Gaze Mechanisms for Situated Interaction with Embodied Agents

by

Sean Andrist

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

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 05/13/2016

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

Bilge Mutlu, Department of Computer Sciences, UW–Madison (co-chair)
Michael Gleicher, Department of Computer Sciences, UW–Madison (co-chair)
Kevin Ponto, Department of Computer Sciences, UW–Madison
David Shaffer, Department of Educational Psychology, UW–Madison
Adriana Tapus, Computer Science Department, ENSTA ParisTech, France

To my parents, *Connie & Eddie*,
for their unconditional support and love.

ACKNOWLEDGMENTS

In the Fall of 2010 I came to the University of Wisconsin–Madison convinced that I would be pursuing cutting edge research in animation. I nervously approached Michael Gleicher about possibly serving as my advisor, and he happily accepted that challenge. I also had a wonderful conversation with a relatively new professor, Bilge Mutlu, that entirely changed the direction of my research career. Mike and Bilge were starting a new project together on designing gaze mechanisms for virtual agents and robots (sound familiar?). The project combined some elements of the animation-y research I was interested in (animating the virtual agents themselves) with an area that I was not all that familiar with yet: human-computer interaction. Bilge was particularly persuasive that I join the project, and I am incredibly happy that I did. This project grew into the bulk of my dissertation work, and I am extremely thankful that I was able to find such a fascinating project so quickly and then pursue continuously for the next six years.

All members of my committee have been of invaluable help throughout my PhD. My advisors, Mike and Bilge, as mentioned above, sparked my interest in agents and robots, and have been there to offer guidance every step of the way since. They taught me how to conduct great research, from surveying existing literature to creating novel technical systems to conducting sound and thorough empirical evaluations. Equally if perhaps not even more importantly, they taught me how to professionally present myself and my research to the world, via papers, posters, and conferences talks; skills that I will make great use of throughout the entirety of my career. I hope they do not mind that I will invariably be asking them for more advice for many years to come.

Kevin Ponto became a friend when he joined the Visual Computing Lab as a post-doctoral researcher, and I was thrilled to have him on my committee when he became a professor, as he has always been a great resource to bounce new ideas off of. David Shaffer entered the scene as a collaborator on one of my later projects, and his unique insights have greatly helped to hone and refine my research and the arguments presented in this document. Adriana Tapus invited me to join her lab in France under an international fellowship for the spring semester of 2014. This was an incredible experience, both personally and professionally, and I am extraordinarily grateful to Adriana for being so warm and welcoming, guiding me through life in France and a fascinating research project.

I made many new lifelong friends (and hopefully, future collaborators) throughout my studies. I was lucky enough to be a part of two different labs at UW–Madison: the Human-Computer Interaction Lab (run by Bilge Mutlu) and the Visual Computing Lab (or as everyone actually calls it, the "graphics lab", run by Mike Gleicher). From the HCI lab,

I would especially like to recognize and thank Allie, Irene, Chien-Ming, Dan, Margaret, Steve, Zhi, Erin, Majid, Danny, Toshikazu, Javi, Faisal, and Erdem. From the graphics lab, I'm thankful to Tomislav, Brandon, Michael, Eric, Nathan, Alper, Danielle, Mike, Danny, and Chris. Having a close network of fellow grad students to commiserate with over a pint or two (or three) was absolutely essential to my survival.

Other collaborators from my research, both my dissertation research and various side projects, include Majd Sakr, Micheline Ziadee, Halim Boukaram, Kerstin Ruhland, and Wesley Collier. I would also be remiss if I didn't recognize my professors from the University of Minnesota for sparking my first interest in research as an undergraduate student: Victoria Interrante and Maria Gini.

Finally, I of course need to thank my family, whose unwavering support form the bedrock of all my current and future accomplishments. I would also like to thank all the new friends I made in Pittsburgh, Paris, Seattle, and conferences around the world, as well as longtime friends, especially Justin and Leah (for keeping my cat, and often myself, alive over the past six years), Pat, Brian, Andrew, Austin, Keith, Peter, Simon, and Sam.

Finally, I would like to recognize the funding resources that made all of this work possible: National Science Foundation awards 1017952, 1149970, and 1208632; the Grace Wahba fellowship; and the Chateaubriand research fellowship.

Significant Student Collaborators

Various aspects of this work benefitted from significant collaboration with other student researchers, which I would like to acknowledge separately here.

- *Tomislav Pejsa*: collaborated on the development of the virtual agent framework and implementation of the gaze shift model presented in Chapter 3.
- *Xiang Zhi Tan*: collaborated on the implementation of the gaze aversion model and study procedure for the Nao robot platform presented in Chapter 4.
- *Wesley Collier*: one of the developers of Epistemic Network Analysis, collaborated on the design and implementation of analyses presented in Chapter 5.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Computer interfaces represented as embodied agents, either virtually as animated characters or physically as humanlike robots, utilize a powerful metaphor of everyday social interaction in order to communicate effectively with human users. One of the most promising features of embodied agents is their ability to embody humanlike attributes and make use of nonverbal conversational cues just as people do. Gaze is a particularly important nonverbal signal in social interactions and is utilized in several rich communication mechanisms with which people are intuitively familiar.

This dissertation proposes the following thesis: *humanlike gaze mechanisms can enable both virtual agents and social robots to more effectively communicate with human users in situated interaction contexts*. To be fully situated, these mechanisms must be tightly linked with and responsive to the user, environment, and context in which they are deployed. Four mechanisms of social gaze are discussed. The first and most basic gaze mechanism, *gaze shifts*, handles the coordination of both head and eye movements to direct an agent's attention from one focal point to another. The next mechanism is *gaze aversion*, specifying when agents should avert their gaze away from their interlocutors and what they might accomplish by doing so. By coordinating its gaze with gaze motions tracked from a human collaborator, the agent can be become more tightly situated in the task and improve collaborative outcomes. This goal is captured in a mechanism referred to as *gaze coordination*. Finally, it is important to consider that a one-size-fits-all approach limits an agent's ability to account for cultural and individual differences across human users. The final mechanism presented in this dissertation, *gaze adaptivity*, demonstrates how the timing of an agent's gaze shifts can be manipulated in order to express extroversion or introversion, and how this personality expressed via gaze can be matched to a user's personality in order to improve motivation in a rehabilitation setting.

This dissertation makes a number of design, systems, and empirical contributions to research on human-robot interaction (HRI), intelligent virtual agents (IVA), human-computer interaction (HCI), multimodal interaction, and human communication. Overall, this dissertation contributes a set of gaze models that embody humanlike gaze mechanisms situated in specific interaction contexts, systems that implement these gaze models on virtual and physical agent platforms, and a number of user studies that demonstrate the effectiveness of these models and the general importance of well-designed gaze behaviors for achieving powerful social and cognitive outcomes for human users.

1 INTRODUCTION

The central thesis of this dissertation is that **humanlike gaze mechanisms can enable both virtual agents and social robots to more effectively communicate with human users in situated interaction contexts**. This dissertation presents new understanding of several complex gaze mechanisms as they are utilized in human-human interactions, new models of gaze that connect low-level gaze variables with high-level social and cognitive processes, implementations of these models on both virtual and physical platforms, and evaluations of the effectiveness of these models across tasks in achieving measurable outcomes.

This dissertation presents the culmination of several projects with the unified goal of empowering virtual agents and social robots—collectively referred to as *embodied agents* due to their embodied representations—to use gaze cues effectively in real-world tasks. Embodied agents are intelligent agents that are assigned virtual or physical bodies in order to interact with a virtual or physical world, as opposed to the traditional disembodied software agents. In order to empower embodied agents to use effective gaze cues, the methodology employed combines formal observational studies to build detailed models of human gaze mechanisms with practical efforts to build controllable computational models. These models are designed with controllable parameters in order to explore how they might be varied to achieve specific positive outcomes. This dissertation will demonstrate how gaze cues generated by these models on embodied agent platforms are able to evoke positive social and cognitive responses, and discuss the applicability of these results across agent representations and task contexts.

This dissertation focuses on *social gaze*—any gaze that can be interpreted as communicative by an observer. Social gaze includes gaze movements utilizing the eyes, head, and/or body, that are intentionally expressive, such as gaze aversions that are designed to communicate thoughtfulness (Emery, 2000). This does not include gaze movements that are not typically perceived by others during social interactions, such as gaze shifts that happen in isolation, or visual processing techniques on an agent's scene camera that do not involve changing the agent camera's point of focus.

A critical component of this thesis is the claim that in order to be truly effective, an agent's gaze behaviors must be *situated*, which means that they must be tightly linked and responsive to the environment and context—including the task, social setting, and participation structure—in which they are deployed (Hendriks-Jansen, 1996). The agent must adapt its gaze to, e.g., the characteristics and behaviors of the human user, the goals of the interaction, the location of relevant objects in its environment, and so on. Generating effective gaze behavior requires more than simply pointing the agent's eyes

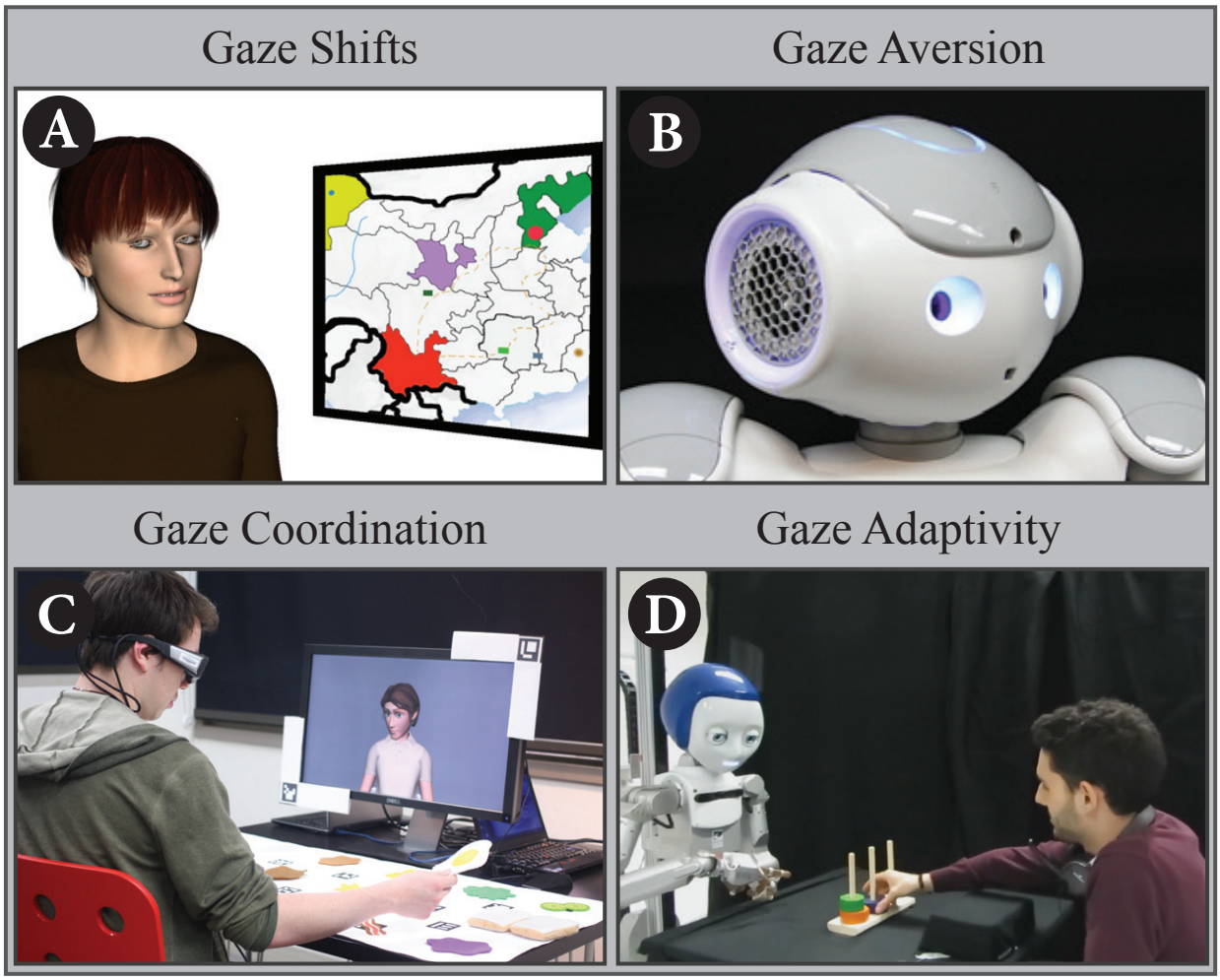

Figure 1.1: This dissertation presents new understanding, models, and evaluation of four gaze mechanisms for virtual agents and social robots: (A) gaze shifts, (B) gaze aversion, (C) gaze coordination, and (D) gaze adaptivity.

in the right direction—gaze comes from a combination of movements of the eyes, head, and body (Emery, 2000). Much of the subtlety and communicativeness of gaze comes from the details of timing, the usage of different body parts, and the degree of subtle fluctuation in the movements. These variables that make up an agent's gaze behavior must be determined contingently on a number of multimodal features in the interaction. Approaches to generating gaze movements for agents must capture and represent these complex contingencies in a manner that provides sufficient fidelity to serve as effective gaze mechanisms and appear natural. In order to be the most useful to interaction designers, these behaviors also need to be highly controllable, adaptable to the agent's goals, and flexible enough to work for both virtual and physical embodiments.

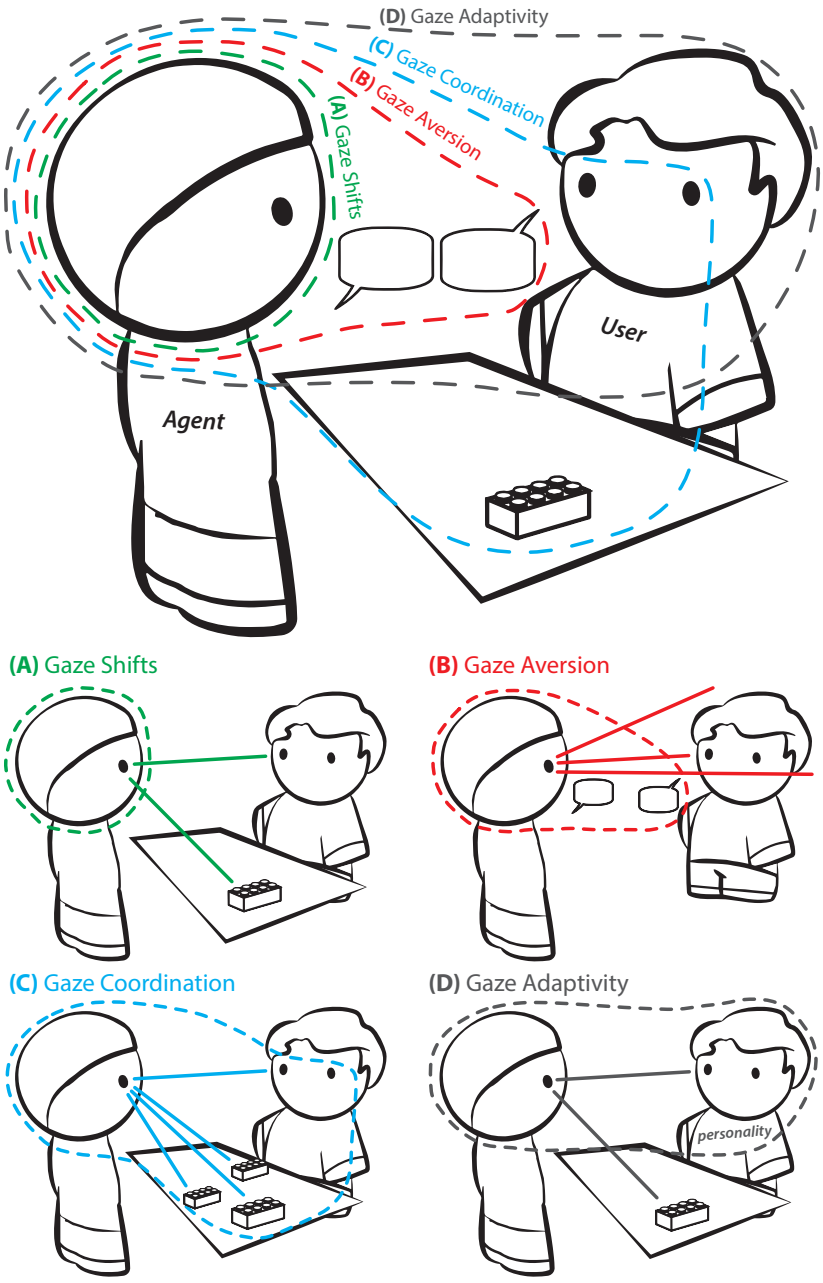

Figure 1.2: The gaze mechanisms presented in this dissertation can be characterized by the number of situated factors that they act contingently on and a context of possible gaze targets. Each progressive mechanism generally widens the scope of contingencies. (A) Gaze Shifts – contingent on internal variables of the agent. Gaze to user and task-relevant object. (B) Gaze Aversion – contingent on speech. Gaze to user and deflections to the side, up, or down. (C) Gaze Coordination – contingent on user gaze. Gaze to user and any task-relevant objects. (D) Gaze Adaptivity – contingent on user personality and task state. Gaze to user and task-relevant object.

