team performance in measures of concurrent activity and idle time, it impaired users' experience with and perceptions of the robot. In line with prior work, this

result indicates that task efficiency and user perceptions of team fluency are not necessarily positively correlated (Hoffman and Breazeal, 2007). It also suggests a lack of correlation between perceived and objective measures of team fluency, as suggested by prior work on human-robot collaboration (Hoffman, 2013; Shah et al., 2011). I therefore argue that team fluency is a complex construct that could not be defined only objectively but that it requires a consideration of partners' perceptions and their task demands.

Team experience and performance are equally essential elements of joint action (Marsh et al., 2009). I found that proactive coordination improved team performance but hurt user experience. I speculate that this result was due to the pressure that the proactive robot may have imposed on users to complete their secondary task, as suggested by the excerpt below.

> *"It [the robot] moved really quickly and I was holding one plate still trying to figure out the problem and feeling like I have to pick up the next plate, so that was stressful."*

Conversely, reactive coordination improved user experience but reduced team performance, as indicated below.

> *"It [the robot] waited until I put the plate down to move again, hmm and that was almost too slow because then I'd have finished the math problem and be waiting to get the next one."*

Finally, adaptive coordination improved team performance while maintaining the support that reactive coordination provided for user experience, as predicted by prior research on human-human joint action (Marsh et al., 2009; Sebanz et al., 2006).

Limitations

This present study has limitations that motivate future investigation. First, I focused on a particular handover application that involved interaction partners unloading dishes. While I speculate that the adaptive strategies of waiting and slowing-down

will be applicable to other applications, the parameters that I derived from the collected data may not be applicable to new applications. This work, however, illustrates a process for building coordination mechanisms for interactive robots. Second, although I demonstrated the potential of a KNN algorithm in predicting user state, its effectiveness and efficiency depend on the size of the training dataset. Although a larger dataset promises better accuracy, it may decrease efficiency due to the number of comparisons that the algorithm has to perform on the fly. Future work can explore alternative models such as decision trees (e.g., (Strabala et al., 2012)) and probabilistic graphical models (e.g., (Grigore et al., 2013)) for effective, efficient prediction of user states. Finally, the robotic manipulator used in this work placed constraints on the different characteristics of the motions it produced, including trajectories, velocity, and noise.

5.4 Summary

In human physical collaborations, interaction partners adaptively coordinate their actions in order to achieve more fluid interactions and greater team performance and experience. In addition to *action observation*, the study presented in this chapter exploited *task-sharing* and *action coordination* in achieving adaptive coordination between the partners.

I contextualized the study in a household application involving partners engaged in handovers to collaboratively unload a dish rack. I first modeled how people adapted their handover actions to the workload of their partners in the household application, identifying two strategies—slowing-down and waiting—people used to achieve fluid adaptation in the joint action. I implemented an autonomous robotic system that took into consideration a real-time awareness of the task status of its user and adapted to its users if necessary in performing handover actions. This awareness was enabled by the implementation of *action observation* and *task-sharing*. Particularly, task-sharing was realized by explicitly representing the partner's task. For example, the robot giver knew that after *taking* over an object the human receiver should *place* the object to the target place. I then conducted a laboratory human-

robot interaction experiment to investigate how the autonomous robotic system interacted with people in a similar household application.

The results from the HRI experiment showed that there was a tradeoff between team performance and user experience. Specifically, I found that solely maximizing team performance (e.g., reducing idle time) impaired user experience. The results highlighted the importance of a joint consideration of team performance and user experience in enabling effective human-robot joint action.

6 GENERAL DISCUSSION

In this chapter, I provide discussions on the lessons I learned from the results from Studies 1–5 (Section 6.1), limitations of this research (Section 6.2), and directions of future work (Section 6.3).

6.1 Lessons learned

6.1.1 Awareness in joint action

Awareness of the interaction partner is vital to joint action. I want to particularly focus my discussion on awareness of the partner here because awareness of the environment is critical even without involving interaction with others. The coordination mechanisms of *joint attention*, *action observation*, *task-sharing*, and *action*

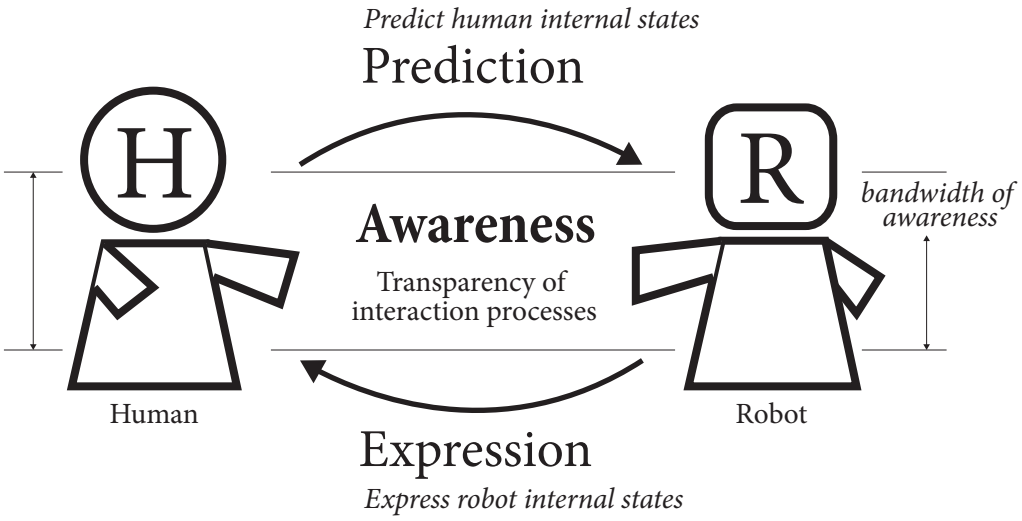


Figure 6.1: The quality of joint action largely depends on the *bandwidth of awareness* between the interaction partners. Two processes—expression and prediction—help increase the bandwidth of awareness. From a robot's point of view, the process of expression is to use behavioral means to reveal internal states to the human partner. The process of prediction is to understand and predict the partner's internal states by monitoring his/her behavioral signals.

coordination explored in this research contribute to increasing awareness of the interaction partner. I would argue that awareness of the partner characterizes the problem of joint action. I would further argue that the quality of joint action largely depends on the *bandwidth of awareness* between the interaction partners (Figure 6.1).

The bandwidth of awareness in joint action is modulated by two processes—*expression* and *prediction*. Expression is the process to use perceivable signals to reveal internal states in a way that are more predictable and comprehensible to the other partners (e.g., Pezzulo et al. (2013)). For example, people use gaze cues to show the partner their current interest, subsequently eliciting an attentional shift from the partner to the joint interest and therefore achieving perceptual common ground. The process of prediction is to comprehend and predict the partner's internal states using perceived expressional cues. People voluntarily and cooperatively participate in these two processes to establish common ground during joint action (Clark, 1996).

The five coordination mechanisms of joint action proposed by Sebanz et al. (2006) arguably revolve around this concept of *awareness* in joint action. *Joint attention* supports the process of expression to signal attention and intention. *Action observation* and *task-sharing* facilitate the process of prediction to understand the partner's action intent and anticipate later actions. These three coordination mechanisms together increase the bandwidth of awareness between the partners and therefore allow for *action coordination*. Finally, *perception of agency* concerns how partners can be aware of their actions and resulting effects in seamless joint action.

In this research, I explored different means to increase the bandwidth of awareness between humans and robots to support effective human-robot joint action. In Studies 1–3, I leveraged the robot's embodied characteristics, particularly gaze and gestural cues, to invite participants to read and align with the robot's internal states. In Study 4 & 5, I utilized an eye-gaze tracker and a Kinect sensor to perceive the participants' behavioral signals, which allowed for the robot to understand and predict their internal states. Moreover, the robot's anticipatory actions (Study 4) and slowing-down and waiting behaviors (Study 5) provided clues to the participants

about its intentions, which participants were able to pick up and attributed them accordingly.

In particular, the findings that people followed the robot's referential gaze cues to task-relevant objects (Study 1) and that people unintentionally signaled their intentions via gaze cues (Study 4) suggested that robot systems aim to increase the bandwidth of awareness during their joint actions with humans should skillfully leverage people's inclination to processing and producing social signals.

Timing is crucial in the processes of expression and prediction. Social signals are only relevant and meaningful when they are produced and perceived at the right time. For example, Studies 1–3 showed that behavioral cues were closely coupled in time and could effectively elicit immediate effects from the interaction partner, such as attentional shifts. Study 4 demonstrated that how anticipation of future events can be possible by processing and analyzing perceived behavioral signals in a timely fashion and that how a robot might respond quickly with that anticipation. Study 5 illustrated that incorporating the partner's work pace into timing the robot's actions could create an adaptive human-robot team. Together, these results suggested the importance of timing in human-robot joint action, especially in producing the robot's actions and in perceiving and predicting the human partner's actions. To further support such careful management of time during joint action, new advances in computing technologies and predictive algorithms are needed.

6.1.2 Evaluation of joint action

The complex, multifaceted nature of joint action makes it challenging to have accurate metrics to evaluate the quality of joint action. While evaluation guidelines for human-robot collaboration (Hoffman, 2013), or general human-robot interaction (Steinfeld et al., 2006), have been proposed, it remains an open problem to quantify what successful joint action is. Having standardized and quantifiable metrics is important not only for benchmarking the performance of current human-robot systems but also for advancing the field as it would allow researchers to compare different methods and systems.

Table 6.1: Objective measures for estimating the quality of human-robot joint action, their descriptions, and in which present studies they were employed.

Objective measures		
Team performance		
Idle time	The duration when interaction partner is not working toward the task goal. This measure can be further devided into user idle time and robot idle time.	Study 5
Concurrent action	How much time when interaction partners both are working toward the task goal.	Study 5
Response time	The time the robot needs to respond to the user's request or action.	Study 4
Predition accuracy	The accuracy of the robot's predictions about the user's intentions or task-relevant actions.	Study 4 & 5
Task performance		
Task time	The time needed to complete the joint task.	Study 1, 4, & 5
Correctness of task	How accurate the task was completed. In an educational/training context, this measure could be information recall.	Study 1, 2, & 3

Despite the lack of a benchmark for human-robot joint action, several objective, subjective, and behavioral measures (e.g., Hoffman (2013)) could be used to estimate the quality of joint action. Below, I summarize and discuss measures used in this research. Based on the findings from this research, I also discuss potential measures for evaluating human-robot joint action.

Objective, subjective, & behavioral measures

Objective measures. Objective measures were used to quantify various aspects of *team* performance (e.g., idle time, concurrent action, and response time) and *task* performance (e.g., task completion time and correctness of task) (Table 6.1). The *efficiency* of a human-robot team could be gauged by team idle time, team concurrent action, and the total task time. These measures are likely to be related. Little idle time possibly leads to more concurrent actions and thus yielding a shorter task

Table 6.2: Subjective scales for estimating user experience in human-robot joint action, their descriptions, and in which present studies they were employed.