Four mechanisms of social gaze are discussed in this dissertation (Figure 1.1 and Figure 1.2). The first step in designing social gaze into artificial systems is to determine the precise mechanics of *how* to carry out gaze motions. Chapter 3 presents the development and evaluation of the most basic gaze mechanism, *gaze shifts*, coordinating both head and eye movements to direct an agent's attention from one focal point to another (Figure 1.2A). This chapter also explores how different parameter settings in the gaze shift model can lead to different high-level outcomes in an educational setting. Once an agent has the ability to execute gaze movements naturally and effectively, the next consideration is *when* to carry out gaze shifts. This is particularly important in dyadic conversations with agents, in which the agent will by default make continuous eye contact with its human interlocutor. Chapter 4 presents the development of a conversationally situated model of gaze, focusing on the mechanism of *gaze aversion*—when agents should *avert* their gaze away from their interlocutors and what they might accomplish by doing so (Figure 1.2B).

In more complex interactions involving physical collaboration over a shared task space, the agent must be able to effectively distribute its gaze across relevant task objects in addition to its human collaborator. By coordinating the agent's gaze with gaze motions tracked from the human collaborator, the agent can become more tightly situated in the task and improve collaborative outcomes. This goal is captured in a mechanism referred to as *gaze coordination*, modeled and evaluated in Chapter 5 (Figure 1.2C). Finally, it is important to consider that a one-size-fits-all approach limits an agent's ability to account for cultural and individual differences across human users. Chapter 6 presents a mechanism of *gaze adaptivity* that demonstrates how the timing and duration of an agent's gaze shifts can be manipulated in order to express extroversion or introversion, and how this personality expressed via gaze can be matched to a user's personality in order to improve motivation in a rehabilitation setting (Figure 1.2D).

1.1 Motivation

Computer interfaces represented as embodied agents, including both virtual agents and humanlike robots, utilize a powerful social interaction metaphor in order to communicate effectively with human users. Embodied agents hold great potential across a range of application domains including education, training, rehabilitation, entertainment, and collaboration. In each of these areas, agents will take on a variety of roles, such as educational assistants, providers of technical support, therapists, companions, coaches, and personal trainers (Breazeal, 2003; Cassell, 2001). These roles each have their own associated interaction goals, such as improving comprehension, managing the flow of conversation,

engendering feelings of rapport, increasing the user's level of motivation, and so on.

Technologies with robotic features in particular have long been envisioned as utilizing social competence to integrate seamlessly into the working and living environments of people. In recent years, this vision has started to become reality, with the development and commercialization of social robotic products such as Jibo[1] and Pepper[2] that serve as family companions in homes, Baxter[3] and Hospi[4] that work alongside people in organizations such as factories and hospitals, and Autom[5] that takes on the role of a personal weight loss coach. Moving forward, social robots are expected to take on even larger roles in areas such as education, collaboration, and rehabilitation.

Virtual agents have also been utilized in a wide array of applications, such as embodied conversational agents that interact with and attempt to develop relationships with humans (Bickmore and Picard, 2005), pedagogical agents in tutoring systems (Rickel and Johnson, 1999), human-controlled avatars in virtual reality or online games (Vilhjálmsson and Cassell, 1998), simulated humans in virtual environments (Shao and Terzopoulos, 2005), interactive computer-controlled characters in video games, and as characters in pre-rendered animated films. Research has shown that virtual agents employing a range of communicative cues can motivate people (Mumm and Mutlu, 2011a), help them learn better (Lusk and Atkinson, 2007), and teach them social skills (Tartaro and Cassell, 2007).

Although they show great promise in these roles and beyond, embodied agent technologies still lack the full range of fine-grained verbal and nonverbal behaviors that would allow them to embody the rich space of human cognitive and communicative mechanisms. One of the most promising features of embodied agents is their ability to embody human-like attributes and make use of nonverbal conversational cues just as people do (Cassell et al., 1999a). These attributes and cues form rich communication mechanisms with which people are intimately familiar. Embodied agents can facilitate intuitive interaction through the use of communicative cues such as speech, facial expressions, gestures, and gaze. By employing such cues in interaction, agents can activate in people key social and cognitive processes, in turn eliciting significant positive outcomes such as improved learning and rapport in key application domains such as education (Lester et al., 2000), collaboration (Rickel and Johnson, 2000), and therapy (Tartaro and Cassell, 2006).

Gaze is a particularly important nonverbal signal that people employ in social interactions, compared with pointing, body posture, and other behaviors. Evidence from

[1] http://www.myjibo.com
[2] http://www.aldebaran.com/en/a-robots/who-is-pepper
[3] http://www.rethinkrobotics.com
[4] http://news.panasonic.net/archives/2013/1105_24824.html
[5] http://myautom.com/about

psychology suggests that eyes are a cognitively special stimulus, with unique "hard-wired" pathways in the brain dedicated to their interpretation (Emery, 2000). Gaze cues are particularly important communicative cues for achieving significant high-level social and communicative goals, such as improving listener comprehension, indicating interest in or appraisal of objects and people, managing the flow of conversation, engendering feelings of rapport and affiliation, expressing complex emotions, and facilitating interpersonal processes (Argyle and Cook, 1976; Kendon, 1967; Bayliss et al., 2006; Mason et al., 2005).

In order to enable artificial agents to achieve such significant effects, interaction designers need models that capture key variables of gaze and generate appropriate gaze behaviors. Well-designed gaze mechanisms—e.g., gazing at turn-taking boundaries during conversation—have been shown to result in more efficient task performance and more positive subjective evaluations (Heylen et al., 2002). However, creating effective gaze cues for agents is still an open challenge, as humans have built and finessed subtle but complex patterns in which they use these cues over thousands of years of evolution and developed a sensitivity to observing them in others (Parke and Waters, 2008). Furthermore, how these design variables might be varied to achieve specific high-level social and cognitive outcomes is an open subject of inquiry in psychology and communication research and thus still not fully understood.

The specific mechanisms of gaze that an agent should utilize will depend on the context and goals of the interaction. An on-screen tutoring agent may want to express attention to and engagement with a user by performing frequent mutual gaze, while a collaborative assembly-line robot may prioritize task-focused gaze that enables joint attention and object reference. Gaze mechanisms that are contextually situated will be the most successful both in terms of objective task performance and subjective perceptions of human users. This dissertation presents several gaze mechanisms for agents designed for contexts including education, conversation, collaboration, and rehabilitation.

1.2 Research Questions

This dissertation poses a number of research questions across each chapter. These questions are listed here, and the rest of this document presents a number of computational models built from human observation, research implementations, user studies, and discussions to address them.

Gaze Shifts (Chapter 3)

- How do people coordinate their head and eyes to carry out shifts in gaze?
- How can we enable embodied agents to carry out humanlike gaze shifts while maintaining control and flexibility for designers?
- What are the positive interaction outcomes of an agent utilizing humanlike gaze shifting, and how do these effects differ when the agent carries out gaze shifts in different ways?

Gaze Aversion (Chapter 4)

- When, why, and in which direction do people avert their gaze from one another in open conversation?
- How can we enable embodied agents to appropriately avert their gaze in conversations with people?
- What are the positive interaction outcomes achievable by having an agent avert its gaze in a humanlike way?

Gaze Coordination (Chapter 5)

- When there are task-relevant objects in the environment, how should an agent distribute its gaze to these objects?
- How do a collaborating dyad's gaze movements—to each other and over a shared task space—coordinate with each other over the course of an interaction?
- How can we design agent gaze behaviors that similarly coordinate with a human collaborator's gaze behaviors?
- What might an agent achieve by coordinating its gaze with user gaze in a humanlike way?

Gaze Adaptivity (Chapter 6)

- How should an agent adapt its gaze to the individual characteristics of its human user?

- How does a person's gaze motions reflect their underlying personality, specifically extroversion?
- How can we design gaze behaviors for agents that similarly signal a specific personality type?
- How do people with different personalities utilize gaze when attempting to motivate others to carry out tasks?
- How can we design gaze behaviors for therapeutic agents that are motivational for human users in rehabilitation with different personalities?

1.3 Methodology

To address the above research questions, this dissertation presents the design of several gaze mechanisms for embodied agents. In order to design a specific mechanism, we first need to understand the social system we are trying to replicate as it occurs in humans. This understanding gives us a computational understanding of the mechanism in question, allowing us to design agents that fit into a sociotechnical system. Once this sociotechnical system is implemented, we must evaluate whether or not it is a good one, measuring interaction outcomes in a similar context as where we studied the initial social system. This dissertation, therefore, utilizes a theoretically and empirically grounded, iterative methodology with three phases—understanding, building, and evaluating—for designing situated gaze mechanisms for embodied agents.

The first phase, *understanding*, involves the combination of social science literature review with data collection in laboratory studies. The goal of this phase is to connect high-level effects documented in the literature to specific mechanisms and low-level cues that can be implemented on agents. Current social-scientific knowledge provides us with descriptive and predictive top-down theories of how and to what purpose social cues are deployed in everyday human interaction, but the theories often lack low-level design specifications. This dissertation seeks to address this knowledge gap by augmenting high-level social-scientific knowledge with controlled human-human interaction lab studies to fill out the low-level details.

In this phase, naive participants are invited to the laboratory to enact a scenario in which embodied agents are envisioned to someday serve a role, e.g., in a collaborative setting where one participant instructs the other on how to correctly complete a specific task. The enactment sessions are recorded using devices such as high-definition cameras,

microphones, gaze trackers, and 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. The collected data usually requires additional processing either to annotate with features of interest (e.g., speech transcription, coding of gaze aversions, marking actions, etc.) or to reduce sensor noise.

In the second phase, *building*, the knowledge gleaned in the understanding phase is converted into computational representations of gaze behavior that use low-level gaze variables as building blocks to construct and control gaze mechanisms that communicate high-level social and cognitive states. The approach to creating these computational representations is to integrate rules from theory (e.g., that a speaker looks away from an addressee to start a speech turn) and findings from empirical data in the understanding phase (e.g., the distribution of timings before the start of the turn that speakers first look away) into hybrid rule-based/stochastic models. These models are designed with controllable parameters that allow investigation of their impact on high-level social and cognitive outcomes in a principled way and are then implemented as behavioral scripts attached to virtual animated characters or behavioral applications executed on a robotic platform. The implemented system usually involves various components that can manage real-time sensor inputs, make decisions about what, how, and when to act, and control an agent based on the decisions.

The third phase, *evaluation*, involves the testing of research hypotheses in human-agent interaction experiments to assess whether the gaze behaviors generated by the control models effectively evoke high-level social and cognitive responses in human users, such as higher rapport, better recall, or better task performance. This evaluation takes the form of controlled laboratory experiments in which participants, different from those who participated in data collection for the understanding phase, are asked to perform interactive tasks with embodied agents. The experimental task is generally similar to that in the data collection.

The goal of this phase is to obtain an understanding of the performance of the system and analyze how the participants respond to manipulations in the agent's gaze behaviors. This understanding is quantified using a variety of objective, subjective, and behavioral measures. Depending on the context, objective measures usually cover the dimensions of task performance (e.g., time, errors); subjective measures can include users' perceptions of the agent in terms of different characteristics (e.g., intelligence, naturalness, and competency); behavioral measures capture how users respond to the robot behaviorally (e.g., duration of gazes to the agent or specific objects).

As reflected in the understanding phase, this methodology reflects an explicit choice to

develop human-centered models of gaze constructed from theory and direct observations of human behavior. However, one could imagine a completely different methodology that results in novel behavioral mechanisms for artificial agents designed from the ground up, completely divorced from "doing what people do." This dissertation focuses on human-centered models for a few reasons. For people who observe or interact with these agents, humanlike gaze behaviors will serve as intuitive, familiar system behaviors that they are hardwired to interpret and that should not require learning. These behaviors can enable agents to tap into specific human perceptual and cognitive mechanisms to elicit specific responses or achieve targeted goals, such as generating the "correct" amount of gaze aversion with users to effectively moderate intimacy levels, ultimately serving as applications that effectively support the goals of their users. As a designer of agent interaction, studying humans provides a starting blueprint for creating agent behaviors which can then be explicitly altered through manipulatable model parameters. One can then develop specific hypotheses (from social science literature) about how people will perceive and respond to these behaviors.

As a longer term goal beyond the scope of this dissertation, research efforts will eventually turn toward developing socially interactive technologies with superhuman competence, not only in terms of their artificial intelligence and reasoning but also their interactional capabilities. Accomplishing such a feat will require modified or wholly new methodologies, e.g., automatically selecting only the "best" interaction exemplars from a corpus of human-human data for modeling and extending upon, or fully design-based development of behavioral mechanisms from the ground up. Current state-of-the-art agents are still very far below human-level interaction competence, so a methodology that targets average human capabilities is a natural choice at this stage in the research field.

Virtual and Physical Embodiments

This dissertation includes research on both virtual and physical platforms. Each embodiment has its own set of strengths and weaknesses, and interesting research questions arise when considering how one might design effective behaviors that work across both types of platforms. The virtual agents and social robotics communities have separately investigated a number of gaze mechanisms and their conversational effects. However, it is often unclear how these mechanisms might be accurately translated between the virtual and physical realms. Robots and virtual agents differ along a number of social dimensions, including realism, social presence, lifelikeness, and physical proximity (Powers et al., 2007). Several studies have demonstrated effects of these differences (Bainbridge et al., 2011; Kidd and

Breazeal, 2004; Powers et al., 2007).

A major challenge for expressing gaze with virtual characters on 2D displays comes from the Mona Lisa effect (Al Moubayed et al., 2012). In general, the perception of 3D scenes on 2D displays is prone to a number of artifacts and illusions (Gregory, 1997). The Mona Lisa gaze effect, named after the famous Mona Lisa painting, refers to the perception that Mona Lisa's gaze rests steadily and simultaneously on any viewer, no matter where they are standing in the gallery room. The same happens for virtual agents on 2D displays. If there are multiple human participants, each observer shares the same perception such that the agent either makes eye contact with everybody simultaneously, or nobody. It is also difficult to direct an agent's gaze to absolute locations in the real world, as the agent's gaze is perceived differently relative to the viewer's location.

Thus, a major opportunity of physical embodiments is that robots have the ability to more easily direct their gaze to participants and objects in the physical world. However, a major challenge of designing gaze for robot platforms lies in the fact that robots are often more restricted than virtual agents in terms of their affordances. For example, robots often do not have articulated eyes. Previous work has demonstrated that this challenge can be overcome; a robot with no articulated eyes can still produce effective gaze cues through appropriate movements of its head alone (Mutlu et al., 2006).

A reasonable strategy when considering both virtual and physical embodiments, and one that is employed in Chapter 4, is to first design a behavior for virtual agents—which are generally more flexible and easier to control—and then retarget that behavior to a less-flexible robot platform. This process requires a consideration of what the minimum representation that behavior would need in order to evoke the desired response (Cassell and Tartaro, 2007). A number of challenges arise when attempting to retarget behaviors from a virtual agent to a humanlike robotic platform, including (1) the acceleration and speed of a robot's movements have both upper and lower limits, (2) due to physical inertia and communication latency, a robot will typically not react instantaneously to a command, and (3) robot expression has fewer degrees of freedom (Lohse and van Welbergen, 2012). The latter challenge is of particular importance if we attempt to retarget gaze mechanisms from a virtual platform in which agents can use both their head and eyes to a robotic platform without articulated eyes.

Research Platforms

Virtual agents — All virtual agent scenarios were implemented using a custom framework built within the Unity game engine[6]. This framework is part of a larger pipeline that

[6]http://www.unity3d.com/

allows for quickly producing sets of character models and importing them into Unity. All agent models were created using DAZ Studio[7] by parametrically modifying a base figure (Figure 1.1A,C). The characters were then exported into Autodesk 3ds Max where they were enriched with a set of keyframe animations (e.g., blinking) and lip-sync animations using the FaceFX 3ds Max plugin. Finally, the characters were exported into Unity, where they were extended with a suite of behaviors (gaze models, blinking, speech, gestures, etc.) implemented as reusable Unity C# scripts and robust enough to scale to characters with varying skeletal structures and morphologies. Agent speech was either pre-recorded (Chapter 3,4) or automatically generated from text using the Microsoft Speech API (Chapter 5,6).

Robots — One target platform for this work was the Nao (Chapter 4), a programmable humanoid robot manufactured by Aldebaran Robotics.[8] The Nao has 25 degrees of freedom, including two in its neck, and a multitude of sensors (Figure 1.1B). All behaviors for the Nao were implemented using the .NET SDK provided by Aldebaran.

The system described in Chapter 6 was implemented on the Meka robot platform (Figure 1.1D). The Robot Operating System (ROS) was used to handle the execution and communication amongst each of the system components. The Meka robot is designed to work in human-centered environments, featuring compliant force control throughout its body, a sensor head, durable and strong hands, and an omnidirectional base. Each arm has seven degree-of-freedom (DOF) series elastic actuators and features high-strength force-controlled actuators and intrinsic physical compliance. The head is a seven DOF robotic active-vision head. The robot utilizes a variety of ROS modules for manipulation, navigation, and user interaction.

1.4 Contributions

This dissertation makes a number of design, systems, and empirical contributions to research on human-robot interaction (HRI), intelligent virtual agents (IVA), human-computer interaction (HCI), multimodal interaction, and human communication. Overall, this dissertation contributes a set of gaze models that embody humanlike gaze mechanisms situated in specific interaction contexts, systems that implement these gaze models on virtual and physical agent platforms, and a number of user studies that demonstrate the effectiveness of these models and the general importance of well-designed gaze behaviors for achieving powerful social and cognitive outcomes for human users.

[7] http://www.daz3d.com/
[8] http://www.aldebaran-robotics.com/

Design Contributions

The design contributions of this dissertation advance our understanding of human gaze mechanisms from a computational point of view. To human communication research, this dissertation contributes new knowledge and computational models of human gaze mechanisms. To HRI, IVA, and HCI research, the design contributions include a set of models and controllable parameters—such as gaze target, gaze triggers, frequency, and duration—that designers can use to create gaze behaviors for agents that can be manipulated to obtain social and cognitive outcomes.

- New knowledge of how low-level gaze variables and gaze mechanisms might achieve high-level social and cognitive effects, connecting current knowledge in the social sciences with a more computational foundation (Chapters 3-6).
- A computational model of gaze shifts—a fundamental building block of overall gaze behavior, intentionally redirecting gaze to specific targets—that coordinates eye and head movements in a humanlike way (Chapter 3). This model serves as a core component of all the subsequent gaze models.
- A computational model of gaze aversion—intentional shifting of gaze away from a partner's face—that specifies when and in which direction agents should avert their gaze when conversing with people in order to appear more thoughtful, regulate turn-taking, and maintain a comfortable level of intimacy (Chapter 4).
- A demonstration of how a new analysis technique, Epistemic Network Analysis, can be used to obtain a detailed and nuanced understanding of coordinated referential gaze patterns arising in dyadic physical collaborations. In particular, this analysis revealed (1) how a collaborating dyad's gaze behaviors unfold over the course of an interaction, (2) how the alignment of gaze behaviors shift throughout the interaction, and (3) how coordinated gaze behaviors differ in interaction sequences that include breakdowns and/or repairs (Chapter 5).
- A computational model of gaze coordination that specifies how an agent should deploy its gaze over a shared collaborative workspace, tightly linked to the human user's gaze, speech, and actions (Chapter 5).
- A computational model of personality-expressive gaze that specifies the frequencies and lengths of gazes toward a shared task space and toward the human partner in order to convey introversion or extroversion (Chapter 6).

Systems Contributions

Each of the computational gaze models listed above was implemented on a virtual agent and/or humanlike robot platform. Each system also required additional competencies to be implemented in order for the agent to autonomously perform the task associated with the scenario it was embedded in. In some cases, implementations were extended in interesting ways or retargeted from one type of embodiment to another.

- A comprehensive virtual agent framework, implemented in the Unity game engine, that was utilized for all virtual agent scenarios presented in this dissertation. This framework includes animated character models as well as behavior modules for gaze, speech, gestures, task logic, and so on (Chapters 3-5).
- An implementation of the gaze shift model for virtual agents in an educational scenario in which the agent gave one-sided lectures to a human listener while periodically gazing toward visual content supportive to the lecture (Chapter 3).
- An implementation of the gaze aversion model that autonomously plans and executes gaze aversions for virtual agents in a scenario where the agent engages in a structured conversation with a human user (Chapter 4).
- An implementation that retargets the gaze aversion model to a robot platform with more limited affordances, requiring a number of new techniques—idle motion, face tracking, and predictive filtering—in order to make it successful in a similar structured conversation scenario (Chapter 4).
- An implementation of the gaze coordination model in a virtual agent system that can autonomously collaborate with human users in a sandwich-making training scenario, including gaze tracking, speech recognition, and action tracking (Chapter 5).
- An extension of the gaze coordination implementation to utilize head pose tracking as a lower fidelity and lower cost proxy for full gaze tracking (Chapter 5).
- An implementation of the gaze coordination model and sandwich-making scenario in head-mounted virtual reality with a virtual agent (Chapter 5).
- An implementation of the personality-expressive gaze model on a robot platform that can autonomously instruct, motivate, and monitor people in a Towers of Hanoi puzzle-solving task (Chapter 6).

Empirical Contributions

In addition to modeling and implementation, each situated gaze mechanism presented in this dissertation was evaluated in one or more user studies, contextualized in specific scenarios. These studies provide a better understanding of the social, cognitive, and behavioral outcomes achievable via carefully deployed gaze mechanisms in human-agent interaction. All studies utilized carefully designed experimental paradigms for studying how subtle manipulations in situated gaze mechanisms can target specific outcomes under various conditions.

- Evidence across all studies that seemingly subtle manipulations in an agent's gaze behavior can lead to powerful high-level interaction outcomes (Chapters 3-6).
- Evidence that the presence of an agent can improve recall in an educational scenario, compared with having the same lecture content expressed through audio alone (Chapter 3).
- Evidence that head alignment, one parameter of the gaze shift model, can be manipulated to achieve either better rapport or better recall in a lecture-style educational scenario (Chapter 3).
- Evidence that virtual agents using gaze aversions generated by the presented computational model were perceived as thinking, elicited more disclosure from human interlocutors, and effectively managed turn-taking (Chapter 4).
- Evidence that gaze aversions expressed by social robots utilizing the presented gaze aversion model are perceived as intentional, and that robots can use gaze aversions to appear more thoughtful and effectively manage the conversational floor (Chapter 4).
- Evidence that the gaze coordination model can improve collaborative outcomes for a virtual agent task-training system in terms of task time, number of errors made, recall, and the subjective preferences of users (Chapter 5).
- Evidence that the gaze coordination model is comparably effective even when gaze tracking is replaced with lower fidelity head pose tracking (Chapter 5).
- Evidence from an online study that the personality-expressive gaze model is effective in conveying introversion or extroversion noticeably for robots (Chapter 6).
- Evidence that matching a robot's personality—introversion or extroversion expressed via gaze alone—to that of a user can positively motivate them to participate longer in

a repetitive task, particularly when their intrinsic motivation going into the task is low (Chapter 6).

1.5 Dissertation Overview

This dissertation presents four threads of research on designing situated gaze mechanisms for embodied agents, described in the following chapters: modeling *gaze shifts* for agents that can be manipulated to target specific high-level interaction outcomes (Chapter 3), modeling *gaze aversion* motions (disengaging mutual gaze) for both virtual agents and robots in a conversational setting (Chapter 4), modeling gaze mechanisms that *coordinate* a virtual agent's gaze with the user's gaze while collaborating on a physical task (Chapter 5), and modeling gaze behaviors that can match a robot's *personality* with the user in order to increase *motivation* in a therapeutic task (Chapter 6).

2 BACKGROUND

The earliest research into communicative gaze was led by the virtual agent community in the 1990s (Thórisson, 1994; Vilhjálmsson and Cassell, 1998; Cassell et al., 1999a). Virtual agents were created with eye gaze motions as a means for capturing attention, maintaining engagement, and increasing conversational fluidity with human users (Cassell, 2000). Roboticists began introducing communicative eye gaze into their systems in the late 1990s, in robots such as Cog (Brooks et al., 1999) and Kismet (Breazeal and Scassellati, 1999). Research in human-robot interaction (HRI) has investigated the positive outcomes achievable through a robot's gaze behavior, including increasing the robot's competence in conversations with people (Mutlu et al., 2009a, 2012), enabling joint attention and referential communication (Huang and Mutlu, 2012; Staudte and Crocker, 2011), and improving upon the robot's ability to hand objects to people (Moon et al., 2014).

This dissertation builds from the history of previous work in order to develop generalizable models of gaze behavior that are effective for virtual and physical platforms. It is informed by literature on social gaze behavior in human communication research, virtual agents, and robotics. The literature survey below encompasses a review of related work from all three literatures, with a focus particularly on understanding how agents can use situated gaze mechanisms to interact more effectively with people. Before that, however, it will be useful to clarify some gaze-related terminology used throughout the dissertation.

- *saccade* — Quick movement of the eyes between two points of fixation (Cassin et al., 1984).
- *fixation* — The maintaining of gaze on a single location (Cassin et al., 1984). Regular human gaze behavior alternates between saccades and fixations, except in smooth pursuit motions that are outside the scope of this dissertation.
- *gaze shift* — An intentional redirection of gaze (moving the eyes, head, and/or torso) toward a particular target in the interaction context (Binder et al., 2009) (Chapter 3).
- *mutual gaze* — Often referred to colloquially as "eye contact," it is eye gaze that is directed from one agent to another's eyes or face, and vice versa (Argyle and Cook, 1976). Face-directed gaze without reciprocity is not mutual gaze.
- *gaze aversions* — Shifts of gaze away from the main direction of gaze, which is typically a partner's face (Doherty-Sneddon and Phelps, 2005). Gaze aversions can occur in any direction, though evidence suggests that the purpose of the aversion influences the direction of the shift (Andrist et al., 2013b) (Chapter 4).

- *referential gaze* — Also called deictic gaze, it is gaze directed at an object or location in space (Moore and Dunham, 1995). Such gaze sometimes occurs in conjunction with verbal references to an object, though it need not accompany speech (Chapter 5).
- *joint attention* – Involves sharing attentional focus on a common object (Moore and Dunham, 1995). It is often composed of several phases, beginning with mutual gaze to establish attention, proceeding to referential gaze to draw attention to the object of interest, and transitioning back to mutual gaze to ensure that the attention is shared.

The remainder of this chapter discusses prior research on the functions of gaze in human communication and presents previous work on designing social gaze mechanisms for both virtual agents and social robots. Each section discusses the perception of gaze cues produced by humans, virtual agents, or robots, what positive interaction outcomes can be achieved with gaze, and how gaze is deployed in situated interactions.

2.1 Gaze in Human Communication

Gaze serves a number of very important functions in everyday human behavior. People use their observations of others' eye gaze to guide everything from conversation (Kleinke, 1986) to speech (Argyle and Cook, 1976) to attention (Frischen et al., 2007). This section surveys previous research on the mechanisms and functions of gaze in everyday human communication. A thorough survey of this topic was also published by Kleinke (1986).

Functions of Gaze as a Social Signal

Gaze can be used as a clear signal of attention, as people generally look toward what they are attending to. In conversations, gaze predicts the target of conversational attention. When someone is listening, the person they look at is the person being listened to around 88% of the time (Vertegaal et al., 2001). When someone is speaking, they are looking at the target of their speech around 77% of the time (Vertegaal et al., 2001). Other studies have found similar results, such as Cappella and Pelachaud (2002) observing that gaze is directed at conversational partners around 80% of the time. Humans also tend to be very particular about who they prefer to gaze at. People look more at those who they like and those with whom they are interpersonally involved (Argyle and Cook, 1976).

Gaze is used to signal availability for interaction. When they must pass objects back and forth, object handovers between people rely on a receiver signalling readiness to receive an object by gazing at their partner (Strabala et al., 2012, 2013). Caregivers in a nursing

home demonstrate their availability to their patients through broadly distributed gaze, and people naturally wait for caregivers to establish mutual gaze before requesting assistance (Yamazaki et al., 2007). Gaze is also an important component of greeting behavior which appears in very diverse cultures and may be innate. Indeed, mutual gaze between two people is often the first step to a social encounter (Argyle and Cook, 1976).

Gaze is a powerful method by which emotions are expressed to others (Izard, 1991). A speaker's gaze behavior affects how their emotions are perceived by the addressee (Pourtois et al., 2004). Gaze is also closely tied to personality, with extroverts commonly engaging in significantly more mutual gaze with their conversational partners than introverts do (Rutter et al., 1972). Chapter 6 further explores the relationship between gaze and personality.

Gaze relates directly to the syntax of speech. People generally look away from their conversational partner when beginning the theme of the sentence—which indicates what the overall sentence is about—and look toward their partner when beginning the rheme of the sentence—which provides information or exposition about the theme (Cassell et al., 1998).

The topic of conversation also affects gaze behavior. People exhibit less mutual gaze when their conversation involves high levels of intimate self-disclosure (Kang et al., 2012). The dynamics of dyadic gaze can be used to determine the context of utterances during interaction, such as conveying a fact or answering a question (Admoni and Scassellati, 2014). When referring to objects or locations around them, a person's gaze is closely tied to the content of their speech. For example, Griffin (2001) and Meyer et al. (1998) demonstrated that deictic gaze shifts generally occur 800 - 1000 ms before the object being gazed at is mentioned in speech. This finding is directly integrated into the gaze shift implementation presented in Chapter 3.