Subjective measures		
Perceptions of the robot		
Naturalness	How natural the robot's behaviors were	Study 1, 2, & 3
Awareness	How aware the robot was about the user	Study 4 & 5
Likability	How much the user like the robot	Study 1 & 3
Competence	How competent the robot was in performing that task	Study 1, 2, & 3
Interaction experience		
Fluency	How fluid the interaction with the robot was	Study 5
Immediacy/engagement	The feeling of closeness with the robot	Study 3

completion time. A quicker response could also possibly lead to a more efficient teamwork. However, if a response was due to a prediction (e.g., in Study 4, a robot proactively prepares actions based on predicted user intentions), how quicker responses could improve the efficiency of teamwork is contingent on the accuracy of prediction. An incorrect prediction might cause an erroneous response, which later could require more time to correct its erroneous action, possibly resulted in inferior task performance. For example, in Study 4 (Figure **??**), the task efficiency was correlated with the prediction accuracy of user intent.

In addition to efficiency, another evaluative aspect to consider is the *correctness* of teamwork. For example, in the collaborative sorting task in Study 1, the robot wished the user not only to complete the task in a shorter time but also to sort the objects to the correct places. In the context of information delivery (e.g., Studies 2 & 3), the quality can be assessed by how well delivered information was received (e.g., retention of information). Joint action should be done both efficiently and correctly.

Subjective measures. Subjective measures were used to probe user experience, including *perceptions of the robot* and *interaction experience* (Table 6.2). Each scale summarized in Table 6.2 consisted of multiple questionnaire items, aiming to capture different facets of that scale. While subjective measures in this research were usually obtained using 7-points rating scales, they could also be acquired using open choice. For example, in Study 3, to understand how participants would describe the robot's behavior, I provided the participants with 20 adjective words, 10 positive and 10 negative words, and asked them to choose the adjectives from the list that would best characterize the robot's behavior (Figure 3.27). This open-choice method aimed to gather a broader understanding of how people felt about the designed robot behavior.

The list of scales presented in Table 6.2 is not meant to be an exhaustive list. Rather it represents different aspects of user experience that I believed to be important to consider in joint action. However, it remains an open problem to create a set of validated scales that would capture key psychological experiences in joint action (Hoffman, 2013).

Behavioral measures. Behavioral measures were employed to collect behavioral evidence regarding how people interacted with the robot during an interaction. To a certain extent, behavioral evidence is most ecologically valid and could faithfully reveal the *quality* of joint action as they emerge naturally from the interaction and unintentionally "leak" the users' internal states (Ekman and Friesen, 1969a; Mutlu et al., 2009b). For example, Study 1 (Figure 3.10) showed that when people could not rely on the robot's gaze cues to identify task-relevant objects, their gaze patterns were resembled to those in human communication when they were asked to identify objects in the absence of the speaker gaze cues (Altmann and Kamide, 2004; Fischer, 1998). This evidence suggested a potential parallel between human-robot and human-human communication.

Moreover, behavioral measures could also be used indirectly to assess the quality of joint action. For instance, the behavioral measure employed in Studies 2 & 3 was to evaluate how well the participants could retell the story previously heard from the robot. Their retelling provided a way to assess their retention of information,

which was different from using a set of pre-designed questions. Pre-designed questions may not faithfully reflect people's acquired knowledge, as the details they remembered might not be tested in the questions and vice versa.

Future research can validate human-robot joint action through neurocognitive measures, such as recorded electrical activity of the brain, probing whether or not naive people have similar neurocognitive experiences when they interact with another person and with a robot. Neurocognitive measures, similar to behavioral measures, hold promise in providing direct, truthful evaluation of joint action.

A composite quality of joint action

Due to the complex system of joint action, the objective, subjective, and behavioral measures do not necessarily correlate with one another positively. For example, Study 5 showed that there was a tradeoff between objective team performance and subjective user experience. In particular, when the robot sought to maximize the team performance by reducing idle time for both the user and itself, the user reported the feeling of stress and rated the robot to have less awareness of and less patience with him or her. This tradeoff reinforced that joint action is a complex construct that should not be assessed by only a limited set of qualities.

Additionally, joint action involves various types of collaboration. The objective measures used in this research and in prior work on human-robot collaboration (e.g., Hoffman (2013); Nikolaidis et al. (2013); Shah et al. (2011)) seemed to mainly capture the success of coordination (e.g., efficiency) in professional teamwork, such as workers in a factory. While these metrics might be applicable to the context involving professional teamwork, they might not be applicable to daily collaborative activities such as unloading a dish rack. For instance, when people unload dishes at home, they might not care how quick they can finish the task. People may dislike working under pressure in these settings. As a result, a proactive robot, as in Study 5, may just be too efficient for a casual joint action and causes negative user experiences.

The above discussion leads to a proposal for evaluation of a composite quality of joint action (Equation 6.1). The quality of joint action can be estimated using a

joint consideration of various objective (indexed by i), subjective (indexed by j), and behavioral (indexed by k) measures. Each measure may have different impact (e.g., weight) on the overall quality. ω_i, δ_j, and φ_k are weights for objective, subjective, and behavioral measures, respectively.

$$\begin{aligned}\text{Quality of joint action} = \sum_{i=1}^{m} \omega_i * Objective_i + \sum_{j=1}^{n} \delta_j * Subjective_j \\ + \sum_{k=1}^{o} \varphi_k * Behavior_k\end{aligned} \tag{6.1}$$

The next question is how to determine the weight for each measure. I discuss two directions here. First, we can follow a system-level evaluation paradigm (Foster et al., 2009; Peltason et al., 2012; Walker et al., 1997) to obtain how the different qualities are represented in human joint action. For example, we can recruit pairs of naive people work together in achieving a common goal. Objective, subjective, and behavioral qualities are obtained from their joint actions. Either a third-party evaluation (e.g., having a judge to evaluate the quality of the joint action) or first-person subjective evaluation of their joint action can be used to obtain the "score" of the quality of joint action. A linear regression approach can then be applied to model how various measured qualities are contributing to the judged quality of the joint action.

Second, a multivariate evaluation, as proposed in Study 2, can be used to explore the effects of direct manipulations in the robot's behaviors on the quality of human-robot joint action. For example, one can manipulate the robot's behaviors in maximizing a particular measure (e.g., increasing the pace of the robot's actions to reduce idle time). By a systematic manipulation of various behaviors that aim for maximizing different measures, we can obtain insights into how various manipulated qualities shape the overall quality of human-robot joint action.

In brief, joint action is a complex construct and therefore requires a joint consideration of various objective, subjective, and behavioral measures in assessing its quality. More research is needed for benchmarking human-robot joint action.

6.1.3 Inevitable disfluencies in joint action

Human joint action, such as communication, involves disfluencies and people continuously repair them (Clark, 1994). Autonomous interactive systems, as those developed in this research, seem to inevitably suffer from disfluencies occurred in both the processes of *prediction* and *expression*. For example, in Study 4, the robot's proactive actions depended on its predictions of the user's intentions, which were shown to be incorrect from time to time. During a post-experiment interview, one participant noted that the robot went to the opposite side of what she was going to order and blamed the robot for the mistaken action.

> *"[The anticipatory robot] shouldn't move before I said what I wanted...so I guess that's [its] fault..."*

As another example, in Study 3, the robot used multimodal behaviors that were learned from the data of human performers to deliver information to participants. The robot's behaviors were driven by the learned model of human behavior, which was constrained in several aspects, including the used learning algorithm, the designed model structure, and the limited data of human performers. Therefore, the robot's behaviors were not necessarily as natural as those of human performers. A quote from an interview showed that a participant noted that the robot did not make much eye contact as it should.

> *"[The robot] seemed not to sometimes make eye contact when I expected it to. There were a few periods where it was talking and talking straight at the screen and not making eye contact."*

While people were sensitive and able to identify "abnormal" behaviors in the robot, they sometimes justified the robot's unnatural behavior. In a comment provided by a participant interacting with the robot showing random behaviors in Study 3, the participant rationalized how people might also behave in an unnatural manner.

"...sometimes maybe [the robot] used hand gestures in a way that I wouldn't necessary to use that ... [but] I feel like people do that sometimes too...."

Overall, the results in this research showed that although the developed system was not perfect people would consider the overall interaction as a whole when asked to evaluate the system and the interaction. However, one speculation is that the participants were paid to participate in the experiments and might feel obligated to provide positive evaluation, or even they might consider the systems as research prototypes and had lower expectations.

Nevertheless, future research is required to address disfluencies in joint action. I here briefly discuss two directions toward addressing this issue. One direction is to continuously improve the development of the autonomous system. More research is needed to enhance the performance of perception (e.g., computer vision and speech recognition) and to improve and develop algorithms for learning and prediction (e.g., machine learning). Advances in these fields will sure help enable effective joint action between humans and robots. Moreover, online learning algorithms can also be applied to continuously learn new interaction policies from experiences. However, no matter how advanced the new technologies and techniques might reach, there will always be errors and mistakes. Even humans are not perfect and make mistakes. Therefore, the other direction embraces these inevitable errors and seeks to mitigate or repair potential errors (Sauppé and Mutlu, 2014). It would require the robot to recognize errors, or deviations from social norms, and to incorporate such realization into its action plans to reduce any negative impact as resulted from the erroneous actions. Future research should explore how to incorporate expected action effects into motion planning and how to quantify social norms as a comparison standard.

6.1.4 Behavioral privacy

To support effective joint action with their partners, robots must monitor the partners' behavioral signals to understand and predict their actions (e.g., *action observation*). However, such monitoring might raise the concern of *behavioral privacy*, as

indicated by results from this research. Behavioral privacy in human-robot joint action concerns how people feel about the robot's ability to sense their behaviors and how the sensory information might be utilized by the robot. This concern was pointed out by a participant in Study 4.

> *"I could tell [the anticipatory robot] was watching my gaze or aware of my gaze... It has awareness... and that almost felt kind of freaky... that it almost could guess what I wanted although it was tracking my intention. I didn't like it as much."*

However, while being "watched" by a Kinect sensor in Study 5, no participants commented negatively on being monitored by the robot. One speculation is that body movements are overt in daily interaction and could be considered as "public" information. In contrast, gaze cues are much covert and are sometimes used in a secretive way, leading to a "personal ownership." Having access to and exploiting these subtle behavioral signals might mean violating behavioral privacy to some people. However, the sense of behavioral privacy varied among people as some participants in Study 4 attributed positive perceptions to the robot that utilized gaze cues to provide anticipatory assistance. Future investigations are needed to explore how robot systems can effectively utilize people's behavioral signals while keeping users comfortable during interactions.

6.2 Limitations

Limitations of this dissertation are discussed to provide directions for future research. Firstly, the research approach (Figure 1.1) employed in this dissertation utilizes a model of human-human joint action to inform the design of human-robot joint action. This approach provides a quantitative understanding of joint action and enables the development of operational models for generating human-inspired robot behaviors to facilitate human-robot joint action. However, it imposes limitations on how the modeled and realized behaviors might be generalized to other

contexts different from the one in which data of human joint action is collected. I here discuss two approaches that might help address this methodological limitation. The first approach involves using meta-models to organize collective knowledge of joint action. A meta-model aims to structure various studied mechanisms in a cognitively and computationally plausible manner. This approach requires new evidence on how different coordination mechanisms are related and work together as well as a computational architecture that can accommodate the cognitive organization of coordination mechanisms. The second approach extends the current research approach to study multiple joint-action scenarios with the goal to identify common design variables for effective joint action. This approach may allow for a more general model that could be potentially used in multiple applications. Ultimately, as collective knowledge of joint action becomes comprehensive, more general models would emerge to drive successful human-robot joint action.

In addition to the methodological limitation, the joint actions studied in this dissertation were modeled using instructed interactions in a laboratory setting. Therefore, the constructed models might not explain the interaction dynamics emerged from spontaneous joint action in a natural environment. For instance, how people perform the task of unloading a dish rack in a laboratory with a strange person might be different from how they would perform the task at home with someone they knew. Similarly, all my HRI evaluations were conducted in a laboratory setting. The way people interact with a robot in a laboratory might be different from the way they would interact with it in a familiarized environment. However, laboratory investigations serve as a systematical method to explore hypotheses and test prototypes. As more developed theories and systems become available, field studies are needed to validate laboratory findings and obtain new insights into joint action in natural interactions.

Additionally, HRI evaluations conducted in this research mostly involved one-time short-period interactions (e.g., usually less than 10 minutes). Such short exposure to the robot and the experimental manipulations might yield results different from those obtained through recurring long-term interactions. An emerging research direction is to deploy robots into natural settings to study long-term

human-robot interaction. To realize natural, effective human-robot joint action, future research is necessary to study long-term deployment of robots in natural human environments and explore how those robots might be integrated into human daily activities.

Moreover, the reported experimental results were subject to the quality and limitations of the sensors and robotic platforms used in the present studies. For instance, each of the Wakamaru robot's arms only has 4 DoF limiting the validity and variety of gestures that it could produce. This limitation in gestural movement might influence how people perceived the performance of the narrative robot (Study 2 & 3). Similarly, the inherent imperfection in eye tracking could negatively impact the accuracy of intention prediction as well as the production of anticipatory actions (Study 4). Furthermore, as discussed in Study 4, systematic errors are likely to be accumulated and multiplied throughout the process of sensing, decision making, and acting. Interpretations of the reported results should, therefore, consider inherent limitations in the research platforms and potentially accumulated errors.

Lastly, contextual factors, such as cultural background, personal disposition, and gender, were not explicitly taken into account in modeling human joint action and in HRI evaluations. Prior research has shown how these contextual factors might influence ways in which people perceive and express behaviors in social interactions. For example, the literature in human communication has suggested cultural differences in shaping how people use and perceive behavior, such as proxemic cues (Argyle, 2013; Evans and Howard, 1973; Hall, 1966) and backchannel signals (Maynard, 1986). Personal disposition (e.g., extroverted or introverted) was linked to the amount of mutual gaze people have with their interaction partners (Rutter et al., 1972). Gender has also be found to influence how social gaze cues are perceived and expressed in interactions (Argyle and Cook, 1976). Similarly, these contextual factors have also been observed to shape human-robot interaction (Bartneck et al., 2007; Mutlu et al., 2006; Walters et al., 2005). Future research is needed to explore how these factors might modulate the dynamics of joint action and how robots might take these factors into account when engaging human users in joint actions.

6.3 Future research

In addition to the future work motivated by the above limitations, this dissertation points toward directions for future research on human-robot joint action, or in general human-robot interaction.

The puzzle of joint action

Joint action is a complex system enabled by various supporting mechanisms. In this dissertation, I focused on the coordination mechanisms of *joint attention*, *action observation*, *task-sharing*, and *action coordination* as proposed by Sebanz et al. (2006). While these mechanisms provide a basis to understand and realize joint action, they are by no means exhaustive. For example, problems occurred in joint action need to be resolved jointly. Interaction partners employ joint strategies that prevent, warn, and repair problems as they interact (Clark, 1994). How are these strategies related to and different from the coordination mechanisms studied in this dissertation? Future research is necessary to study how robots might repair and manage problems that emerge from joint action. Additionally, future research should, for example, also consider how robots might perceive socially relevant information (Allison et al., 2000), infer human partners' mental states (Frith and Frith, 1999), and regular emotion (Barsade and Gibson, 2007) while interacting with humans in joint action. Moreover, each studied mechanism in this research is richly composite. The present studies only focused on some aspects of the mechanisms. For instance, I only focused on the temporal aspect in *planned action coordination* (Study 5), while assuming users would adapt to the robot spatially and not considering any emergent coordination (Knoblich et al., 2011). Future research is needed to further explore the complex process of joint action emerged from human-robot interaction.

While findings from psychology, cognitive science, and neuroscience have provided insights into the understanding of joint action, most of the results were derived from studies in which participants performed tasks that did not involve direct physical joint action. For instance, in investigating how people coordinate

with each other in a joint task, Knoblich and Jordan (2003) utilized a task in which two participants worked together to keep a circular tracker on top of a horizontally moving dot on a computer screen by pressing two separate key controls. One participant had the control of acceleration while the other participant had the control of braking. While this type of investigations shed light on the process of joint action, it lacks an embodied account of physical joint action.

Hypothetically, direct physical joint action might involve different coordination mechanisms as it inherently involves a greater use of embodied movements and their interactions with the physical world. Further understanding of embodied physical joint action in human interaction is necessary to uncover how embodiment shapes coordination in joint action. Such understanding will not only expand current knowledge of joint action but also inform better designs of robot systems to take the full advantage of embodied characteristics to support humans in joint action. With the advances in sensory technologies (e.g., Kinect and eye-tracker), researchers are now able to study the dynamics of physical joint action from a behavioral perspective. However, obtaining such understanding from a cognitive and neuroscience perspective remains challenging.

Towards human-robot joint action in the wild

As discussed above, laboratory investigations allow for systematic, controlled tests on research hypotheses and system prototypes. However, to ultimately enable robots to interact with humans in daily joint actions, we need a better understanding of how people work together in their natural environments and how robots might be integrated into their joint action processes. To gain such understanding, field studies—both *field observation* of human joint action and *field deployment* of robots to join in human teams—are necessary. The research approach used in this dissertation (Figure 1.1) can be extended to include field studies to increase its applicability in enabling human-robot teams in daily situations (Figure 6.2). Such inclusion of field studies bridges the gap between controlled laboratories and real-world environments.

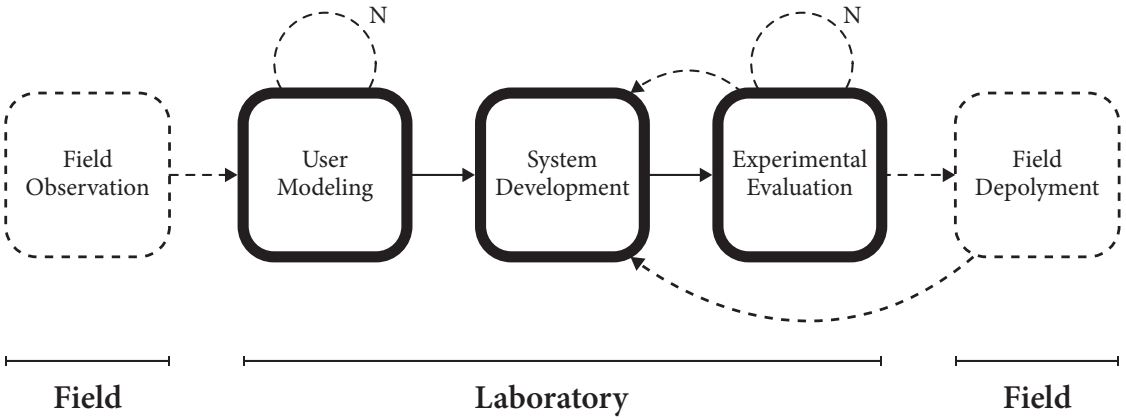


Figure 6.2: A research approach towards human-robot joint action in the real-world. This dissertation research has focused on the laboratory modeling, development, and evaluation (Solid line). The research approach used in this dissertation can be augmented with field observation and deployment, as well as modeling and evaluation with multiple application scenarios, to enrich the applicability of the developed systems. Future research is outlined by dashed lines.

An investigation can start with field observations to gain an understanding of how people perform joint actions in their environments. Field observations could identify key design variables and processes in joint action and inform laboratory user modeling studies. During the user modeling phase, those identified variables and processes are further studied to construct computational models that characterize the particular aspects of joint action in a controlled laboratory. The computational models are implemented into robotic systems, which can be first evaluated in a controlled laboratory and later deployed into human environments for field evaluation. Results from laboratory and field evaluations can inform further improvements for the robotic systems. Through iterative refinement, the robotic systems are enhanced to better support real-world human-robot joint action.

Although the above proposed research approach might serve as a start point for realizing human-robot joint action in the wild, integrating robots into human daily environments to support joint activities is very challenging. The real world presents numerous problems that challenge current robotic systems. For example, sensing, such as visual tracking and speech recognition, in the real-world environment is

non-trivial. There will be lighting changes and obscured objects that make visual sensing difficult. There will also be ambient noises or multiple people talking nearby that cause failures in speech recognition. In addition to sensing difficulties, joint actions in the real world are unstructured and often interfered with unexpected events. For instance, someone may join or leave the group anytime, causing changes in the group dynamics. Unexpected events might also cause the joint action to temporarily pause or fully stop. For example, the human partner might need to take a call or leave for other tasks. Moreover, a real-world joint action is usually embedded in a larger social interaction, where the start and end of the joint action are blended with other joint actions or processes. In addition to these technical challenges in handling uncertainties and variabilities, how to evaluate human-robot teams in the wild is difficult as well. More research is needed to address issues in real-world deployment and evaluation to advance toward enabling joint actions between humans and robots in everyday activities.