Gaze is important for facilitating turn-taking and conversational floor management (Kendon, 1967). Gaze functions here as a signal regulating the exchange and maintenance of speaker roles. A typical pattern of interaction when two people converse with each other consists of the listener maintaining fairly long gazes at the speaker, interrupted only by short glances away (Kendon, 1967). The listener will usually spend about 75% of the time looking at the speaker (Argyle, 1988). When exchanging speaking turns in conversation, a pattern is frequently observed in which the first speaker finishes speaking, looks toward their interlocutor and engages in momentary mutual gaze, and finally the second speaker averts their gaze and begins their speaking turn (Novick et al., 1996). By looking away at the beginning of an utterance, the speaker strengthens his or her claim over the speaking turn. Looking away during a pause in speech is also used to indicate that the conversational floor is being held and that the speaker should not be interrupted (Kendon, 1967). The

timing of gaze motions to facilitate turn-taking is explored further in Chapter 4.

Perceiving and Responding to Human Gaze Cues

From an evolutionary perspective, distinguishing another individual's visual perspective from one's own—perceiving and recognizing their gaze—is thought to be an important step in interpreting their intentions and thoughts about the world (Emery, 2000). Thus, people are very highly tuned to others' gaze direction. Three-month-olds shift their attention in the direction of an adult's gaze (Hood et al., 1998). In adults, seeing someone's eyes directed to the side—even in a still photograph—evokes rapid, reflexive attention shifts in the direction of the gaze (Langton and Bruce, 1999; Frischen et al., 2007). A series of controlled experiments tested this reflexive attention shift and found that it is resistant to conscious control. In these experiments, participants are shown a picture of a face gazing to one side. Even when they are told that they should look in the opposite direction of the gaze, their attention is still initially drawn to the direction of the gaze, slowing reaction times to the non-gaze-cued location (Driver et al., 1999; Downing et al., 2004; Friesen et al., 2004).

Achieving Positive Interaction Outcomes

In general, being gazed at by another person can result in a number of interesting effects. Gaze is predominantly interpreted as being attended to, and depending on the context of interaction, being gazed at can lead to discomfort from feeling observed, or lead to genuine social interaction (Argyle and Cook, 1976). A person who makes increased eye contact is associated with greater perceived dynamism, likeability, and believability (Beebe, 1976). In an interview scenario, people who spend more time gazing at an interviewer receive higher socioemotional evaluations (Goldberg et al., 1969).

An important construct in the study of nonverbal behavior is *immediacy*, defined as the degree of perceived physical or psychological closeness between people (Mehrabian, 1966; Christophel, 1990). Gaze has been found to be a significant component of immediacy, especially in the context of improving educational outcomes (Harris and Rosenthal, 2005). Students from primary school age through college have been shown to learn better when they are gazed at by the lecturer (Burgoon et al., 1986; Fry and Smith, 1975; Sherwood, 1987; Otteson and Otteson, 1980). Learning in these cases was usually measured by the students' performance on recall tasks. In a similar study, gazing into the camera during a video link conversation was shown to increase the recall of the viewer at the other end (Fullwood and Doherty-Sneddon, 2006). Gaze's positive effect on recall is usually attributed to its role as an

arousal stimulus, which increases attentional focus and therefore enhances memory (Kelley and Gorham, 1988). Chapter 3 seeks to further explore the impact of gaze on participant recall when gaze motions are parameterized into different behavioral profiles.

Nonverbal behaviors, especially gaze, have long been recognized in the social sciences literature as useful tools in persuading others to comply with requests or demands. When a collector of money for charity engaged in mutual gaze with possible donors, rather than looking at the collecting tin, they were more successful in receiving donations (Bull and Gibson-Robinson, 1981). Using a combination of gaze and touch has been found to be more successful than not using these nonverbal behaviors in getting people in malls to participate in interviews (Hornik and Ellis, 1988). Students are more likely to comply or at least partially comply when their teachers use more immediacy cues (including gaze), and are more likely to choose to reject requests made by teachers without such cues (Burroughs, 2007). Similarly, attraction and dominance increase compliance and cues of such are often expressed through gaze (Peters, 2007). Manipulating gaze to achieve higher compliance with agent requests is explored further in Chapter 6.

People do not make constant eye contact with each other, and averting one's gaze serves a number of important functions as well. For example, averting gaze has the practical benefit of improving cognitive performance. The frequency of speaker gaze aversions in conversation has been shown to be related to the difficulty of cognitive processing (Glenberg et al., 1998). This is because an interlocutor's face is rich in social information, and is a cognitively demanding visual target. With these aversions, a speaker signals to the listener that cognitive processing is occurring, creating the impression that deep thought or creativity is being undertaken in formulating their speech (Argyle and Cook, 1976). Gaze aversions are discussed in depth in Chapter 4.

Gaze in Situated Interaction

When people are manipulating objects, their eye gaze is tied to their task and intended action. Eye gaze typically reaches the object of interest before any movement of the hands has started (Land and Hayhoe, 2001). Though people fixate the same object while they act on it, eyes often shift to the next object in the task sequence before the action is completed on the current object (Land and Hayhoe, 2001). Objects not related to the task at hand are much more rarely fixated (Hayhoe and Ballard, 2005). The timing of gaze to task-relevant objects in relation to actions is a key part of the gaze coordination mechanism presented in Chapter 5.

In spontaneous, interactive dialogues relating to a common visual scene, participants'

eye movements are tightly coupled (Richardson et al., 2007). Moving the eyes closely in step with a speaker allows the listener to use spatial structure to organize information in the same way as the speaker (Richardson and Dale, 2005). Addressees also make use of the speaker's gaze as a cue for disambiguating references (Hanna and Brennan, 2007). Conversational partners' shared gaze toward referents is higher while they speak about those objects (Bard et al., 2009). Dual eye-tracking methods can be employed to better understand the role gaze plays as a conversational resource (Clark and Gergle, 2012). These methods are employed extensively in Chapter 5.

Summary

The above work shows that gaze as a whole is an important cue to create positive interactions. Gaze motions serve a number of different functions; they can act as intentionally communicative signals, illustrating the referent of a remark, disambiguating deictic expressions such as "this one" or "that one," expressing intimacy or dominance, or communicating various emotional states. But gaze shifts need not be intentionally communicative at all, such as when people avert their gaze upwards when thinking. Given all the different functions and the range of meaning the eyes and head might express, it is remarkable how people are able to precisely interpret what another's gaze actually means. How do people disambiguate gaze cues? Langton et al. (2000) explored the different spatial and temporal properties that can be used to disambiguate the meaning of different gaze shifts, but these cues are very subtle and timing is particularly important. For example, children with autism spectrum disorder (ASD) make approximately the same *quantity* of eye-contact overall, but not at the same *times* as children without autism (Baron-Cohen et al., 1995). Thus, if we are to design effective gaze mechanisms for artificial agents, it is important to get the details right.

2.2 Gaze Mechanisms for Virtual Agents

This dissertation builds on previous research in the area of virtual agents, sometimes called embodied conversational agents. Giving these agents more complex and human-like behaviors has been a longstanding goal, for example to make inhabited interfaces and agents in virtual reality environments more effective (Nijholt et al., 2000). The ability to use non-verbal communicative behavior is very important as it increases positive feelings of copresence and familiarity, and in general makes agents more effective communicators (Bailenson et al., 2006). Numerous approaches have been taken to increase the behavioral

realism of agents. For example, an approach called "digital chameleons" involves creating virtual agents that simply mimic the human interlocutor's nonverbal behavior, such as head movements (Bailenson and Yee, 2005). Other work has focused on modeling subsets of behaviors such as social dialogue (Bickmore and Cassell, 2001) and emotional expression (de Melo et al., 2011) to create positive outcomes such as increased feelings of rapport (Wang and Gratch, 2010) and trust (Bickmore and Cassell, 2001).

Perceiving and Responding to Agent Gaze Cues

Virtual agents that use turn-taking gaze during conversations are evaluated as more natural and more pleasant, and their conversation is rated as more engaging, than agents that use random gaze or no gaze in their communication (Garau et al., 2001). In an immersive virtual reality setting, researchers confirmed that people have more positive subjective evaluations of an agent when it performs conversationally-driven gaze than when it performs random gaze, but that effect depends on the agent's appearance (Garau et al., 2001). More realistic agents benefit greatly from humanlike conversational gaze, but low-fidelity agents, such as stick figures, are negatively perceived when using very humanlike gaze behavior (Garau et al., 2003).

A large body of work on virtual agents involves the modeling of nonverbal feedback signals for artificial listeners (Heylen et al., 2007) (see Bevacqua (2013) for an overview). In these situations, gaze has been studied in the context of backchannels during conversation, i.e., non-intrusive visual and acoustic signals provided to the speaker by listeners during their turn (Yngve, 1970). The Rapport Agent, developed by Gratch et al. (2007), generates nonverbal backchannels on the listening virtual agent. Conversely, Hjalmarsson and Oertel (2012) found that listening humans are more prone to provide backchannels of their own when the gaze of a virtual agent is directed toward them.

The effects of cultural and individual differences, including gender, on the conversational behavior of agents are also of importance. Jan et al. (2007) have simulated different cultural parameters related to gaze, overlap in turn-taking, and proxemics for rating by native speakers of North American English, Mexican Spanish, and Arabic. Studies in immersive and augmented reality environments have shown that users provide more personal space to agents that engage in mutual gaze with them (Bailenson et al., 2003), and users have a higher physiological arousal towards virtual agents that violate norms associated with their cultural backgrounds (Obaid et al., 2012). Furthermore, Vala et al. (2011) have utilized gaze in the creation of a model for varying the communication style of agents based on gender.

Achieving Positive Interaction Outcomes

An agent's gaze behavior has been found to play a key role in achieving rich interactions. Well-designed gaze mechanisms—e.g., shifting gaze at turn boundaries during conversation—result in increased task performance and more positive subjective evaluations (Heylen et al., 2002). Poor gaze behavior can be worse than the absence of gaze behavior. The positive effects of having an embodied agent—as opposed to only audio or text—can be completely lost if gaze is very poor or random (Garau et al., 2001). Gaze cues are particularly important in multiparty interactions, where a virtual agent can use its gaze to selectively engage human users in a group setting (Andrist et al., 2013a).

Wang and Gratch (2010) have shown that a virtual agent that continuously gazes toward a human interlocutor is able to increase perceptions of rapport when the gaze is accompanied by nonverbal indicators of positivity and coordination. Continuous gaze without these accompanying behaviors had a negative social impact. Bailenson et al. (2001) further examined the relationship between interpersonal space and eye contact (two indicators of intimacy) in virtual environments, finding that more interpersonal distance is maintained when the amount of mutual gaze is high.

In an interactive storytelling scenario, a virtual agent that modulates mutual gaze by shifting its gaze in reaction to a user's gaze is able to improve user perceptions of social presence and rapport (Bee et al., 2010). An agent can also use gaze, along with other nonverbal behaviors, to effectively shape a conversation with multiple participants according to its own intentions by directing gaze towards specific participants or averting gaze from participants in order to signal turn-taking in the interaction (Bohus and Horvitz, 2010). Cafaro et al. (2012) investigated how nonverbal behaviors—including smiling, gaze, and proximity—of agents during the approach to interaction lead to the formation of impressions related to personality and interpersonal attitude, highlighting particularly important relationships between gaze and judgements of friendliness.

Exactly how agents should perform gaze motions is an important consideration for successfully conveying emotional states to users (Fukayama et al., 2002). Previous approaches have sought to analyze gaze motions in animated films to create animation models that can automatically map between emotions and gaze animation characteristics (Lance et al., 2004; Queiroz et al., 2007; Lance and Marsella, 2010a). Some research has also considered the expression of emotions through gaze shifts that involve movements of both the torso and the head (Lance and Marsella, 2007). Empirical studies have been performed in order to link low-level gaze attributes from non-verbal behavior literature with observers' attribution of emotional states (Lance and Marsella, 2010b; Queiroz et al., 2008). For example, Cig et al. (2010) conducted user studies to show that changes in gaze behavior led to changes in user

perceptions of the arousal and dominance levels of agents. Li and Mao (2012) describe a rule-based approach to generate emotional eye movements based on the Geneva Emotion Wheel to enable virtual agents to convey different emotional expressions to users through eye movements. A data-driven approach was adopted by Busso et al. (2008) to generate expressive head movements from speech data.

Gaze in Situated Interaction

Previous research has explored how a virtual agent can become more situated in the user's physical environment through the use of a web camera and making gaze motions towards salient locations in the environment (Itti et al., 2004; Picot et al., 2007). Such saliency-based approaches are based on a neurobiological model of visual attention (Itti, 2000) and have been popular for animating the gaze movements of agents (Itti et al., 2006; Oyekoya et al., 2009). More recent efforts have focused on specific aspects of visual attention, e.g., the role of object relevance (Oyekoya et al., 2011) and task constraints (Kokkinara et al., 2011) on gaze behavior. Attentive presentation agents (Eichner et al., 2007) have been developed that utilize a user's visual attention, based on tracked eye fixations, to detect objects of interest and guide the behaviors of two virtual presentation agents who describe the items in further detail. A key challenge in all of these efforts is balancing bottom-up (low-level saliency) and top-down (high-level goal-based) visual attention for gaze behavior (Mitake et al., 2007).

A number of efforts have aimed to detect gaze direction from users to inform the behavior of virtual agents. Some of these efforts use the position of a user's head to determine an agent's reactive gaze behavior (Kipp and Gebhard, 2008), while others recognize an individual's gaze aversion behaviors given contextual information (Morency and Darrell, 2007) or identify an addressee during multiparty conversations to ensure that an agent responds appropriately when it has been addressed (Nakano et al., 2013b).

Existing Gaze Models and Systems

Previous work has created gaze models for virtual agents that are derived from human gaze behaviors in order to improve agents' abilities in turn management (Beskow, 1997; Cassell et al., 1999a,b; Pelachaud and Bilvi, 2003; Thórisson, 2002), determine where agents should be looking and why they should be looking there (Khullar and Badler, 2001; Itti et al., 2006; Lee et al., 2007; Poggi et al., 2000; Fukayama et al., 2002), and enable agents to make random eye saccades in idle situations (Cafaro et al., 2009) and face-to-face conversations (Lee et al., 2002). A variety of approaches have been used in prior work to model head and

eye motions for virtual agents, including both data-driven (Deng et al., 2005; Ma and Deng, 2009; Heck, 2007) and procedural (Peters, 2010) approaches. The Expressive Gaze Model (EGM) takes a hybrid data-diven/procedural approach to the animation of gaze shifts and focuses on communicating emotion by executing gaze shifts in different ways (Lance and Marsella, 2010a). In general, existing research has proven that more realistic gaze behavior for humanoid avatars results in considerably improved and fluid communication (Garau et al., 2001). At the very least, simply *conveying* gaze direction eases turntaking and is essential to establishing who is talking and listening to whom in multiparty mediated communication (Vertegaal, 1999).

A number of systems use speech as an input from which to generate facial expressions involving the head and eyes (Albrecht et al., 2002). Zoric et al. (2011) automatically generate facial gestures in real time from the prosodic information obtained from speech signals. Nods, head movements, blinks, eyebrow gestures, and gaze were generated using hidden markov models and stochastic parameters. Gaze direction was generated to lower at hesitation pauses and rise at the end of utterances to obtain listener feedback.