Multidisciplinary nature of human-robot joint action

Throughout the discussion and presentation in this dissertation, it should be clear that enabling natural, effective human-robot joint action is an interdisciplinary problem, requiring applications of various techniques, methods, and theories. This research therefore also motivates future work in related fields, such as social signal processing (e.g., Vinciarelli et al. (2009)), affective computing (e.g., Picard and Picard (1997)), and action recognition (e.g., Poppe (2010)). Advances in social signal processing will allow robots to better understand and utilize social signals displayed by human partners in joint action. Social signals carry rich information about a person's internal states. Accurate interpretations of the partner's social signals help increase the *bandwidth of awareness* in joint action. Similarly, a better interpretation and presentation of affect will facilitate the establishment of common ground in joint action as well as building rapport between the partners. Moreover, this research highlights the importance of action recognition and anticipation in aiding in robots' ability to work adaptively and collaboratively with human partners

in joint action. Development in these fields are enabled by fundamental advances in artificial intelligence, machine learning, computer vision, and more.

In addition to motivating research that allows better perceptions and understanding of behavioral signals expressed by humans, this research also encourages robot designs in both appearance and motion to support richer expression. Prior research has explored designing "eyes" for anthropomorphic robots (e.g., Diana and Thomaz (2011)) as well as for non-anthropomorphic robots (e.g., Hoffman and Weinberg (2010)) to support better non-verbal communication with human partners. Research has also focused on designing expressive motions for robots to leverage spatiotemporal affordances during interactions with humans (Hoffman and Ju, 2014). This current research provides additional empirical evidence to invite further research to enrich robots' capacity in expressing themselves to increase the *bandwidth of awareness* in joint action.

Beside the related fields mentioned above, the current research approach also inspires future work on automatic data annotation and processing. Right now, this research requires a certain amount of data annotation for obtaining behavioral parameters or training models of joint action. The annotation process usually involves coders to go through the data, such as recorded videos and system logs, and label features of interest. This process is particularly laborious and limits how much data is available for developing an understanding of human joint action as well as training a learning algorithm. While crowdsourcing data annotation (Lasecki et al., 2014; Park et al., 2012) seemed to be a plausible alternative, a self-supervised process of data annotation can greatly improve the current research barrier and allows for online incremental learning (e.g., Bohus and Horvitz (2014)).

In brief, realizing human-robot joint action involves interdisciplinary efforts. Further understanding of human joint action in cognitive science and neuroscience, technological advancement in perceiving and reasoning about human actions, innovative designs for expressive robot appearance and motion, and improvement in research approaches will together push forward the frontiers of human-robot joint action.

Speak in "robot language" ?

The premise of this dissertation research is that robots using human-inspired behaviors and coordination mechanisms can improve their interactions with humans during joint action. This premise is based on people's tendency and ability to read, attribute, and display behaviors that they are all familiar with. Therefore, this research, along with a large body of prior work, has been motivated to design robots to "speak in human language"—interacting with humans using human interaction manners. While the results from this research as well as prior work have provided evidence to support such premise, I here discuss a different perspective to leverage people's ability in "speaking robot language."

Humans are adaptive. People are tuned to adapt to others during interactions. Substantial evidence has shown that people mimic their interaction partners' behaviors during interactions, including the use of syntactic structures in talking (Branigan et al., 2000), accents (Coupland et al., 1991), gestures (Chui, 2014), and body movement (Shockley et al., 2003). Such mimicry and behavioral adaptation create fluid interactions and increase liking between the partners (Chartrand and Bargh, 1999).

I also found that participants adapted to robots in this research. For example, in Study 4, there was one participant who did not follow the task instruction to request five different fruits one by one for the smoothie order. Instead, he requested his five fruit choices at once in a single utterance. However, when the robot only picked up the first fruit that he requested, he adjusted his behavior to place one request at a time. As another example in Study 5, the participants were not informed how the robot would present the plates and cups during handovers. In the first trial, some participants approached the robot to try to take the plate from the robot while it was moving toward the participants to hand off the object. After the first trial, participants knew how the robot would present the object and adapted their behaviors to wait for the robot to present objects in the subsequent trials. While these instances might be due to the robot not expressing itself clearly to help the participants to understand it, the participants were able to quickly adjust their mental models of the robot. These observations suggested that people tend to adapt

their behaviors to fill the gap between their mental model of the robot behavior and the observed robot behavior.

While designing robots to "speak in human language" promises intuitive human-robot interaction (e.g., reducing users' effort in learning new interaction styles for robots), future robotic systems can consider how users might be able to "speak in robot language" to overcome technological limitations that prevent the realization of perfect human-language-speaking robots.

7 CONCLUSION

This dissertation investigated how human-inspired coordination mechanisms might be realized for robotic systems to support natural, effective human-robot joint action. To this end, I drew on the mechanisms of *joint attention*, *action observation*, *task-sharing*, and *action coordination* proposed in research cognitive science in developing autonomous robot systems that engage people in joint tasks. My research approach followed a human-centered design process to leverage inspirations from human-human joint action to design human-robot joint action. Specifically, my research approach involved *modeling* human behaviors in joint action, *implementing* autonomous robotic systems that utilized models of human behaviors in producing robot behaviors, and *evaluating* the effectiveness of the robotic systems in interacting with human users in joint tasks. In a series of five user studies, I followed this approach to design human-inspired coordination mechanisms for robot systems to interact with human users in joint tasks. Overall, the results from the five studies showed that robots utilizing human-inspired coordination mechanisms when interacting with humans in joint action were able to elicit greater team performance and improved user experience.

Below, I summarize the system, methodological, and empirical contributions of this research. I then provide final remarks to conclude this dissertation.

7.1 System contributions

This dissertation made contributions to enabling autonomous interactive robots that aim to support human partners in joint action. First, it introduced the *repertoire of robot behaviors*, a framework that allows generating robot social behaviors from a repository of specifications of social acts. Chosen specifications are organized and consolidated to form coherent behaviors using guidelines informed by Activity Theory (Leontjev, 1978). I implemented the *Robot Behavior Toolkit*, a software realization of the repertoire of robot behaviors. I demonstrated that the Robot Behavior

Toolkit could generate effective gaze behaviors (Study 1) and multimodal behaviors involving gaze and gestures (Study 2 & 3) in supporting effective communication. The Toolkit was implemented in the framework of Robot Operating System (ROS) and released as an open-source package to contribute to a broader community.

In investigating how robots might provide anticipatory assistance to users (Study 4), I proposed a Support-Vector-Machine-based model that quantified predictive relationships between gaze patterns and user intentions and an algorithm for anticipatory motion planning and execution (Algorithm **??**). Additionally, I implemented a robotic system that utilized the predictive model and the algorithm. The system was shown to interact with naive users effectively in a joint task in real-time. The system served as a prototype to demonstrate that the plausibility of utilizing user gaze cues obtained from an eye-tracker to infer their intention in a joint task and that robots could leverage such predicted intentions to provide anticipatory assistance to increase team efficiency. The predictive model, anticipatory algorithm, and the prototype system provided practical insights to developing robotic systems aim to deliver anticipatory support for users.

In another investigation to study how robots might adapt to user actions in a joint activity (Study 5), I proposed an algorithm that utilized two adaptive strategies identified in human joint action to generate adaptive coordination behavior for a robot (Algorithm 3). I proposed a K-Nearest-Neighbors-based model to predict user actions based on human joint data obtained from a Kinect sensor. The algorithm and model were implemented into a robot system to produce robot actions that adapted to user actions. The robotic system was shown to interact with naive users effectively in a collaborative task in real time. It served as a prototype to demonstrate that robots could take user actions into account in planning its actions to produce adaptive teamwork. The adaptation algorithm, predictive model, and the prototype system provided practical insights to developing robotic systems aim to work with users adaptively.

In brief, in this dissertation research, I built several real-time autonomous interactive systems to interact with naive users. These system building efforts not only demonstrated concepts of effective human-robot joint action but also revealed prac-

tical knowledge on the possibilities and limitations of current technologies. I sought to use these experiences to continuously make progress toward the realization of robot partners to support everyday activities.

7.2 Methodological contributions

This dissertation made methodological contributions to modeling, generating, and evaluating social behaviors for robots. I introduced a method for modeling and generating multimodal social behaviors. In particular, the method utilizes a dynamic Bayesian network (DBN) to jointly model behaviors in multiple channels. This method provides an implicit representation of temporal relationships among multimodal behaviors. In Study 3, I showed that such a DBN was able to model the temporal dynamics among speech, gaze, and gestures and further to generate those behaviors for a robot. In a human-robot interaction experiment, participants rated the robot using the behaviors generated by a proposed DBN to be more engaging than the robot that did not show gaze and gestures and the robot that showed random behaviors. Moreover, participants rated the robots using the behaviors generated by the proposed DBN and by a conventional, inspection-based approach to have similar qualities. While not explicitly demonstrated, this method holds promise in modeling complex temporal relationships among a large number of behaviors, which would be almost impossible using a conventional, inspection-based method.

The *repertoire of robot behaviors* was also an innovation in generating robot social behaviors (Study 1). Research in human communication and psychology has provided many empirical findings on how people use social behaviors in interaction. The repertoire of robot behaviors allows researchers and robot behavior designers to utilize those findings in a systematic way to produce coherent behaviors to achieve intended interaction goals. Moreover, it supports a community-based repository of social acts that would continue to expand as new findings on human behavior are revealed.

I proposed a *multivariate evaluation* approach to assessing how various behavioral variables (e.g., different types of gestures) that are systematically manipulated in a robot might shape outcomes of human-robot interaction. The multivariate evaluation approach is based on multiple linear regression to uncover the relationships between the behavioral manipulations and measured outcomes. Different from the system-level evaluation approach (e.g., Foster et al. (2009); Peltason et al. (2012)), the proposed approach allows direct manipulations in robot behaviors. When considering several behavioral variables at the same time (e.g., designing multimodal behaviors), such direct and systematic manipulations enables the possibility of in-person human-robot interaction evaluation, which would be much challenging using a common categorical evaluation approach. The proposed approach offers a systematic exploration of complex relationships among behavioral variables and their interactions with measured outcomes.

Together, these innovations in modeling, generating, and evaluating social behaviors serve as research tools for designing social behaviors for robots.

7.3 Empirical contributions

This dissertation provided empirical evidence and findings about human-human and human-robot interactions, showing how the coordination mechanisms of joint action might be computationally understood and realized. Such new knowledge of interaction would inform further research directions and development of robot systems.

Joint attention — The robot using referential gaze cues that were properly coupled with speech in time (Griffin, 2001; Meyer et al., 1998) effectively directed the participants attention to task-relevant artifacts in joint action. Such achievement of joint attention helped reduce the time needed by the participants to locate task objects in a shared workspace and enhanced the participants' retention of information delivered by the robot (Study 1). Moreover, deictic pointing gestures were also demonstrated to be particularly effective in manipulating people's attention. I found that participants performed better in recalling information delivered by the

robot using more deictic gestures during its presentation (Study 2). These findings were in line with prior research suggesting that people naturally use multimodal behaviors, including referential gaze cues, pointing gestures, and speech references, in aligning common ground with others to achieve effective communication (Bangerter, 2004; Baron-Cohen, 1997; Kaplan and Hafner, 2006). Therefore, I further studied how gaze, gestures, and speech might integrated together to produce coherent communicative behaviors (Study 3). In line with findings from Study 1 & 2, I found that a robot using coherent multimodal behaviors during its presentation of information was perceived to be more engaging compared to the robot that did not employ gaze and speech and the robot showed random behaviors. I also found that the participants performed better in retelling the story presented by the robot when the robot used coherent multimodal behaviors as opposed to random behaviors. Together, these findings suggested the effectiveness of gaze cues and gestures in manipulating people's attentional focus and how such attention manipulation can facilitate joint action between humans and robots.

In studying the role of gestures in effective communication (Studies 2 & 3), I developed an empirical model of human use of gestures. The model specified how deictic, iconic, and metaphoric gestures are temporally and linguistically linked to their corresponding speech references. In addition, the model described how gaze cues are distributed across various targets when performing different types of gestures. This model provided a quantitative understanding of how people employ speech, gaze, and gestures in communication.

Action observation — In my investigation into how a robot might monitor a user's behavior to predict his or her task intent (Study 4), I applied a machine learning approach to quantify the predictive relationship between gaze patterns and intentions. In an offline analysis, I showed that gaze patterns, particularly during and frequency of gaze cues, were effective in predicting a user's intentions (76% accuracy). Moreover, the predictive model could make correct predictions approximately 1.8 seconds before an explicit realization of user intention. Additional qualitative analyses revealed sequential patterns that might bear semantic meanings and signify user intention.

In a laboratory experiment, the predictive model was implemented in a robot system to provide anticipatory assistance to users. The results from the experiment showed that the predictive model could achieve comparable efficacy when considering implementation limitations of online interactions (e.g., projection errors). The results also showed that the robot could leverage predictions of user intentions to respond to the user quicker and to complete the task in a shorter time (2.5-second advantage). Furthermore, the participants attributed the robot's anticipatory actions to its awareness of themselves.

Task-sharing & action coordination — In addition to monitoring behavioral signals, I explored how a robot might explicitly represent its partner's task and integrate the understanding of current partner action and task progress into planning its actions to achieve adaptive coordination (Study 5). I identified two adaptive strategies—slowing-down and waiting—that people used in coordinating actions with others. I developed another machine learning model to predict a user's current task state based on the sensed user motion. The results from a laboratory human-robot interaction experiment showed that a robotic system, utilizing the adaptive strategies and the predictive model, was able to produce adaptive actions in response to users' task progress. The results also highlighted that the importance of a joint consideration of team performance and user experience in evaluating human-robot joint action.

7.4 Final remarks

In this dissertation, I present a series of five studies to demonstrate how the coordination mechanisms of *joint attention*, *action observation*, *task-sharing*, and *action coordination* can be realized for robots to effectively engage humans in joint action. The results from the studies provide evidence showing the *bandwidth of awareness* between a human and a robot during joint action can be increased by the processes of *expression* and *prediction*. In particular, Studies 1–3 showed that robots can utilize expressive behaviors, such as gaze and gestures, to help human partners to understand the robots' internal states. Studies 4 & 5 showed that robots can exploit

human partners' expressive behaviors to predict their intentions and actions in achieving anticipatory and adaptive joint action. This dissertation motivates future research on and makes system, methodological, and empirical contributions to enabling natural, effective joint actions between humans and robots.

REFERENCES

Admoni, Henny, Anca Dragan, Siddhartha S Srinivasa, and Brian Scassellati. 2014. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction (HRI)*, 49–56. ACM.

Admoni, Henny, and Brian Scassellati. 2014. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *Proc. ICMI'14*.

Aglioti, Salvatore M, Paola Cesari, Michela Romani, and Cosimo Urgesi. 2008. Action anticipation and motor resonance in elite basketball players. *Nature neuroscience* 11(9):1109–1116.

Alami, Rachid, Aurélie Clodic, Vincent Montreuil, Emrah Akin Sisbot, and Raja Chatila. 2006. Toward human-aware robot task planning. In *AAAI spring symposium: To boldly go where no human-robot team has gone before*, 39–46.

Aleotti, Jacopo, Vincenzo Micelli, and Stefano Caselli. 2012. Comfortable robot to human object hand-over. In *Proc. RO-MAN'12*.

———. 2014. An affordance sensitive system for robot to human object handover. *International Journal of Social Robotics* 6:653–666.

Alibali, M.W., and M.J. Nathan. 2007. *Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson*, 349–365. Cambridge U Press.

Allison, Truett, Aina Puce, and Gregory McCarthy. 2000. Social perception from visual cues: role of the sts region. *Trends in cognitive sciences* 4(7):267–278.

Altmann, G., and Y. Kamide. 2004. Now you see it, now you don't: Mediating the mapping between language and the visual world. In *The interface of language, vision, and action: Eye movements and the visual world*, 347–386. Psychology Press.

Andrist, Sean, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction (HRI)*, 25–32. ACM.

Argyle, M., and M. Cook. 1976. *Gaze and mutual gaze*. Cambridge U Press.

Argyle, Michael. 2013. *Bodily communication*. Routledge.

Astington, Janet W. 1993. *The child's discovery of the mind*, vol. 31. Harvard University Press.

Baldwin, Dare A. 1993. Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental psychology* 29(5):832.

Bangerter, Adrian. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science* 15(6):415–419.

Baron-Cohen, Simon. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry* 42(2):241–251.

Barresi, John, and Chris Moore. 1996. Intentional relations and social understanding. *Behavioral and brain sciences* 19(01):107–122.

Barsade, Sigal G, and Donald E Gibson. 2007. Why does affect matter in organizations? *The Academy of Management Perspectives* 21(1):36–59.

Bartneck, Christoph, Tomohiro Suzuki, Takayuki Kanda, and Tatsuya Nomura. 2007. The influence of people's culture and prior experiences with aibo on their attitude towards robots. *Ai & Society* 21(1-2):217–230.

Basili, Patrizia, Markus Huber, Thomas Brandt, Sandra Hirche, and Stefan Glasauer. 2009. Investigating human-human approach and hand-over. In *Human centered robot systems*, 151–160.

Bauer, Andrea, Dirk Wollherr, and Martin Buss. 2008. Human–robot collaboration: a survey. *International Journal of Humanoid Robotics* 5(01):47–66.

Becchio, Cristina, Luisa Sartori, and Umberto Castiello. 2010. Toward you the social side of actions. *Current Directions in Psychological Science* 19(3):183–188.

Beer, Jenay M, Cory-Ann Smarr, Tiffany L Chen, Akanksha Prakash, Tracy L Mitzner, Charles C Kemp, and Wendy A Rogers. 2012. The domesticated robot: design guidelines for assisting older adults to age in place. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction*, 335–342. ACM.

Bekkering, Harold, Ellen RA De Bruijn, Raymond H Cuijpers, Roger Newman-Norlund, Hein T Van Schie, and Ruud Meulenbroek. 2009. Joint action: Neurocognitive mechanisms supporting human interaction. *Topics in Cognitive Science* 1(2): 340–352.

Berlin, Matt, Jesse Gray, Andrea Lockerd Thomaz, and Cynthia Breazeal. 2006. Perspective taking: An organizing principle for learning in human-robot interaction. In *Proceedings of the national conference on artificial intelligence*, vol. 21, 1444. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Blakemore, Sarah-Jayne, and Jean Decety. 2001. From the perception of action to the understanding of intention. *Nature Reviews Neuroscience* 2(8):561–567.

Bohus, Dan, and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, 2–9. ACM.

Branigan, Holly P, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition* 75(2):B13–B25.

Breazeal, Cynthia, Andrew Brooks, Jesse Gray, Guy Hoffman, Cory Kidd, Hans Lee, Jeff Lieberman, Andrea Lockerd, and David Mulanda. 2004. Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots* 1(2):1–34.

Breazeal, Cynthia, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent robots and systems, 2005.(IROS 2005). 2005 IEEE/RSJ international conference on*, 708–713. IEEE.

Breazeal, Cynthia, and Brian Scassellati. 1999. How to build robots that make friends and influence people. In *Intelligent robots and systems, 1999. IROS'99. proceedings. 1999 IEEE/RSJ international conference on*, vol. 2, 858–863. IEEE.

Bremner, P., A. Pipe, C. Melhuish, M. Fraser, and S. Subramanian. 2009. Conversational gestures in human-robot interaction. In *Proc. SMC'11*, 1645–1649.

———. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *Proc. HUMANOIDS'11*, 458–465.

Brennan, Susan E, Xin Chen, Christopher A Dickinson, Mark B Neider, and Gregory J Zelinsky. 2008. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106(3):1465–1477.

Brethes, L., P. Menezes, F. Lerasle, and J. Hayet. 2004. Face tracking and hand gesture recognition for human-robot interaction. In *Proc. ICRA'04*, 1901–1906.

Brscić, Drazen, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. 2015. Escaping from children's abuse of social robots. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, 59–66. ACM.

Bruce, Allison, Jonathan Knight, Samuel Listopad, Brian Magerko, and Illah R Nourbakhsh. 2000. Robot improv: Using drama to create believable agents. In *Robotics and automation, 2000. proceedings. ICRA'00. IEEE international conference on*, vol. 4, 4002–4008. IEEE.

Buccino, Giovanni, Ferdinand Binkofski, Gereon R Fink, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, Rüdiger J Seitz, Karl Zilles, Giacomo Rizzolatti, and H-J Freund. 2001. Action observation activates premotor and parietal areas in a somatotopic manner: an fmri study. *European journal of neuroscience* 13(2):400–404.

Burgard, W., A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. 1999. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence* 114(1):3–55.

Butler, Samantha C, Albert J Caron, and Rechele Brooks. 2000. Infant understanding of the referential nature of looking. *Journal of Cognition and Development* 1(4): 359–377.

Butterworth, George. 1991. The ontogeny and phylogeny of joint visual attention.

Cakmak, Maya, Siddhartha S Srinivasa, Min Kyung Lee, Jodi Forlizzi, and Sara Kiesler. 2011a. Human preferences for robot-human hand-over configurations. In *Proc. IROS'11*.

Cakmak, Maya, Siddhartha S Srinivasa, Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011b. Using spatial and temporal contrast for fluent robot-human handovers. In *Proceedings of the 6th international conference on human-robot interaction (HRI)*, 489–496. ACM.

Chan, Wesley P, Yohei Kakiuchi, Kei Okada, and Masayuki Inaba. 2014. Determining proper grasp configurations for handovers through observation of object movement patterns and inter-object interactions during usage. In *Proc. IROS'14*.

Chan, Wesley P, Chris AC Parker, HF Van der Loos, and Elizabeth A Croft. 2012. Grip forces and load forces in handovers: implications for designing human-robot handover controllers. In *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction (HRI)*, 9–16. ACM.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chartrand, Tanya L, and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology* 76(6): 893.