Several frameworks have been developed to express and coordinate multimodal behaviors, including gaze, on virtual agent systems. Foremost among these initiatives is SAIBA (Situation, Agent, Intention, Behavior, Animation) (Kopp et al., 2006). Behavior Markup Language (BML) (Vilhjálmsson et al., 2007), developed as one of the three stages of SAIBA, defines multimodal behaviors such as gaze, head, face, body, gesture, and speech in a human-readable XML format. BML allows the definition of multimodal behaviors by specifying temporal details for primitive action elements (see Krenn et al. (2011) for an overview).

Considering Eyelid Movement

The eyelids are not strictly part of the overall gaze mechanisms discussed in this dissertation, but their motion does interact with the human oculomotor system and must be considered particularly in virtual agent embodiments with very humanlike facial features. Normal eye blinks can be broken into spontaneous, voluntary, and reflexive subclasses, all of which have slightly different eyelid dynamics (VanderWerf et al., 2003). Endogenous blinks—blinks that are spontaneous—are the most interesting for human-agent interaction designers, because their frequency is linked to cognitive state and activity (Stern et al., 1984; Skotte et al., 2007). Various studies have linked blink occurrence to attentional processes (Nakano et al., 2013a), fatigue (Johns et al., 2007; Anderson et al., 2010), lying (Buller and Burgoon, 1998), and speech production (Nakano and Kitazawa, 2010). Blink rates are highly variable,

however. Studies have found a range of rates during, e.g., reading (1.4–14.4 blinks per minute) and conversation (10.5–32.5 blinks per minute) (Doughty, 2001).

The occurrence of individual blinks has been modeled in previous work as a Poisson process (Greene, 1986). Blinks also very often occur almost simultaneously with the onset of gaze movements, particularly large ones over 30° (Evinger et al., 1994). Previous research has noted that eyelid motion within a single blink is not uniform; the downward velocity is approximately twice as fast as the upward velocity, and the durations of up and down blink phases are nonlinear (Evinger et al., 1991; Guitton et al., 1991). Eyelid displacements, called *lid saccades* (Becker and Fuchs, 1988), always accompany vertical eye saccades. Lid saccades do not exhibit as much motion asymmetry between down and up phases as do blinks (Evinger et al., 1991; Guitton et al., 1991).

For virtual agents and robots with articulated eyelids, several blink models have been produced in previous work in which eyelid dynamics are modeled according to physiological data and account for both endogenous blinks and eyelid saccades (Steptoe et al., 2010; Trutoiu et al., 2011). Some of these models focus on higher level aspects of blinking, such as the timing and synchronization of blinks during head movements and conversations (Lee et al., 2002; Gu et al., 2008; Masuko and Hoshino, 2007). Peters (2010) compared the realism of different methods for simulating blink timing and confirmed the overall importance of regular blinking to improve people's perceptions of agents.

Summary

Virtual agents hold tremendous potential for computer interfaces with their unique ability to embody humanlike attributes (Cassell et al., 1999a). These attributes form rich communication mechanisms with which people are intimately familiar and afford intuitive interactions. Variations in these attributes activate key social and cognitive processes, in turn eliciting significant positive outcomes such as improved learning and rapport in key application domains such as education (Lester et al., 2000), collaboration (Rickel and Johnson, 2000), and therapy (Tartaro and Cassell, 2006). A deeper understanding of how humans use these attributes in social interactions might enable designers to create more effective virtual agents in these domains.

Previous research has explored a number of different gaze models for virtual agents at varying levels of controllability, interactivity, complexity, and effectiveness. These models are primarily generated with the goal of creating more engaging interactions and studying people's responses to particular gaze motions. While previous research has explored how gross manipulations in gaze behavior might shape user experience and perceptions of

virtual characters, how specific parameters in the control space for gaze might be mapped to specific outcomes in interaction has not been explored. This dissertation seeks to fill this knowledge gap.

2.3 Gaze Mechanisms for Social Robots

Robot gaze has been shown to increase the fluidity of conversation (Mavridis, 2015) or effectively direct a user's attention to relevant information in a tutoring setting (Johnson et al., 2000). Gaze can be used by socially assistive robots to demonstrate their engagement with and attention to a user (Tapus et al., 2007). Gaze can also reveal a robot's mental states, including its knowledge and goals (Fong et al., 2003). This section reviews related previous work on gaze from the robotics and human-robot interaction literatures.

Perceiving and Responding to Robot Gaze Cues

Seeing a robot perform natural gaze motions improves people's perceptions of that robot (Emery, 2000; Liu et al., 2012; Mutlu et al., 2006, 2012; Staudte and Crocker, 2009). When a robot is a listener in a multiparty conversation, seeing the robot track the conversation with its gaze elicits higher evaluations of that robot's comprehension and naturalness than seeing the robot perform random gaze turns between speakers (Kousidis and Schlangen, 2015). A robot that displays gaze focused on its conversational partner, but occasionally responding to motion in the background, is evaluated as more natural, humanlike, and attentive than a robot that exclusively focuses on the partner or that distributes its gaze randomly (Sorostinean et al., 2014).

In order for people to make use of and react to a robot's social gaze, they must first perceive it. In a previous study measuring the effects of a robot's gaze, a number of human participants and a single robot were placed in a circular arrangement (Imai et al., 2002). Participants were shown to notice when the robot was gazing at them. Although a seemingly simple finding, it serves as a basis for further work on the effects of gaze in embodied agents. However, it should be noted that people noticed the robot's gaze when it looked at or near them, but not when it gazed at somebody else nearby (Imai et al., 2002). Further research has revealed that people have stronger feelings of being looked at when a robot gazes toward them using short, frequent glances rather than longer, less frequent fixations (Admoni et al., 2013).

People are also sensitive to robot gaze when that gaze is directed toward objects or locations in the environment. For example, in object selection games, people can use

referential gaze cues to make predictions about which objects to select, even when they are not consciously aware of those cues (Mutlu et al., 2009b). With a back-projected robot head with animated gaze and facial expressions, people can predict the target location of the robot's gaze almost as accurately as that of a human's gaze, although this accuracy suffers when the robot is viewed from the side or when its gaze only involves head movements with static eyes (Al Moubayed and Skantze, 2012).

Achieving Positive Interaction Outcomes

Previous research has shown that robots can greatly benefit from the utilization of human-like gaze mechanisms (Emery, 2000; Liu et al., 2012; Mutlu et al., 2006, 2012; Staudte and Crocker, 2009). Research in human-robot interaction (HRI) has investigated the positive outcomes achievable through a robot's gaze behavior, such as improving the robot's competence in conversational interactions with people (Mutlu et al., 2009a, 2012), enabling joint attention and referential communication (Huang and Mutlu, 2012; Staudte and Crocker, 2011), and increasing the robot's ability to hand objects to people (Moon et al., 2014). In a user study involving a map route drawing task, Skantze et al. (2013) showed that appropriate robot gaze behaviors can improve task performance and reduce cognitive load by helping to disambiguate referring expressions to objects in a shared scene and manage the flow of interaction.

Modeling humanlike gaze mechanisms enables conversational robots to signal different participant roles to human interlocutors, manage turn-exchanges, and shape how users perceive the robot and the conversation (Mutlu et al., 2012). Human comprehension of robot speech is improved when a robot gazes to objects it is speaking about in a similar way and with similar timings as found in human gaze behavior (Staudte and Crocker, 2009; Huang and Mutlu, 2012). Robots that use gaze cues are perceived as possessing mental states and intentionality, as evidenced by previous work on gaze leakage in a human-robot guessing game (Mutlu et al., 2012). A positive learning effect of gaze has also been observed in HRI; increased gaze from a storytelling robot facilitates greater recall of story events (Mutlu et al., 2006).

Robots that engage in mutual gaze with users have been shown to improve people's subjective evaluations of them. For example, mutual gaze from a stuffed animal companion robot leads to favorable evaluations of it (Yonezawa et al., 2007). When a robot is learning from human demonstration, displaying mutual gaze influences people to view the robot as more intentional than displays of random gaze; people spend more time teaching the robot, pay more attention to it, and speak more with it (Ito et al., 2004). Robot tour guides can

influence people's experience of a tour by how often they direct gaze toward each listener (Karreman et al., 2015). When a robot favors a particular person by gazing at them longer than other members of the group, that person reports greater feelings of likability towards the robot (Karreman et al., 2013).

Gaze in Situated Interaction

To be truly effective, a robot's gaze mechanisms need to be employed contingently based on the behaviors and unique characteristics of its human interlocutors. A robot that gazes responsively toward a human user is capable of eliciting a stronger feeling of being looked at than a robot that uses non-responsive—i.e., static or random—gaze (Yoshikawa et al., 2006). HRI researchers have developed a responsive, rule-based system for generating robot head nods, tilts, and gazes based on discrete dialogue acts in conversation, such as those associated with turn-taking, backchannels, and conversational fillers (Liu et al., 2012).

The detection of user gaze and attention has also been used to inform and trigger interactive robot behaviors. For example, Das et al. (2014) classify human gaze motions as either spontaneous or scene-relevant. The robot interrupts the user to explain aspects of the scene only if it detects that they are looking at something of scene relevance. This work illustrates the need for employing conversational gaze mechanisms responsively to the behaviors of the robot's human interlocutor.

Unlike virtual agents, robots may also need to use gaze to obtain information if its cameras are embedded in its eyes. For example, in conversations with multiple participants, previous research has explored how a robot's scanning behavior can serve dual functions of indicating attention to all conversational partners while simultaneously updating the robot's knowledge of the partners that are occasionally outside the range of its cameras (Bennewitz et al., 2005).

An important part of collaboration involves referencing task-relevant objects in the environment. Joint attention from a companion robot has been shown to effectively draw a user's attention to where the robot is looking (Yonezawa et al., 2007). Gaze can also act as a reinforcement of deictic pointing gestures (Sauppé and Mutlu, 2014). A robot can use gaze to support its speech in a cooperative object selection task, in which a human user must quickly select an object referenced by the robot (Admoni et al., 2014; Boucher et al., 2012). People can recognize and respond to predictive gaze that indicates spatial references, completing the task faster than if they had been relying on the robot's speech alone.

Tour guide robots can also use deictic gaze to have a positive effect on listener engagement and attention. When a robot displays deictic gaze that reflects the subject of its speech,

people display more nodding and mutual gaze, signaling increased engagement, than they do when the robot's deictic gaze occurs at random points in its speech (Kuno et al., 2007; Yamazaki et al., 2010). When a robot uses deictic gaze in addition to verbalized object references, people become more engaged, spend more time interacting with the robot, and display more coordinated gaze behaviors than when the robot speaks without supportive gaze (Sidner et al., 2004). A robot that orients its body and accompanying gaze toward an exhibit can more easily engage its listeners than a robot that orients toward the audience, but people lose interest in the robot and its narrative more often when the robot looks only toward the exhibit and not at its audience (Karreman et al., 2015). This suggests that agents must distribute their gaze between listeners and informationally interesting objects, and how this might be done to achieve specific effects is explored in Chapter 3.

Both detecting and producing gaze cues can improve the efficiency of robot-to-human handovers. For example, handovers are improved when a robot monitors its partner's eye gaze for attention and engagement, only releasing the object when the user's focus of attention has turned to that object (Grigore et al., 2013). In multiparty scenarios, robots can also use gaze to nonverbally select a member of the crowd to hand an object to (Kirchner et al., 2011). During a handover, people begin reaching for an object earlier when a robot continuously looks at the projected position in space where the handover will occur, than when it looks away from that location (Moon et al., 2014). People reach for the object even earlier when the robot continually gazes at their faces than when it looks at the handover location (Zheng et al., 2015). Gazes that transition between the user's face and handover location do not objectively improve the quickness with which people begin reaching, although people subjectively report that these gazes communicate the handover timing more effectively than continuous gazes (Zheng et al., 2015).

Existing Gaze Models and Systems

In multiparty conversations, robot gaze can influence people to take on certain conversational roles. Several studies have found that a robot can use gaze behaviors to manipulate certain members of a group into taking conversational roles such as bystander, active participant, or listener (Kirchner et al., 2011; Mutlu et al., 2009a, 2012). A robot's gaze behaviors are successful at influencing people to conform to the intended roles as much as 97% of the time (Mutlu et al., 2012).

Robots have been envisioned in a number of socially assistive roles (see also Chapter 6). For socially assistive robots that act as therapy assistants to children with autism spectrum disorder (ASD), gaze can be a particularly important cue because of the difficulties in social

gaze perception and detection that are often typical in people with this disorder (Scassellati et al., 2012). Some children with ASD show spontaneous social gaze behaviors in response to robots, including increased eye gaze and shared attention during robot interactions as compared to human interactions (Tapus et al., 2012). However, there is large variability in responses, and other children do not necessarily demonstrate the same increase in gaze behavior (Tapus et al., 2012).

The context of conversation also influences what kind of gaze works best for robots. In conversations about emotionally neutral topics, robots that make eye contact are seen as more sociable and intelligent than robots that avoid it, but this effect is reversed when the topic of conversation is embarrassing, with eye contact avoidance rated more highly (Choi et al., 2013). In persuasive communication, natural gaze behaviors improve a robot's persuasiveness (Ham et al., 2015), even more than using expressive vocalizations (Chidambaram et al., 2012). Gaze also interacts with other nonverbal behaviors for persuasiveness. Persuasive gestures improve a robot's overall persuasiveness only when accompanied by natural eye gaze, and have the opposite effect when performed without eye gaze, actually hindering the robot's persuasiveness (Ham et al., 2015).

Gaze behavior can be used to express recognizable personalities and emotions. Mutual gaze can express feelings of trust, while inappropriate levels of gaze aversion can express feelings of distrust (Normoyle et al., 2013). Specific manipulation of features of gaze—such as amount of gaze, duration of gaze, and the points of fixation during gaze aversion—yields consistent impressions of dominance and friendliness in a robot (Fukayama et al., 2002).

Revealing mental states through gaze can make cooperative task performance faster, with errors detected more quickly and handled more effectively than purely task-based nonverbal communication (Breazeal et al., 2005). Indicating engagement and providing feedback through subtle gaze behaviors improves performance of a human-robot team (Jung et al., 2013). Users also report understanding the robot better during their collaboration when it makes its mental models explicit (Breazeal et al., 2005). Expressive gaze behavior derived from animation principles can make intentions and desires more explicit, for example by looking at a door handle when wanting to open a door (Takayama et al., 2011). Even when users are unaware of the intended communication, robots can leak their intentions through gaze, influencing human behavior in measurable ways (Mutlu et al., 2009b).

Summary

Just as for virtual agents, a number of gaze models and systems have been proposed to help make robots more effective communicators. Gaze has been utilized to improve learning, enable joint attention, improve task-based referencing, express conversational roles, improve handovers, convey mental states, and more. While previous research has explored how both virtual agents and robots can use gaze to achieve positive social outcomes, a precise account of how and when agents should deploy gaze mechanisms contingently on a number of social variables and what positive outcomes these mechanisms might achieve is still needed. This dissertation seeks to address this knowledge gap from both design-driven and empirical perspectives through the application of existing social-scientific knowledge and a study of human dyadic conversations to design gaze behaviors for embodied agents. The next section compares virtual to physical embodiments, elucidating the important differences that must be accounted for when designing behaviors that target both types of agents.

2.4 Virtual vs Physical Embodiments

Robots and virtual agents differ along a number of social dimensions, including realism, social presence, lifelikeness, and physical proximity (Powers et al., 2007). Several studies have demonstrated effects of these differences (Bainbridge et al., 2011; Kidd and Breazeal, 2004; Powers et al., 2007).