Chen, Tiffany L, and Charles C Kemp. 2010. Lead me by the hand: Evaluation of a direct physical interface for nursing assistant robots. In *Proceedings of the 5th acm/ieee international conference on human-robot interaction (HRI)*, 367–374. IEEE Press.

Chen, Yi-Wei, and Chih-Jen Lin. 2006. Combining svms with various feature selection strategies. In *Feature extraction*, 315–324. Springer.

Choi, Young Sang, Travis Deyle, Tiffany Chen, Jonathan D Glass, and Charles C Kemp. 2009. A list of household objects for robotic retrieval prioritized by people with als. In *Rehabilitation robotics, 2009. ICORR 2009. IEEE international conference on*, 510–517. IEEE.

Christensen, Henrik, T Batzinger, K Bekris, K Bohringer, J Bordogna, G Bradski, O Brock, J Burnstein, T Fuhlbrigge, R Eastman, et al. 2009. A roadmap for us robotics: From internet to robotics. *Computing Community Consortium and Computing Research Assoc*.

Chui, Kawai. 2014. Mimicked gestures and the joint construction of meaning in conversation. *Journal of Pragmatics* 70:68–85.

Clark, Herbert H. 1994. Managing problems in speaking. *Speech communication* 15(3):243–250.

———. 1996. *Using language*, vol. 4. Cambridge University Press.

———. 2005. Coordinating with each other in a material world. *Discourse studies* 7(4-5):507–525.

Clark, Herbert H, and Susan E Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13(1991):127–149.

Clark, Herbert H, and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50(1):62–81.

Cohen, Philip R, and Hector J Levesque. 1990. Intention is choice with commitment. *Artificial intelligence* 42(2):213–261.

Cortes, Corinna, and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Coupland, Justine, Nikolas Coupland, and Howard Giles. 1991. Accommodation theory. communication, context and consequences. *Contexts of Accommodation. Cambridge & Paris: Cambridge University Press & Éditions de la maison des sciences de lâŁ™homme* 1–68.

Csikszentmihalyi, Mihaly. 1999. If we are so rich, why aren't we happy? *American psychologist* 54(10):821.

Daprati, Elena, Nicolas Franck, Nicolas Georgieff, Joëlle Proust, Elisabeth Pacherie, Jean Dalery, and Marc Jeannerod. 1997. Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition* 65(1):71–86.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society* 1–38.

Diana, Carla, and Andrea L Thomaz. 2011. The shape of simon: creative design of a humanoid robot shell. In *Chi'11 extended abstracts on human factors in computing systems*, 283–298. ACM.

Doshi, Anup, and Mohan M Trivedi. 2009. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *Intelligent Transportation Systems, IEEE Transactions on* 10(3):453–462.

Dragan, Anca, and Siddhartha Srinivasa. 2013. Generating legible motion.

Dragan, Anca D, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction (HRI)*, 51–58. ACM.

Duffy, B.R., M. Dragone, and G.M.P. O'Hare. 2005. Social robot architecture: A framework for explicit social interaction. In *Android science: Towards social mechanisms, cogsci 2005 workshop*.

Duncan, Starkey. 1972. Some signals and rules for taking speaking turns in conversations. *J. Personality and Social Psychology* 23(2):283–292.

Edsinger, Aaron, and Charles C Kemp. 2007. Human-robot interaction for cooperative manipulation: Handing objects to one another. In *Proc. RO-MAN'07*.

Ekman, Paul, and Wallace V Friesen. 1969a. Nonverbal leakage and clues to deception. *Psychiatry* 32(1):88–106.

———. 1969b. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture* 57–106.

Evans, Gary W, and Roger B Howard. 1973. Personal space. *Psychological bulletin* 80(4):334.

Farrer, Chlöé, and Chris D Frith. 2002. Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *Neuroimage* 15(3):596–603.

Fischer, Burkhart. 1998. Attention in saccades. In *Visual attention*, 289–305. New York: Oxford University Press.

Flanagan, J Randall, and Roland S Johansson. 2003. Action plans used in action observation. *Nature* 424(6950):769–771.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42(3):143–166.

Foster, Mary Ellen, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp: Volume 2-volume 2*, 879–887. Association for Computational Linguistics.

Frischen, Alexandra, Andrew P Bayliss, and Steven P Tipper. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin* 133(4):694.

Frith, Chris D, and Uta Frith. 1999. Interacting minds–a biological basis. *Science* 286(5445):1692–1695.

———. 2006. How we predict what other people are going to do. *Brain research* 1079(1):36–46.

Frith, Christopher D, Daniel M Wolpert, et al. 2000. Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society B: Biological Sciences* 355(1404):1771–1788.

Gallese, Vittorio, and Alvin Goldman. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences* 2(12):493–501.

Garrod, Simon, and Martin J Pickering. 2004. Why is conversation so easy? *Trends in cognitive sciences* 8(1):8–11.

Georgieff, Nicolas, and Marc Jeannerod. 1998. Beyond consciousness of external reality: a "who" system for consciousness of action and self-consciousness. *Consciousness and cognition* 7(3):465–477.

Goldin-Meadow, S. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3(11):419–429.

———. 2003. *Hearing gesture: How our hands help up think*. Harvard U Press.

Goodwin, C. 1981. *Conversational organization: Interaction between speakers and hearers*. Academic Press New York.

Graham, J.A. 1975. A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology* 10:57–67.

Griffin, Zenzi M. 2001. Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82(1):B1–B14.

Grigore, Elena Corina, Kerstin Eder, Anthony G Pipe, Chris Melhuish, and Ute Leonards. 2013. Joint action understanding improves robot-to-human object handover. In *Proc. IROS'13*.

Haji-Abolhassani, Amin, and James J Clark. 2014. An inverse yarbus process: Predicting observers' task from eye movement patterns. *Vision research* 103:127–142.

Hall, Edward Twitchell. 1966. The hidden dimension.

Hawkins, Kelsey P, Shray Bansal, Nam N Vo, and Aaron F Bobick. 2014. Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In *Robotics and automation (ICRA), 2014 ieee international conference on*, 2215–2222. IEEE.

Hayashi, Kotaro, Masahiro Shiomi, Takayuki Kanda, and Norihiro Hagita. 2012. Friendly patrolling: A model of natural encounters. In *Proc. RSS*, 121.

Heal, Jane. 2005. Joint attention and understanding the mind. *Joint attention: Communication and other minds* 34–44.

Heath, Christian. 1992. Gesture's discreet tasks: Multiple relevancies in visual conduct and in the contextualisation of language. *The contextualization of language* 101:127.

Henderson, A., F. Goldman-Eisler, and A. Skarbek. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech* 9:207–216.

Hoffman, Guy. 2013. Evaluating fluency in human-robot collaboration. In *Proc. HRI'13 workshop on human robot collaboration*.

Hoffman, Guy, and Cynthia Breazeal. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proc. HRI'07*.

Hoffman, Guy, and Wendy Ju. 2014. Designing robots with movement in mind. *Journal of Human-Robot Interaction* 3(1):89–122.

Hoffman, Guy, Rony Kubat, and Cynthia Breazeal. 2008. A hybrid control system for puppeteering a live robotic stage actor. In *Robot and human interactive communication, 2008. ro-man 2008. the 17th ieee international symposium on*, 354–359. IEEE.

Hoffman, Guy, and Gil Weinberg. 2010. Shimon: an interactive improvisational robotic marimba player. In *CHI'10 extended abstracts on human factors in computing systems*, 3097–3102. ACM.

———. 2011. Interactive improvisation with a robotic marimba player. *Autonomous Robots* 31(2-3):133–153.

Hollan, J., E. Hutchins, and D. Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction* 7(2):174–196.

Holroyd, Aaron, Charles Rich, Candace L Sidner, and Brett Ponsler. 2011. Generating connection events for human-robot collaboration. In *RO-MAN, 2011 ieee*, 241–246. IEEE.

Huang, C.-M., and B. Mutlu. 2012. Robot behavior toolkit: Generating effective social behaviors for robots. In *Proc. HRI'12*, 25–32.

Huang, Chien-Ming, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015a. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6.

Huang, Chien-Ming, Maya Cakmak, and Bilge Mutlu. 2015b. Adaptive coordination strategies for human-robot handovers. In *Proceedings of the robotics: Science and systems conference, RSS'15*.

Huang, Chien-Ming, Takamasa Iio, Satoru Satake, and Takayuki Kanda. 2014. Modeling and controlling friendliness for an interactive museum robot. In *Proceedings of the robotics: Science and systems conference, RSS'14*.

Huang, Chien-Ming, and Bilge Mutlu. 2013a. Modeling and evaluating narrative gestures for humanlike robots. In *Proc. RSS'13*.

———. 2013b. The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *Journal of Human-Robot Interaction* 2(2): 80–102.

———. 2014a. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction (HRI)*, 57–64. ACM.

———. 2014b. Multivariate evaluation of interactive robot systems. *Autonomous Robots* 37(4):335–349.

Huang, Chien-Ming, and Andrea Lockerd Thomaz. 2011. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *RO-MAN, 2011 IEEE*, 65–71. IEEE.

Huber, Markus, Aleksandra Kupferberg, Claus Lenz, Alois Knoll, Thomas Brandt, and Stefan Glasauer. 2013. Spatiotemporal movement planning and rapid adaptation for manual interaction. *PloS one* 8(5):e64982.

Huber, Markus, Markus Rickert, Alois Knoll, Thomas Brandt, and Stefan Glasauer. 2008. Human-robot interaction in handing-over tasks. In *Proc. RO-MAN'08*.

Imai, Michita, Tetsuo Ono, and Hiroshi Ishiguro. 2003. Physical relation and expression: Joint attention for human-robot interaction. *Industrial Electronics, IEEE Transactions on* 50(4):636–643.

Jaffe, J., L. Cassotta, and S. Feldstein. 1964. Markovian model of time patterns of speech. *Science* 144:884–886.

Jeannerod, Marc. 1999. To act or not to act: perspectives on the representation of actions. *The Quarterly Journal of Experimental Psychology: Section A* 52(1):1–29.

Johansson, Roland S, Göran Westling, Anders Bäckström, and J Randall Flanagan. 2001. Eye–hand coordination in object manipulation. *the Journal of Neuroscience* 21(17):6917–6932.

Johnson, Matthew, and Yiannis Demiris. 2005. Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems* 2(4):301–308.

Julnes, George, and Lawrence B Mohr. 1989. Analysis of no-difference findings in evaluation research. *Evaluation Review* 13(6):628–655.

Jung, Malte F, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on computer supported cooperative work*, 1555–1566. ACM.

Kanda, T., M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita. 2009. An affective guide robot in a shopping mall. In *Proc. HRI'09*, 173–180.