There is some disagreement in the human-robot interaction (HRI) and intelligent virtual agent (IVA) communities on whether and how physically embodied systems elicit different responses than animated agents or identical video representations of the same physical systems. Some researchers have found that physically embodied systems improve interactions over virtual systems. Children spend more time looking at a robot tutor that is physically embodied than at a virtual representation of that robot (Kennedy et al., 2015), and adults retain lessons about a cognitive puzzle better when they've been tutored by a physically embodied robot than by a video representation of that robot (Leyzberg et al., 2012). People also fulfill unusual requests from a robot more frequently when that robot is physically embodied than when it is displayed on a video monitor (Bainbridge et al., 2011). Physically embodied agents have been shown to be rated more positively (Powers et al., 2007; Wainer et al., 2007) and attributed greater social presence (Lee et al., 2006a) than their virtual counterparts.

One study which compared physical robots with virtual agents within a specific in-

teraction context found that participants liked the robot more than the virtual agent, but disclosed less and remembered fewer key pieces of information when interacting with the robot versus the virtual agent (Powers et al., 2007). Participants in a block-stacking task found interaction with a robot to be more engaging, enjoyable, informative, and credible than interaction with a virtual agent (Kidd and Breazeal, 2004). In a collaborative book-moving task, participants were more likely to fulfill a trust-related request from a physically co-located robot than from a robot presented on a video display, and overall had more positive interaction with the physically present robot (Bainbridge et al., 2011).

Not all research has supported the benefit of physical embodiment over virtual presence. In a tutoring interaction involving sorting, children did not exhibit any differences in learning between physical and virtual robot tutors (Kennedy et al., 2015). In an interaction with a healthcare robot, people remembered less information provided by a physically co-located robot than information provided by a virtual representation of that robot (Powers et al., 2007). Virtual agents can also allow for fine control over the appearance and timing of gaze behaviors, such as subtle eyelid, eyebrow, and eye ball movements. These fine-grained movements are difficult to achieve with the physical motors on robots. Though some hyper-realistic humanoid robots—such as Geminoid (Sakamoto et al., 2007) and FACE (Zaraki et al., 2014)—strive to achieve human-like face actuation, most robots do not achieve the level of facial expressiveness available in animated characters. Virtual agents are thus a more promising platform for studying the effects of finely controlled, subtly expressive motions of social eye gaze.

Virtual agents and social robots generally differ in their physical affordances available for carrying out gaze motions. Overall differences in geometry mean that gaze motions and control laws need to be adapted (Pejsa et al., 2013). Even more critically, robots often lack articulated eyes altogether and must rely solely on head motions to convey gaze motions, making it unclear if they are capable of eliciting the same positive conversational outcomes found in the virtual agent work. Previous research points to the possibility of such capabilities, including work which has shown that people are capable of recognizing a robot's gaze according to its head orientation (Imai et al., 2002) and that robots can use head motions alone to gaze effectively in a storytelling scenario (Mutlu et al., 2006).

2.5 Chapter Summary

This chapter provided background on social gaze behavior from research on human communication, and related work on the design of social gaze behavior for virtual agents and humanlike robots. The next four chapters will describe the design, implementation, and

evaluation of four situated gaze mechanisms useful for both virtual agents and robots, including gaze shifts, gaze aversions, gaze coordination, and adaptive gaze. These gaze mechanisms are contextualized in scenarios of education, conversation, collaboration, and rehabilitation. The next chapter focuses on the most fundamental of these mechanisms, presenting a model of how agents and robots should carry out overt shifts in gaze in a humanlike and communicative way.

3 GAZE SHIFTS

An agent's overall gaze behavior can be characterized as a sequence of *gaze shifts*—the intentional redirection of gaze toward a particular target in the context of interaction. While the gaze shift is a seemingly simple element, it is produced and interpreted in complex ways. Gaze shifts are a fundamental building block of human gaze behavior. Through subtle variation in timing and movement of the head and eyes in shifting the gaze, individuals construct a range of complex communicative behaviors. When animating an embodied agent, control mechanisms must synthesize the wider range of such movements so that the agent displays natural communicative behaviors, yet provide sufficient control over the subtleties of the movements to allow for individual variation and expressions. Creating fine-granulated control mechanisms for gaze that achieve a combination of communicative effectiveness, naturalness, and parametric control remains an open challenge.

This chapter presents the modeling and evaluation of gaze shifts for embodied agents to employ when speaking to human users. This model is parameterized to generate gaze shifts contingently on the characteristics of the agent itself and the goals of the interaction, such as to engender feelings of affiliation or to more effectively convey lecture-style information. These parameters include, e.g., *head alignment*, defined as the degree to which the head is aligned with the gaze target at the end of the gaze shift, and *head latency*, which is the time delay from the beginning of the gaze shift before the head starts moving. This model of gaze shifts serves as the basis for the construction of all higher-level gaze behaviors presented in later chapters.

The model presented in this chapter has a number of key properties: (1) it is grounded in research on human physiology, (2) it provides parameters that allow for generating a range of gaze behaviors, and (3) it has been validated empirically, confirming its effectiveness in generating gaze shifts that communicate gaze direction and appear natural and realistic. The testing also shows that this model outperforms the state-of-the-art model in achieving naturalness and realism.

This chapter also demonstrates how virtual agents can use subtle variations in gaze to achieve significant high-level outcomes. A user study with 20 participants was conducted to investigate how a virtual agent might manipulate low-level gaze parameters to achieve high-level effects such as improved learning and feelings of rapport in a storytelling context. The head alignment parameter of the model was manipulated to create *affiliative gaze*—maintaining a head orientation toward the participant to emphasize the social interaction (Frischen et al., 2007)— and *referential gaze*—maintaining a head orientation toward shared visual space to emphasize the information to which the speaker is referring (Langton and

Bruce, 1999). In both cases, the agent looks at the same targets; the difference lies in the timing and the degree to which the agent uses its head or eyes to look at these targets. The results showed that these manipulations generated significant social and cognitive effects; the use of affiliative gaze improved the perceptions of the agent and the use of referential gaze increased recall performance.

The remainder of this chapter provides related work on gaze from a number of perspectives, describes the model and an online validation study, outlines the design and results of the in-person user study evaluation, and discusses the main findings and their implications for the design of effective gaze mechanisms for embodied agents.[1]

Research Questions

- How do people coordinate their head and eyes to carry out shifts in gaze?
- How can we enable embodied agents to carry out humanlike gaze shifts while maintaining control and flexibility for designers?
- What are the positive interaction outcomes of an agent utilizing humanlike gaze shifting, and how do these effects differ when the agent carries out gaze shifts in different ways?

3.1 Related Work

This section reviews existing models for gaze synthesis from research in computer graphics and human-computer interaction (HCI), as well as the neurophysiological research that informs the model of gaze shifts.

Models for Gaze Synthesis

Numerous gaze models have been proposed in the literature, each with different methods, goals, and contributions. For example, data-driven models (Deng et al., 2005; Heck, 2007) are a common and powerful approach. However, it is often difficult to manipulate parameters or incorporate known constraints in these models without providing new hard-to-find or hard-to-create examples. The Expressive Gaze Model (EGM) takes a hybrid

[1] Portions of this chapter were published in Andrist et al. (2012a) and Andrist et al. (2012b). I would like to acknowledge Tomislav Pejsa, a graduate student collaborator, for his contributions in developing the virtual agent framework and implementing the gaze shift model described in this chapter.

data-diven/procedural approach to the animation of gaze shifts and focuses on communicating emotion by executing gaze shifts in different ways (Lance and Marsella, 2010a). A library of Gaze Warping Transformations (GWTs) comprises the data-driven part of the model and converts a baseline neutral gaze shift (only the head's contribution) into an expressive one using motion capture data. The movements of the eye are procedurally generated separately from those of the head, which relies on simple heuristics and inverse kinematics to compute a highly stereotyped motion. Because this approach handles the head and eyes separately, it does not cover the complexities of eye-head synchronization during gaze shifts as the model presented below does.

Another set of gaze models considers *where* the agent should be looking (Khullar and Badler, 2001; Itti et al., 2006), for instance, by simulating biological visual processing (Itti et al., 2006). Other models seek to explain *why* the agent might be looking toward particular targets, considering gaze signals according to their functional meaning rather than their physical actions (Lee et al., 2007; Poggi et al., 2000; Fukayama et al., 2002). As these models employ simple mechanisms to realize gaze movements, they complement the model in this chapter determining how the head and the eyes work together to produce gaze shifts.

An important consideration in animating natural gaze behavior in agents is *idle gaze* when the agent is not actively performing gaze shifts, e.g., for agents idling in a populated public virtual setting (Cafaro et al., 2009). Idle gaze is particularly important in face-to-face conversations. Lee et al. (2002) have developed a stochastic model of random eye saccades based on whether the agent is talking or listening and maintaining mutual gaze with or gazing away from the user. While the work presented in this chapter focuses on directed gaze shifts, the Lee et al. (2002) model was included as part of the final implementation to achieve natural gaze behavior during idle periods.

Neurophysiological Models of Gaze

Neurophysiologists have studied how humans and other primates coordinate head and eye movements during gaze shifts, in the process revealing potential parameters of gaze that might eventually lead to the positive social outcomes designers wish to create with embodied agents.

Research in neurophysiology has studied how humans carry out gaze shifts by coordinating head and eye movements in a tightly connected dynamic process. In most existing models of directed gaze shifts, kinematics of saccadic movements, such as duration and peak velocity, are simplified, producing movements that depend only on the amplitude and direction of directed eye movements towards the target. In reality, concurrent head

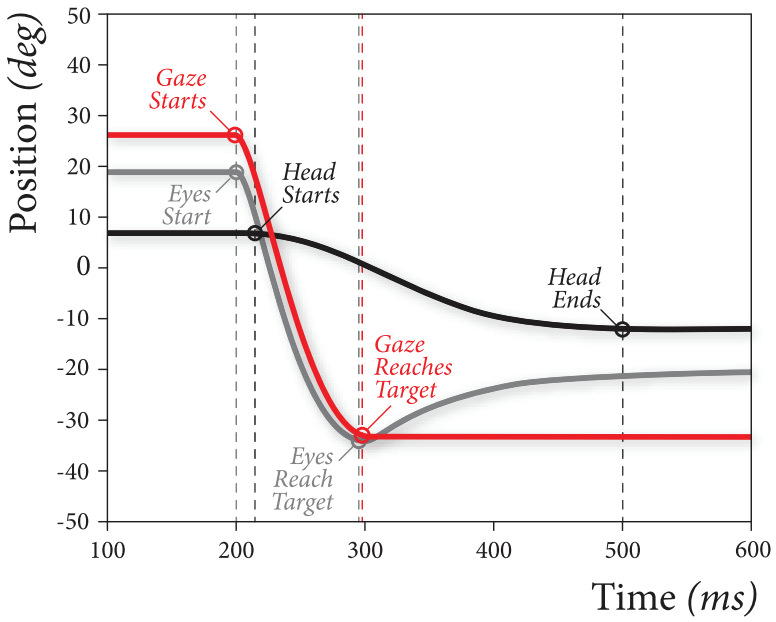

Figure 3.1: Eye, head, and overall gaze trajectories through the archetypal gaze shift (adapted from Freedman and Sparks (2000)).

movements affect eye movements significantly (Freedman and Sparks, 2000). This causal relationship holds in both directions; for example, head movement amplitude decreases as the eyes movements start at increasingly contralateral positions (i.e., oriented away from the target in relation to head direction). Shifts that start at such positions require the eyes to increase their contribution to the shift (McCluskey and Cullen, 2007).

Whether the gaze target is auditory or visual affects how individuals orient their attention toward a target. Studies that compare auditory and visual targets show that eyes tend to lead the head most often when people orient to visual targets, whereas the head tends to lead the eyes most often when people orient to auditory targets (Goldring et al., 1996; Goossens and Opstal, 1997). This finding has been attributed to the differences in initial reference frames used by humans to localize visual and auditory stimuli; visual stimuli are initially localized using eye-centered or retinotopic coordinates, while auditory stimuli are localized in head-centered or crainotopic coordinates.

The degree to which individuals use their heads in performing a gaze shift is highly idiosyncratic. The neurophysiological research literature describe some people as "head-movers," i.e., individuals who move their head fully to align with the gaze target every time, and some as "non-head-movers" (Fuller, 1992). From a bio-mechanical standpoint, humans should universally be "non-head-movers," as fully moving the head—which is almost a hundred times heavier than the eyes—is not an economical solution (Kim et al.,

2007). This neurally based idiosyncratic characteristic of head and eye movements is not captured by naïve inverse kinematics solutions for animating gaze shifts.

While gaze shifts are formed and affected by numerous factors, a certain type of gaze shift that has been found to be the most common (Freedman and Sparks, 2000). This archetypal gaze shift starts with the initiation of an eye saccade, followed very shortly by the onset of head movement. Both the eyes and head then move until the eyes reach the target. When the eyes converge on the target, the vestibulo-ocular reflex moves the eyes in the opposite direction relative to head motion to keep the eyes locked with the target. Figure 3.1 provides a graphical representation of this archetypal gaze shift.

3.2 Modeling Gaze Shifts

In the model presented here, the parameters of a specific gaze shift (e.g., the target direction) and parameters of the character (e.g., maximum head velocity) are used to compute a number of internal timing parameters at the onset of the gaze movement. Table 3.1 and Table 3.2 list all input and internal parameters included in the model respectively. Once the internal parameters are computed, the gaze shift begins, and the eyes and head are rotated directly towards the target at dynamically-changing angular velocities. The velocities are recomputed in each frame of animation based on the progress of the gaze shift and the current rotations of the eyes and the head, allowing the model to react to perturbations of the head position or target during motion. The model calculates the rotations for the eyes independently in order to achieve convergence.

Given the current configuration of the head and eyes and input parameters indicating movement characteristics and target gaze direction, this model computes a trajectory for the head and eyes. It first computes a few key variables for the movement, and then computes velocities for the head and eye rotations for each frame of the animated movement, keeping the gaze on target until the eyes and head have reached their final target rotations. The model is presented graphically in Figure 3.2, and includes six main components: (A) head latency, (B) velocity profiles for head and eye motion, (C) oculomotor range (OMR) specifications, (D) head alignment preferences, and (E) the vestibulo-ocular reflex (VOR).

Internal Parameter Computation

The first phase in generating gaze shifts is to determine the latency of the onset of head movement, $\mathtt{hl}$, in relation to the onset of the eye movement (Figure 3.2A). This head latency can vary from person to person and task to task. Factors such as the vigilance of the agent

Table 3.1: All input parameters to the gaze shift model

Parameter	Symbol	Description
Gaze Target Position	GT_{pos}	The destination of the gaze shift.
Target Predictability	GT_{pred}	A value between 0 and 1, where 0 and 1 correspond to unpredictable and predictable targets, respectively. Could be manually manipulated or driven by the agent's cognitive system.
Target Saliency	GT_{sal}	A value between 0 and 1, where 0 and 1 correspond to low and high saliency, respectively. Could be manually manipulated or driven by the agent's vision system.
Target Modality	GT_{mod}	A binary value that indicates whether the target is visual or auditory.
Agent Vigilance	AG_{vig}	A value between 0 and 1, where 0 and 1 correspond to low and high vigilance, respectively. Could be manually manipulated or driven by the agent's cognitive system.
Agent Intent	AG_{int}	A value between 0 and 1, where 0 and 1 correspond to natural and forced gaze shifts, respectively. Could be manually manipulated or driven by the agent's cognitive system.
Oculomotor Range	OMR	The maximum pitch and yaw of the agent's eyes in degrees.
Head Alignment	h_{align}	A value between 0% and 100%, where 0% corresponds to minimum head alignment and 100% corresponds to full head alignment.
Initial Eye Position	IEP	Initial gaze configuration of the eyes and head. Presented as a rotational offset of the current eye rotation from a central (in-head) orientation. Employed only when the rotational offset is contralateral (on the opposite side of center) to the target.