Kanda, Takayuki, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction* 19(1):61–84.

Kanda, Takayuki, Masayuki Kamasima, Michita Imai, Tetsuo Ono, Daisuke Sakamoto, Hiroshi Ishiguro, and Yuichiro Anzai. 2007a. A humanoid robot that pretends to listen to route guidance from a human. *Autonomous Robots* 22(1): 87–100.

Kanda, Takayuki, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007b. A two-month field trial in an elementary school for long-term human–robot interaction. *Robotics, IEEE Transactions on* 23(5):962–971.

Kaplan, Frederic, and Verena V Hafner. 2006. The challenges of joint attention. *Interaction Studies* 7(2):135–169.

Kendon, A. 1978. Looking in conversation and the regulation of turns at talk: A comment on the papers of g. beattie and d. r. rutter et al. *British Journal of Social and Clinical Psychology* 17:23–24.

———. 1980. *Gesticulation and speech: Two aspects of the process of utterance*. Mouton De Gruyter.

———. 2004. *Gesture: Visible action as utterance*. Cambridge U Press.

Kendon, A., and A. Ferber. 1973. A description of some human greetings. *Comparative ecology and behavior of primates* 591–668.

Kendon, Adam. 1994. Do gestures communicate? a review. *Research on language and social interaction* 27(3):175–200.

Kidokoro, Hiroyuki, Takefumi Kanda, Drazen Brscic, and Masahiro Shiomi. 2013. Will i bother here?-a robot anticipating its influence on pedestrian walking comfort. In *Human-robot interaction (HRI), 2013 8th acm/ieee international conference on*, 259–266. IEEE.

Kilner, James M, Claudia Vargas, Sylvie Duval, Sarah-Jayne Blakemore, and Angela Sirigu. 2004. Motor activation prior to observation of a predicted movement. *Nature neuroscience* 7(12):1299–1301.

King, Chih-Hung, Tiffany L Chen, Advait Jain, and Charles C Kemp. 2010. Towards an assistive robot that autonomously performs bed baths for patient hygiene. In *Intelligent robots and systems (IROS), 2010 ieee/rsj international conference on*, 319–324. IEEE.

Knoblich, Günther, Stephen Butterfill, and Natalie Sebanz. 2011. Psychological research on joint action: theory and data. *Psychology of Learning and Motivation-Advances in Research and Theory* 54:59.

Knoblich, Günther, and Jerome Scott Jordan. 2003. Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(5):1006.

Koene, Ansgar, Anthony Remazeilles, Miguel Prada, Ainara Garzo, Mildred Puerto, Satoshi Endo, and Alan M Wing. 2014. Relative importance of spatial and temporal precision for user satisfaction in human-robot object handover interactions. In *Proc. new frontiers in human-robot interaction '14*.

Kollar, Daphne, and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT Press.

Kollar, Thomas, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *Human-robot interaction (HRI), 2010 5th acm/ieee international conference on*, 259–266. IEEE.

Koppula, Hema, and Ashutosh Saxena. 2013. Anticipating human activities using object affordances for reactive robotic response. In *In RSS*.

Kuffner, James J, and Steven M LaValle. 2000. Rrt-connect: An efficient approach to single-query path planning. In *Robotics and automation, 2000. proceedings. icra'00. ieee international conference on*, vol. 2, 995–1001. IEEE.

Kuhl, Patricia K, Jean E Andruski, Inna A Chistovich, Ludmilla A Chistovich, Elena V Kozhevnikova, Viktoria L Ryskina, Elvira I Stolyarova, Ulla Sundberg, and Francisco Lacerda. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science* 277(5326):684–686.

Kuutti, Kari. 1996. Activity theory as a potential framework for human-computer interaction research. *Context and consciousness: Activity theory and human-computer interaction* 17–44.

Land, Michael, Neil Mennie, Jenny Rusted, et al. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception-London* 28(11): 1311–1328.

Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

Lasecki, Walter S, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual acm symposium on user interface software and technology*, 551–562. ACM.

Lave, J. 1988. *Cognition in practice: Mind, mathematics, and culture in everyday life*. Cambridge University Press.

Lee, Jina, and Stacy Marsella. 2012. Modeling speaker behavior: a comparison of two approaches. In *Proc. IVA'12*, 161–174.

Lee, Min Kyung, Jodi Forlizzi, Sara Kiesler, Maya Cakmak, and Siddhartha Srinivasa. 2011. Predictability or adaptivity?: Designing robot handoffs modeled from trained dogs and people. In *Proc. HRI'11*.

Lee, M.K., S. Kiesler, and J. Forlizzi. 2010. Receptionist or information kiosk: how do people talk with a robot? In *Proc. HRI'10*.

Lenz, Claus, Suraj Nair, Markus Rickert, Alois Knoll, Wolgang Rosel, Jurgen Gast, Alexander Bannat, and Frank Wallhoff. 2008. Joint-action for humans and industrial robots for assembly tasks. In *Robot and human interactive communication, 2008. RO-MAN'08. the 17th ieee international symposium on*, 130–135. IEEE.

Leontjev, A.N. 1978. *Activity. consciousness. personality*. Englewood Cliffs, Prentice Hall.

Leslie, Alan M. 1987. Pretense and representation: The origins of "theory of mind.". *Psychological review* 94(4):412.

Libet, Benjamin, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl. 1983. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain* 106(3):623–642.

Lozano, S.C., and B. Tversky. 2006. Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language* 55(1):47–63.

Mainprice, Jim, and Dmitry Berenson. 2013. Human-robot collaborative manipulation planning using early prediction of human motion. In *Intelligent robots and systems (IROS), 2013 ieee/rsj international conference on*, 299–306. IEEE.

Mainprice, Jim, Mamoun Gharbi, Thierry Siméon, and Rachid Alami. 2012. Sharing effort in planning human-robot handover tasks. In *Proc. RO-MAN'12*.

Malle, Bertram F, and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33(2):101–121.

Marin-Urias, Luis Felipe, E Akin Sisbot, and Rachid Alami. 2008. Geometric tools for perspective taking for human–robot interaction. In *Artificial intelligence, 2008. micai'08. seventh mexican international conference on*, 243–249. IEEE.

Marsh, Kerry L, Michael J Richardson, Reuben M Baron, and RC Schmidt. 2006. Contrasting approaches to perceiving and acting with others. *Ecological Psychology* 18(1):1–38.

Marsh, Kerry L, Michael J Richardson, and RC Schmidt. 2009. Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science* 1(2):320–339.

Maynard, Senko K. 1986. On back-channel behavior in japanese and english casual conversation. *Linguistics* 24(6):1079–1108.

McCall, Joel C, David P Wipf, Mohan M Trivedi, and Bhaskar D Rao. 2007. Lane change intent analysis using robust operators and sparse bayesian learning. *Intelligent Transportation Systems, IEEE Transactions on* 8(3):431–440.

McNeill, David. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Mehrabian, Albert. 1971. *Silent messages*. Wadsworth.

Meltzoff, Andrew N. 1995. Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Developmental psychology* 31(5):838.

Meltzoff, Andrew N, and Rechele Brooks. 2001. Like me" as a building block for understanding other minds: Bodily acts, attention, and intention. *Intentions and intentionality: Foundations of social cognition* 171–191.

Meyer, Antje S, Astrid M Sleiderink, and Willem JM Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition* 66(2): B25–B33.

Montreuil, Vincent, Aurélie Clodic, Maxime Ransan, and Rachid Alami. 2007. Planning human centered robot activities. In *Systems, man and cybernetics, 2007. isic. ieee international conference on*, 2618–2623. IEEE.

Moon, AJung, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction (HRI)*, 334–341. ACM.

Morales, Michael, Peter Mundy, and Jennifer Rojas. 1998. Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development* 21(2):373–377.

Morency, Louis-Philippe. 2010. Modeling human communication dynamics. *IEEE Signal Processing Magazine* 27(5):112–116.

Morency, Louis-Philippe, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20(1):70–84.

Morris, Desmond, and G Desebrock. 1977. *Manwatching: A field guide to human behaviour*. HN Abrams New York.

Mundy, Peter, and Lisa Newell. 2007. Attention, joint attention, and social cognition. *Current Directions in Psychological Science* 16(5):269–274.

Murphy, Kevin, et al. 2001. The bayes net toolbox for matlab. *Computing science and statistics* 33(2):1024–1034.

Murphy, Kevin Patrick. 2002. Dynamic bayesian networks: representation, inference and learning. Ph.D. thesis, University of California.

Mutlu, Bilge, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th ieee-ras international conference on*, 518–523. IEEE.

Mutlu, Bilge, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1(2):12.

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

Mutlu, Bilge, Allison Terrell, and Chien-Ming Huang. 2013. Coordination mechanisms in human-robot collaboration. In *Proc. HRI'13 workshop on collaborative manipulation*.

Mutlu, Bilge, Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009b. Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. In *Proceedings of the 4th acm/ieee international conference on human robot interaction*, 69–76. ACM.

Narahara, H., and T. Maeno. 2007. Factors of gestures of robots for smooth communication with humans. In *Proc. robocomm'07*, 44:1–4.

Nardi, Bonnie A. 1996. Studying context: A comparison of activity theory, situated action models, and distributed cognition. *Context and consciousness: Activity theory and human-computer interaction* 69–102.

Ng-Thow-Hing, V., P. Luo, and S. Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *Proc. IROS'10*, 4617–4624.

Nickel, K., and R. Stiefelhagen. 2007. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing* 25(12):1875–1884.

Nikolaidis, Stefanos, Przemyslaw Lasota, Gregory Rossano, Carlos Martinez, Thomas Fuhlbrigge, and Julie Shah. 2013. Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action. In *Robotics (isr), 2013 44th international symposium on*, 1–6. IEEE.

Ognibene, Dimitri, Eris Chinellato, Miguel Sarabia, and Yiannis Demiris. 2013. Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspiration & biomimetics* 8(3):035002.

Ognibene, Dimitri, and Yiannis Demiris. 2013. Towards active event recognition. In *Proceedings of the twenty-third international joint conference on artificial intelligence*, 2495–2501. AAAI.

Okuno, Y., T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. 2009. Providing route directions: design of robot's utterance, gesture, and timing. In *Acm/ieee international conference on human-robot interaction (HRI)*, 53–60.

Otsuka, Kazuhiro, Hiroshi Sawada, and Junji Yamato. 2007. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances. In *Proc. ICMI'07*, 255–262.

Park, Sunghyun, Gelareh Mohammadi, Ron Artstein, and Louis-Philippe Morency. 2012. Crowdsourcing micro-level multimedia annotations: The challenges of

evaluation and interface. In *Proceedings of the acm multimedia 2012 workshop on crowdsourcing for multimedia*, 29–34. ACM.