(AG_{vig}), the target salience (GT_{sal}), the eccentricity of the target ($GAmp$), the predictability of the target location (GT_{pred}), the modality of the target—visual or auditory—(GT_{mod}), and the intent of the agent—forced or natural shift—(AG_{int}) have an effect (Pelz et al., 2001; Goldring et al., 1996; Goossens and Opstal, 1997; Zangemeister and Stark, 1982). These factors can be stochastically combined to determine the gaze shift's head latency based on findings from (Zangemeister and Stark, 1982).

Each of the factors is associated with a different likelihood ratio of leading to a head-first versus eyes-first gaze shift. A ratio value r is defined as $\frac{P_h}{1-P_h}$, where P_h is the probability of a head-first gaze shift. These ratios are summarized in Table 3.3. For example, gaze shifts with large amplitudes (greater than 30°) are 3.05 times more likely to involve a head-first approach. When considered in isolation from other parameters, auditory targets always

Table 3.2: All internal parameters of the gaze shift model

Parameter	Symbol	Description
Gaze Amplitude	$GAmp$	The rotational difference between the current and the final gaze alignment. Determined from the gaze target.
Head Latency	hl	A value that varies between -100 ms and 100 ms. Stochastically determined from the above inputs.
Effective OMR	OMR_{eff}	Neurally determined limits on eye motion. Updated throughout gaze shift.
Maximum Eye Velocity	EV_{max}	Computed based on a positive linear relationship with the amplitude of the intended gaze shift, which saturates at about 500°/sec for gaze shifts at or beyond the OMR of the character (Guitton and Volle, 1987).
Eye Velocity	EV	Follows the velocity profile described in the text.
Maximum Head Velocity	HV_{max}	Computed based on a positive linear relationship with the amplitude of the intended gaze shift (Guitton and Volle, 1987). For example, a gaze shift of 24° will result in a maximum head velocity of 50°/sec in our implementation.
Head Velocity	HV	Follows the velocity profile described in the text.

produce head-first gaze shifts, hence the ratio of ∞ (Zangemeister and Stark, 1982). A positive latency results in a period of eye motion during which the head remains fixed, while a negative latency result in a period of head motion while the eyes remain fixed.

Rewriting for P_h in the ratio definition above, we get

$$P_h = \frac{r}{r+1}.$$

We can compute the final probability of a head-first gaze shift by linearly interpolating between the probability of a head-first gaze shift (P_h) and an eyes-first gaze shift ($1 - P_h$). Probabilities for each factor are sampled to determine whether each suggests a head-first or an eyes-first movement. These various "votes" on the type of movement are combined to determine the head latency. If f is the number of factors that vote to determine a head-first gaze shift and n is the total number of parameters, then the ratio s can be computed as $\frac{f}{n}$. We can use s to compute the head latency (hl) for the gaze shift by linearly interpolating between the head latency of a purely head-first gaze shift, hl_h, and the head latency of a purely eyes-first gaze shift, hl_e. This model utilizes values of -100ms for hl_h and 100ms for hl_e based on the range of values proposed in the neurophysiology literature (Zangemeister and Stark, 1982).

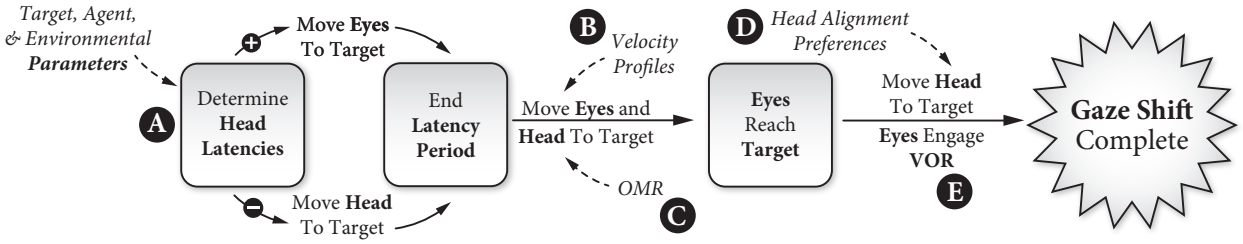

Figure 3.2: A visual representation of the gaze shift model. Key input variables and processes include (A) head latency, (B) velocity profiles for head and eye motion, (C) oculomotor range (OMR) specifications, (D) head alignment preferences, and (E) the vestibulo-ocular reflex (VOR).

Generating Gaze Motion

Once the model determines the head latency, it initiates the movements of the eyes and/or the head. Each eye and the head move towards the target, following velocity profiles that resemble standard ease-in and ease-out functions (Figure 3.2B). These movements prevent unnatural head motions caused by high-frequency signals (Ma et al., 2011). The maximum velocity of the eyes and head, EV_{max} and HV_{max}, are computed based on positive linear relationships with the amplitude of the intended gaze shift (Guitton and Volle, 1987). For example, HV_{max} is 50°/sec for gaze shifts of 24°, and 100°/sec for gaze shifts of 54° in the current implementation. Similarly, EV_{max} increases from a baseline value of 100°/sec to a saturation value of 500°/sec for gaze shifts at or beyond OMR. A piecewise polynomial function was derived to approximate the full velocity profile for both the head and the eyes determined from the literature (Lee et al., 2002; Kim et al., 2007). This polynomial function can be expressed as follows, where g is the proportion of the gaze shift completed, V_{max} is the maximum velocity, and V is the current calculated velocity.

Table 3.3: Ratio of *likeliness of head-first shift / likeliness of eyes-first shift*

Parameter	**Ratio Meaning**	**Ratio Value**
Amplitude	$\frac{large}{small}$	3.05
Intent	$\frac{forced}{natural}$	2.20
Predictability	$\frac{high}{low}$	1.6
Vigilance	$\frac{high}{low}$	1.33
Target Saliency	$\frac{high}{low}$	0.53
Target Modality	$\frac{auditory}{visual}$	∞

$$V = \begin{cases} 2V_{max} \cdot g & g \in [0, 0.5) \\ 4V_{max} \cdot g^2 - 8V_{max} \cdot g + 4V_{max} & g \in [0.5, 1] \end{cases}$$

Eyes are handled independently in order to achieve convergence. Gaze shifts toward each target are continually recomputed, allowing the model to react to perturbations of the head position or target during the motion.

Humans are mechanically limited in their ability to rotate their eyes, a limitation referred to as the oculomotor range (OMR) (Figure 3.2C). The human OMR has been estimated to be between 45° and 55°. An embodied agent's baseline OMR can be empirically determined based on the size of its eye cavities and the size of its pupils and irises (if they exist). However, merely encoding these OMR values as static parameters is not sufficient, as the effective OMR may fluctuate during the course of a single gaze shift. The fluctuation is a product of the neural (as opposed to mechanical) nature of the limitation imposed on eye motion.

At the onset of a gaze shift, OMR_{eff} is computed based on the initial eye position (IEP) and the OMR. IEP is measured in degrees as a rotational offset of the current eye orientation from a central (in-head) orientation. This value is only non-zero when the rotational offset is contralateral (on the opposite side of center) to the target. When the eyes begin the gaze shift at these angles, the OMR_{eff} has a value close to the original baseline OMR. When the eyes begin the gaze shift closer to a central orientation in the head, the OMR_{eff} diminishes (Freedman and Sparks, 1997). This relationship is approximated in the model with the following function:

$$OMR_{eff} = OMR \cdot (\frac{1}{360} IEP + 0.75).$$

The agent's eye rotations are not allowed to surpass the current effective OMR, OMR_{eff}, which is updated throughout the gaze shift at every time step according to the concurrent instantaneous head velocity, HV. As the head moves faster, the OMR_{eff} diminishes (Guitton and Volle, 1987). This relationship was approximated with the following function, where OMR_{IEP} is the value for OMR_{eff} that was computed in the previous equation at the onset of the gaze shift.

$$OMR_{eff} = OMR_{IEP} \cdot (\frac{-1}{600} HV + 1)$$

The next component of the model is a user-defined parameter specifying the *head alignment* preferences of the agent (Figure 3.2D). Head alignment (h_{align})—how much individuals use their heads in performing a gaze shift—is highly idiosyncratic, and creates

a differential directness of gaze at the target (Fuller, 1992). A parameter value of 0% for head alignment indicates that once the eyes have reached the gaze target, the head stops moving, resulting in gaze at the target out of the corner of the eyes. On the other hand, at a 100% parameter value for head alignment, the head continues rotating until it is completely aligned with the target, with concomitant compensatory eye movement to keep the eyes directed toward the target, resulting in gaze with the eyes and head both fully directed towards the target. Head alignment values between these two extremes can be computed using spherical linear interpolation between the two corresponding rotational values. This parameter can be manipulated to create affiliative gaze–keeping high head alignment with a conversational partner and low head alignment with everything else–or referential gaze–keeping high head alignment with information being referred to in the environment and low head alignment with the conversational partner. Head alignment is the most directly controllable input parameter to the model, and as such is the parameter focused on in the experiment presented at the end of this chapter.

Aligning the head with a target takes into account the translational offset from the head joint origin to a point between the eyes. Thus, the alignment occurs for a ray pointing straight out of the head from this point, not from the head joint itself (which may be arbitrarily defined for different character models). The model assumes that the eyes always reach the target first, which is a valid assumption for humanlike gaze given the relative rotational speeds of the eyes and the head.

Once the eyes have reached their target, the vestibulo-ocular reflex (VOR) keeps them locked to the target while the head finishes its rotation (Figure 3.2E). This component of gaze, the final component of the model, is handled by rotating the eyes in the opposite direction of head motion as the head completes its portion of the gaze shift, keeping the eyes fixated on the gaze target even as the head keeps rotating.

Ancillary Components of the Model

This gaze model also includes a blink controller that serves two key functions: generating gaze-evoked blinking as described by Peters (Peters, 2010) and idle blink behavior. The virtual character's eyelids also move with vertical shifts of the eyes, rising up as the eyes pitch upwards and dropping down as the eyes pitch downward (Steptoe and Steed, 2008).

When the agent is not actively engaging in a gaze shift following the model, the eyes are controlled by an implementation of the model presented by Lee et al. (2002). This implementation of subtle random eye movements has been shown to dramatically increase the realism of the character.

Finally, the completed implementation also includes ambient facial and bodily cues for the agents to prevent the unnatural rigidity inherent to a static character. When the character is in an idle state, these cues include smiles and slight movements of the arms and the head.

3.3 Model Validation

Development of the model was followed by an empirical evaluation of the communicative accuracy and perceived naturalness of gaze shifts that the model generates. Gaze shifts generated by the above model were compared against those generated by a baseline model and those displayed by a human. In addition, this validation explored the effect that participant gender and the gender of the virtual character might have on user perceptions of the gaze shifts, as gender has been previously shown to shape gaze perception in artificial agents (Mutlu et al., 2006).

Participants

Ninety-six participants (46 females and 50 males) took part in the validation study. The participants were recruited through Amazon's Mechanical Turk online marketplace, following crowd-sourcing best practices to minimize the risk of abuse, achieve a wide range of demographic representation, and mitigate issues from differences in participants' viewing parameters, such as screen resolution (Kittur et al., 2008; Ipeirotis, 2010). Only users who are residents of the U.S. with an approval rating of 95% or greater were allowed to participate. IP-number-based filtering techniques ensured that the participants could not perform the experiment more than once. Participants received $2.50.

Experimental Setup and Task

Each participant was shown 32 videos of a virtual character or human. In the videos, the agents or the confederates gazed toward the participant, announced that they are about to look toward an object with a specific color on the table, shifted their gaze toward the object, and moved their gaze back toward the participant. This simplified scenario allowed for focusing the evaluation on the effectiveness and naturalness of gaze shifts, while minimizing contextual and interactional factors and facilitating the matching of animated and real world conditions. Participants observed the agent from the perspective of a collaborator seated across from the agent or the human confederate. The objects on the desk were separated into four groups, distinguished by color and shape. Still images

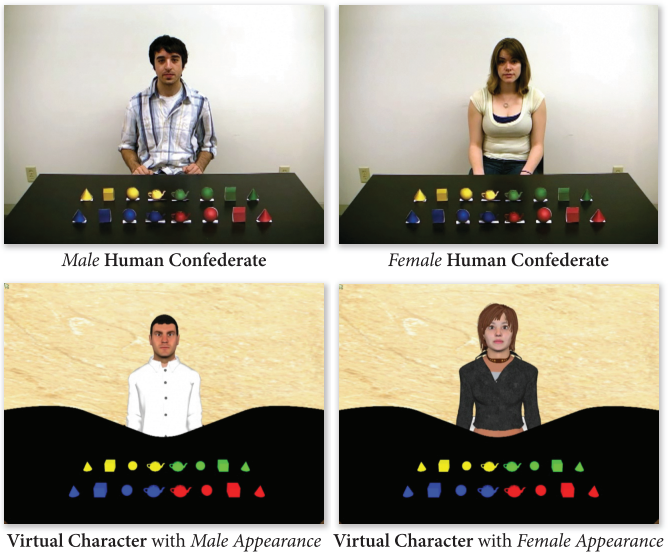

Figure 3.3: Still images from the videos presented to the participants.

from the videos are shown in Figure 3.3. Following each video, the participants filled out a questionnaire for subjective evaluation. Participants observed each gaze model generating gaze shifts toward all object types, colors, and positions. Each video was 10 seconds long, with the overall study lasting approximately 20 minutes.

Study Design

The study followed a two-by-four split-plot design. The factors were gender—participant matched with agent—varying between participants, and model type, varying within participants. The model type independent variable included gaze shifts generated by a baseline gaze model from Peters (2010), and those produced by the model of this chapter. The model type independent variable also included two control conditions. In the first control condition, a virtual agent with a male or female appearance maintained gaze toward the participant (i.e., the camera) without producing any gaze shifts at all. In the second control condition, a male or female human confederate presented gaze shifts toward the object on a desk in front of him/her. The order in which the conditions were presented to each participant was counterbalanced.