Peltason, J., N. Riether, B. Wrede, and I. Lütkebohle. 2012. Talking with robots about objects: a system-level evaluation in hri. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction (HRI)*, 479–486.

Peltason, Julia, and Britta Wrede. 2010. Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In *Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue*, 229–232. Association for Computational Linguistics.

Pérez-D'Arpino, Claudia, and J Shah. 2015. Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. *Currently under review*.

Pezzulo, Giovanni, Francesco Donnarumma, and Haris Dindo. 2013. Human sensorimotor communication: a theory of signaling in online social interactions. *PloS one* 8(11):e79876.

Pezzulo, Giovanni, and Dimitri Ognibene. 2012. Proactive action preparation: Seeing action preparation as a continuous and proactive process. *Motor control* 16(3):386–424.

Picard, Rosalind W, and Roalind Picard. 1997. *Affective computing*, vol. 252. MIT press Cambridge.

Poppe, Ronald. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28(6):976–990.

Quigley, Morgan, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. Ros: an open-source robot operating system. In *Proc. ICRA'09 workshop on open source software*.

Ramnani, Narender, and R Christopher Miall. 2004. A system in the human brain for predicting the actions of others. *Nature neuroscience* 7(1):85–90.

Richardson, Daniel C., and Rick Dale. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science* 29(6):1045–1060.

Richmond, V. 2002. *Teacher nonverbal immediacy: Use and outcomes*, 65–82. Allyn and Bacon.

Riek, Laurel D, Tal-Chen Rabinowitch, Paul Bremner, Anthony G Pipe, Mike Fraser, and Peter Robinson. 2010. Cooperative gestures: Effective signaling for humanoid robots. In *Human-robot interaction (HRI), 2010 5th acm/ieee international conference on*, 61–68. IEEE.

Rizzolatti, Giacomo, and Laila Craighero. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.* 27:169–192.

Rizzolatti, Giacomo, Leonardo Fogassi, and Vittorio Gallese. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* 2(9):661–670.

Roth, W.-M. 2001. Gestures in teaching and learning. *Review of Educational Research* 71(3):365–392.

Rutter, DR, Ian E Morley, and Jane C Graham. 1972. Visual interaction in a group of introverts and extraverts. *European Journal of Social Psychology* 2(4):371–384.

Sakita, Kenji, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura, and Katsushi Ikeuchi. 2004. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *Intelligent robots and systems, 2004.(IROS). proceedings. 2004 IEEE/RSJ international conference on*, vol. 1, 846–851. IEEE.

Salem, Maha, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4(2):201–217.

Sauppé, Allison, and Bilge Mutlu. 2014. Effective task training strategies for instructional robots. In *Proceedings of the 10th annual robotics: Science and systems conference*.

Scassellati, Brian. 1999. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for metaphors, analogy, and agents*, 176–195. Springer.

Schegloff, E. 1984. *On some gestures' relation to speech*, 266–296. Cambridge U Press.

van Schie, Hein T, Rogier B Mars, Michael GH Coles, and Harold Bekkering. 2004. Modulation of activity in medial frontal and motor cortices during error observation. *Nature neuroscience* 7(5):549–554.

Sebanz, Natalie, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences* 10(2):70–76.

Sebanz, Natalie, and Chris Frith. 2004. Beyond simulation? neural mechanisms for predicting the actions of others. *Nature neuroscience* 7(1):5–6.

Sebanz, Natalie, and Guenther Knoblich. 2009. Prediction in joint action: what, when, and where. *Topics in Cognitive Science* 1(2):353–367.

Sebanz, Natalie, Günther Knoblich, and Wolfgang Prinz. 2003. Representing others' actions: just like one's own? *Cognition* 88(3):B11–B21.

———. 2005. How two share a task: corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance* 31(6):1234.

Shah, Julie, and Cynthia Breazeal. 2010. An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Human Factors* 52(2):234–245.

Shah, Julie, James Wiken, Brian Williams, and Cynthia Breazeal. 2011. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proc. HRI'11*.

Shi, Chao, Masahiro Shiomi, Christian Smith, Takayuki Kanda, and Hiroshi Ishiguro. 2013. A model of distributional handing interaction for a mobile robot. In *Robotics: Science and systems (RSS)*.

Shibata, Satoru, Benlamine Mohamed Sahbi, Kanya Tanaka, and Akira Shimizu. 1997. An analysis of the process of handing over an object and its application to robot motions. In *Proc. SMC'97*.

Shibata, Satoru, Kanya Tanaka, and Akira Shimizu. 1995. Experimental analysis of handing over. In *Proc. RO-MAN'95*.

Shiomi, Masahiro, Takayuki Kanda, Dylan F Glas, Satoru Satake, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Field trial of networked social robots in a shopping mall. In *Intelligent robots and systems, 2009. IROS 2009. IEEE/RSJ international conference on*, 2846–2853. IEEE.

Shiomi, Masahiro, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2006. Interactive humanoid robots for a science museum. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on human-robot interaction (HRI)*, 305–312. ACM.

Shockley, Kevin, Marie-Vee Santana, and Carol A Fowler. 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* 29(2):326.

Shrout, Patrick E, and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86(2):420.

Sidnell, Jack. 2006. Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction* 39(4):377–409.

Sidner, Candace L, Cory D Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on intelligent user interfaces*, 78–84. ACM.

Siepmann, Frederic, and Sven Wachsmuth. 2011. A modeling framework for reusable social behavior. In *De silva, r., reidsma, d., eds.: Work in progress workshop proceedings icsr*, 93–96.

Sisbot, Emrah Akin, Luis F Marin-Urias, Rachid Alami, and Thierry Simeon. 2007. A human aware mobile robot motion planner. *Robotics, IEEE Transactions on* 23(5): 874–883.

Sivakumar, Prasanna Kumar, Chittaranjan S Srinivas, Andrey Kiselev, and Amy Loutfi. 2013. Robot-human hand-overs in non-anthropomorphic robots. In *Proc. HRI'13*.

Staudte, Maria, and Matthew W. Crocker. 2009. Visual attention in spoken human-robot interaction. In *Proc HRI'09*, 77–84.

Steinfeld, Aaron, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction*, 33–40. ACM.

Strabala, Kyle, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S Srinivasa. 2012. Learning the communication of intent prior to physical collaboration. In *Proc. RO-MAN'12*.

Streeck, J. 1988. The significance of gestures: How it is established. *Papers in Pragmatics* 2(1/2):60–83.

———. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communications Monographs* 60(4):275–299.

Streeck, Jürgen. 2010. The significance of gesture: How it is established. *Papers in pragmatics* 2(1).

Stroop, J Ridley. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology* 18(6):643.

Sugiyama, O., T. Kanda, M. Imai, H. Ishiguro, and N. Hagita. 2007a. Natural deictic communication with humanoid robots. In *Proc. IROS'07*.

Sugiyama, Osamu, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2007b. Natural deictic communication with humanoid robots. In *Intelligent robots and systems, 2007. IROS 2007. IEEE/RSJ international conference on*, 1441–1448. IEEE.

Sung, JaYoung, Henrik Christensen, Rebecca E Grinter, et al. 2009. Sketching the future: Assessing user needs for domestic robots. In *Robot and human interactive communication, 2009. RO-MAN 2009. the 18th IEEE international symposium on*, 153–158. IEEE.

Takayama, Leila, Doug Dooley, and Wendy Ju. 2011. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on human-robot interaction*, 69–76. ACM.

Tellex, Stefanie, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*.

Thomas, Frank, Ollie Johnston, and Frank. Thomas. 1995. *The illusion of life: Disney animation*. Hyperion New York.

Thomaz, Andrea Lockerd, Matt Berlin, and Cynthia Breazeal. 2005. An embodied computational model of social referencing. In *Robot and human interactive communication, 2005. ROMAN 2005. IEEE international workshop on*, 591–598. IEEE.

Thompson, L.A., D. Driscoll, and L. Markson. 1998. Memory for visual-spoken language in children and adults. *Journal of Nonverbal Behavior* 22:167–187.

Tomasello, Michael. 1995. Joint attention as social cognition. *Joint attention: Its origins and role in development* 103–130.

———. 2009. *Why we cooperate*. MIT press.

Trafton, J Gregory, Nicholas L Cassimatis, Magdalena D Bugajska, Derek P Brock, Farilee E Mintz, and Alan C Schultz. 2005. Enabling effective human-robot interaction using perspective-taking in robots. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 35(4):460–470.

Trautman, Peter. 2013. Robot navigation in dense crowds: Statistical models and experimental studies of human robot cooperation. Ph.D. thesis, California Institute of Technology.

van Ulzen, Niek R, Claudine JC Lamoth, Andreas Daffertshofer, Gün R Semin, and Peter J Beek. 2008. Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. *Neuroscience letters* 432(2):88–93.

Unhelkar, Vaibhav V, Ho Chit Siu, and Julie A Shah. 2014. Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks. In *Proc. HRI'14*.

Vinciarelli, Alessandro, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27(12): 1743–1759.

Vygotsky, L.S. 1979. Consciousness as a problem in the psychology of behavior. *Journal of Russian and East European Psychology* 17(4):3–35.

Walker, Esteban, and Amy S Nowacki. 2011. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine* 26:192–196.

Walker, M.A., D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proc. EACL'97*, 271–280.

Walter, Henrik, Mauro Adenzato, Angela Ciaramidaro, Ivan Enrici, Lorenzo Pia, and B Bara. 2004. Understanding intentions in social interaction: the role of the anterior paracingulate cortex. *Cognitive Neuroscience, Journal of* 16(10):1854–1863.

Walters, Michael L, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay, Christina Kaouri, Sarah Woods, Chrystopher Nehaniv, David Lee, and Iain Werry.

2005. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *Robot and human interactive communication, 2005. ROMAN 2005. IEEE international workshop on*, 347–352. IEEE.

Wegner, Daniel M, Valerie A Fuller, and Betsy Sparrow. 2003. Clever hands: uncontrolled intelligence in facilitated communication. *Journal of personality and social psychology* 85(1):5.

Wegner, Daniel M, and Thalia Wheatley. 1999. Apparent mental causation: Sources of the experience of will. *American Psychologist* 54(7):480.

White, Sheida. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in society* 18(01):59–76.

Wilcox, Ronald, Stefanos Nikolaidis, and Julie Shah. 2013. Optimization of temporal dynamics for adaptive human-robot interaction in assembly manufacturing. *Robotics* 441.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research* 5:975–1005.

Yamazaki, Akiko, Keiichi Yamazaki, Yoshinori Kuno, Matthew Burdelski, Michie Kawashima, and Hideaki Kuzuoka. 2008. Precision timing in human-robot interaction: coordination of head movement and utterance. In *Proceedings of the sigchi conference on human factors in computing systems*, 131–140. ACM.

Yi, Weilie, and Dana Ballard. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics* 6(03):337–359.