Measures

The study involved two dependent variables: *communicative accuracy* and *perceived naturalness*. Communicative accuracy was measured by capturing whether participants correctly

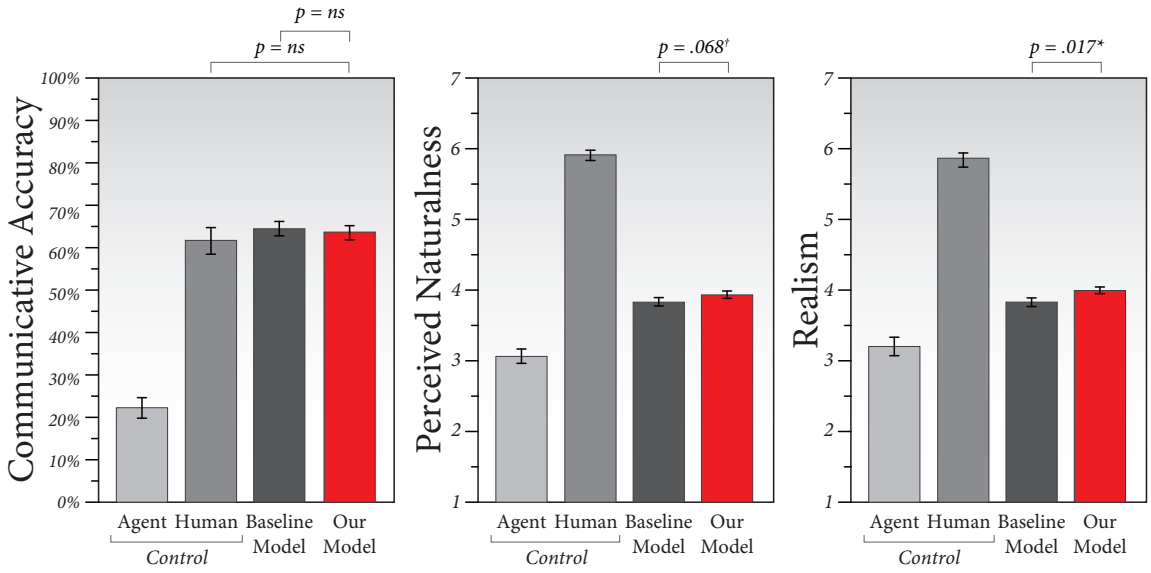

Figure 3.4: Results from the communicative accuracy, perceived naturalness, and realism measures. The baseline model refers to a previously published gaze model (Peters, 2010) used for comparison.

identified the object toward which the gaze shift of the human confederates or the virtual characters was directed. To measure perceived naturalness, a scale was constructed using four items that measured *naturalness*, *humanlikeness*, *lifelikeness*, and *realism*. These items were highly correlated (Cronbach's $\alpha = .94$). Participants rated each video for each item using a seven-point rating scale.

Results

A mixed-model analysis of variance (ANOVA) was conducted to determine the effect that different gaze models had on how accurately participants identified the object that the agents or the human confederates looked toward and the perceived naturalness of the gaze shifts. Overall, model type had a significant effect on communicative accuracy, $F(7, 1958) = 16.77, p < .001$, and perceived naturalness, $F(7, 1958) = 151.03, p < .001$. Detailed comparisons across conditions for each factor are described in the next paragraphs. The experiment-wise error rate was controlled by using Tukey-Kramer HSD in all post-hoc tests.

Communicative Accuracy – Pairwise comparisons found no significant differences in the *communicative accuracy* of the gaze shifts produced by this model and those produced by human confederates, $F(1, 1958) = 0.03, p = ns$. Similarly, no differences in accuracy were found between the current model and the baseline model, $F(1, 1958) = 0.17, p = ns$. The

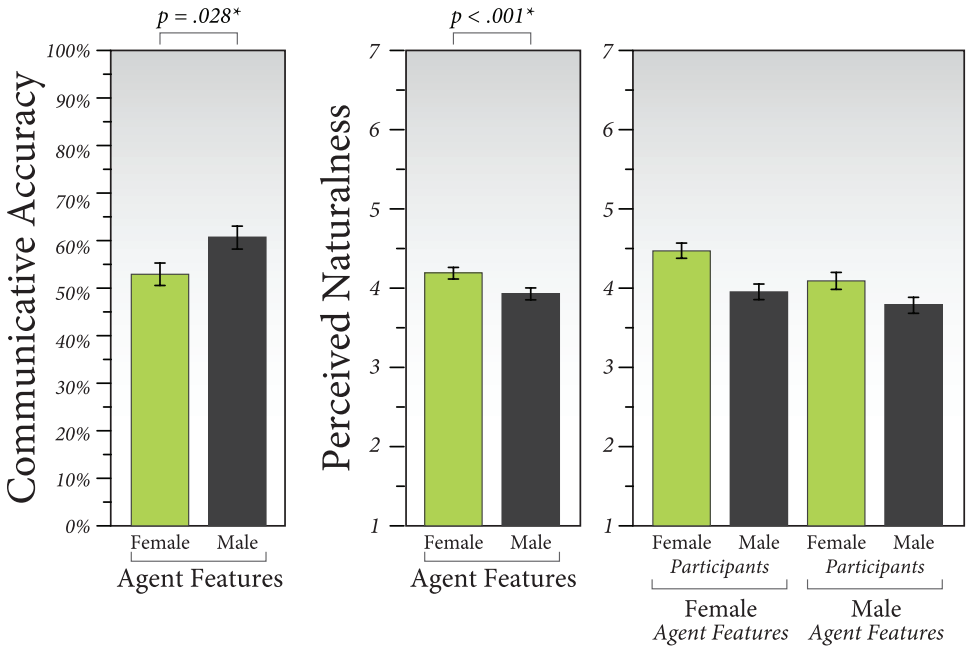

Figure 3.5: Communicative accuracy and perceived naturalness across agents with male and female features.

results suggest that the gaze shifts generated by the proposed gaze shift model are just as accurate as those performed by human confederates and those generated by the baseline model.

Perceived Naturalness – Comparisons across conditions showed that participants rated gaze shifts generated by the current model as marginally more *natural* than those generated by the baseline model, $F(1, 1958) = 3.34, p = .068$. Comparisons over *realism* (one of the items included in the perceived naturalness scale) found that gaze shifts produced by the current model were rated as significantly more realistic than those generated by the baseline model, $F(1, 1958) = 5.75, p = .017$. Results on the communicative accuracy, perceived naturalness, and realism measures are illustrated in Figure 3.4.

Gender – Comparisons across gender showed that participants rated gaze shifts performed by the agent with female features as significantly more natural than those performed by the agent with male features, $F(1, 1958) = 17.17, p < .001$. On the other hand, communicative accuracy of the gaze shifts performed by the agent with male features was significantly higher than that of the shifts performed by the agent with female features, $F(1, 1958) = 4.85, p = .028$. Finally, the analysis found a marginal interaction between participant gender and the gender features of the virtual character, $F(1, 1958) = 3.20, p = .074$. Figure 3.5 illustrates these results.

Discussion of Model Validation

The model presented above builds on findings from neurophysiology and procedurally specifies combined head and eye movements to create humanlike gaze shifts. The neurophysiological basis of the model helps achieve effective communication and subjective naturalness, while the procedural implementation allows for parametric control over the gaze shifts generated by the model. The results from the validation study show that gaze shifts generated by the current model communicate gaze direction as accurately as do human gaze shifts and those generated by a baseline model. The small gaze targets in the task make it challenging one, with performance between 60-70%. However, this is considerably better than the 25% performance expected by chance, and exhibited in the agent control condition. The results also show that gaze shifts generated by this model are perceived as marginally more natural and significantly more realistic than those generated by the baseline model. While the results suggest a gender effect from the virtual character, the measured difference may be due to any one of a number of differences in the characters' designs.

This neurophysiologically inspired model of head-eye coordination for gaze shifts provides a mechanism for synthesizing a key building block of gaze behavior. A key advantage of the model is that it provides parametric controls that may be used to achieve communicative outcomes. The experimental evaluation presented in the next section demonstrates how the use of the alignment parameter can be used in an educational scenario to provide a tradeoff in the high-level outcomes of affiliation and learning.

3.4 Experimental Evaluation

This section presents an investigation into how subtle parameters of gaze might be manipulated to achieve high-level outcomes. One of these parameters is the alignment of the head, which plays a substantial role in shifting a viewer's visual attention (Hietanen, 1999). Gazing at someone with the head fully aligned (affiliative gaze) might be perceived differently than gazing at someone out of the corner of the eyes with the head aligned towards information of interest in the context of interaction (referential gaze). An embodied agent might be able to manipulate this, and other parameters, to achieve specifically desired effects.

The validated model of gaze shifts presented above provides parameters that allow exploration of how manipulation of these low-level parameters can achieve valuable high-level effects. The main study of this chapter explores manipulating the *head alignment*

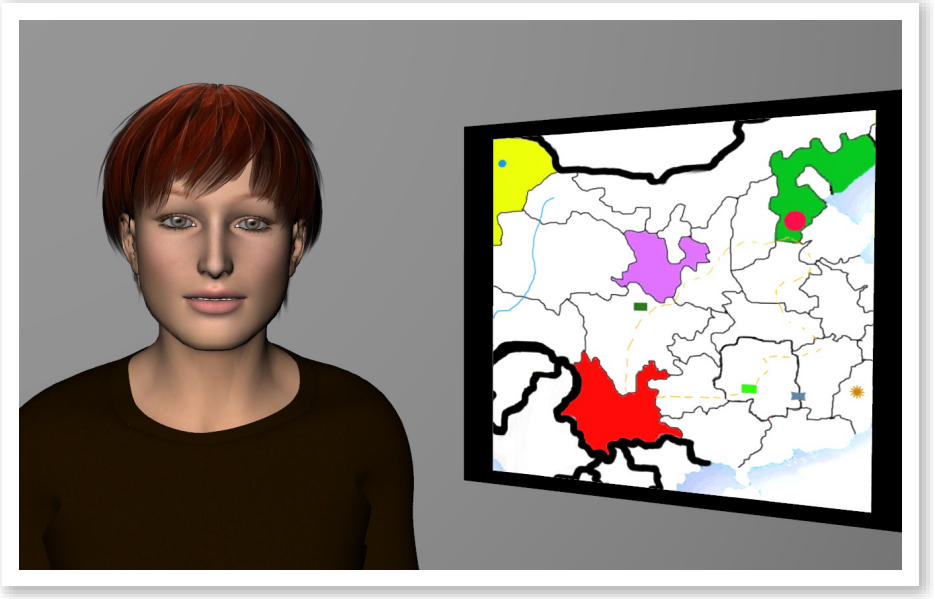

Figure 3.6: One of the humanlike virtual agents used in a study which examined how agents could use their gaze effectively in an educational scenario. Here the agent is giving a lecture on geographical locations of ancient China.

parameter in order to create more affiliative or more referential gaze cues, with the goal of increasing feelings of connection with the agent and increasing learning respectively. The experiment was carried out in the context of an educational scenario, with the virtual agent serving as a lecturer to the human participant. The virtual agent taught the participant about a specific subject from ancient Chinese history. A map of China was used as a reference to facilitate the descriptions of geographical locations. An example of the interface presented to participants is shown in Figure 3.6.

Hypotheses

The experiment was designed to test three hypotheses. First, it is important to confirm that the mere presence of an agent will elicit better learning than only audio.

Hypothesis 1 — The presence of an embodied agent will result in better recall performance than only hearing audio with no accompanying agent.

The literature strongly supports the first hypothesis, in which it is shown that teachers who gaze at their students lead the students to learn better (Otteson and Otteson, 1980; Sherwood, 1987). This evaluation seeks to show these same effects can be achieved by virtual agents using a humanlike model of gaze shifts.

Hypothesis 2 — An agent which employs more affiliative gaze (maintains higher head

alignment with the participant) will garner higher subjective evaluations than one which uses more referential gaze (maintains higher head alignment with the information being referred to).

This hypothesis is supported by the fact that people are perceived as being more intelligent, more trustworthy, and more friendly if they make more direct eye contact (Argyle and Cook, 1976; Fullwood and Doherty-Sneddon, 2006). This evaluation extends this work to show that head alignment makes an impact on the positive subjective effects of mutual gaze, and that the current model is capable of achieving these effects.

Hypothesis 3 — Viewing an agent using more referential gaze should result in better recall performance. This will especially be true when the information to be recalled relies on building associations to objects in the environment.

Referential gaze helps build associations between information and objects in the environment. In communication between a human and virtual agent with a shared environment, being able to perceive the agent's eye gaze to different objects in the environment reduces the time and verbal communication needed for grounding references (Liu et al., 2011). Head alignment has the potential to strengthen or weaken this effect since it plays a substantial role in shifting a viewer's visual attention (Hietanen, 1999). When the agent's head is aligned more fully with the object under consideration, this should serve as a stronger referential gaze cue than if the head is not aligned. This evaluation demonstrates that the current gaze shift model can achieve this effect.

Participants

Twenty participants were recruited for this study (10 females and 10 males), with ages ranging from 19 to 65 ($M = 27.8$, $SD = 14.3$). Fifteen participants were students and five were working members of the community. All were native English speakers. Student participants came from a number of different fields, including psychology, engineering, and business. All participants were recruited using a combination of campus flyers and on-line student job forums.

Study Design

This study employed a within-subjects design. The experiment involved one factor, *type of gaze*, with four levels, each of which is defined as follows:

- **Audio**: The agent is shown briefly during its introduction. Then the lights in the virtual scene are extinguished for the duration of the lecture, except for a spotlight on

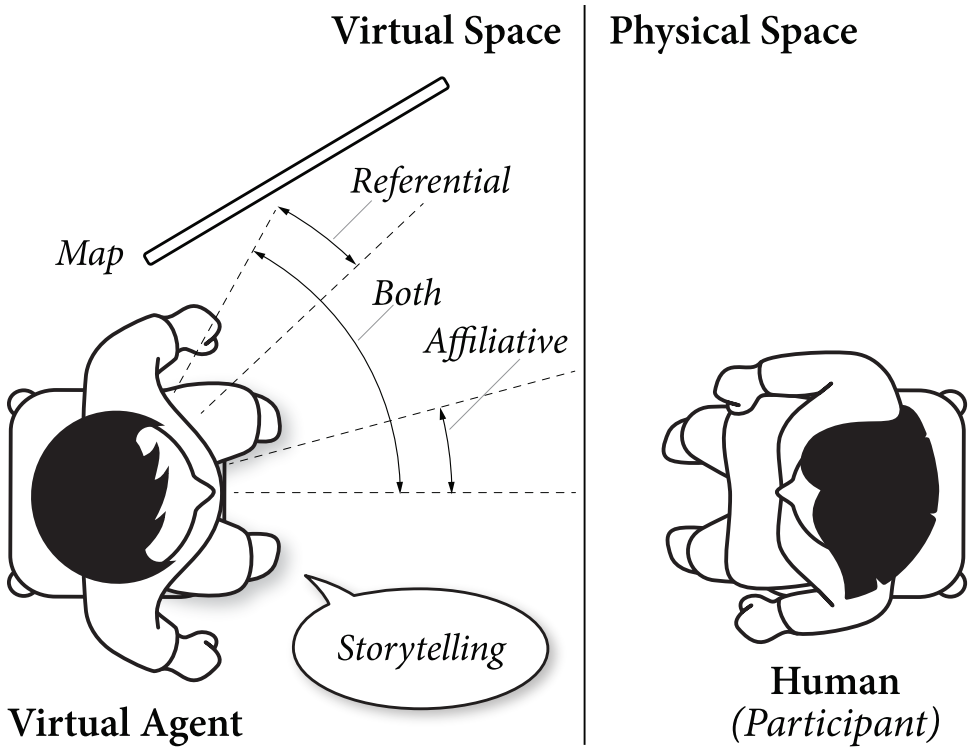

Figure 3.7: A diagram of the setup of the study showing the range of the agent's head movements for each gaze condition. The agent's eye motions (not depicted) always move the full distance from eye contact with the participant to eye contact with each map location being referred to.

the map, so only the map can be seen and the agent is in complete darkness. At the end of the lecture, the scene lights turn back on for the agent to give its instructions to the participant on taking the subjective evaluation and quiz for the lecture.

- **Affiliative**: The agent keeps its head aligned with the participant as much as possible during the lecture. When the agent is making direct eye contact with the participant, the head is fully aligned with the participant. When the agent shifts its eye gaze to refer to something on the map, the head aligns as little as possible with the map—as much as the agent's OMR will allow—so as to keep the head aligned towards the participant.

- **Referential**: The agent keeps its head aligned with the map as much as possible during the lecture. When the agent is gazing at the map, the head is fully aligned with the map. When the agent shifts its eye gaze back to the participant, the head aligns as little as possible with the participant—as much as the agent's OMR will allow—so as to keep the head aligned towards the map. The referential condition

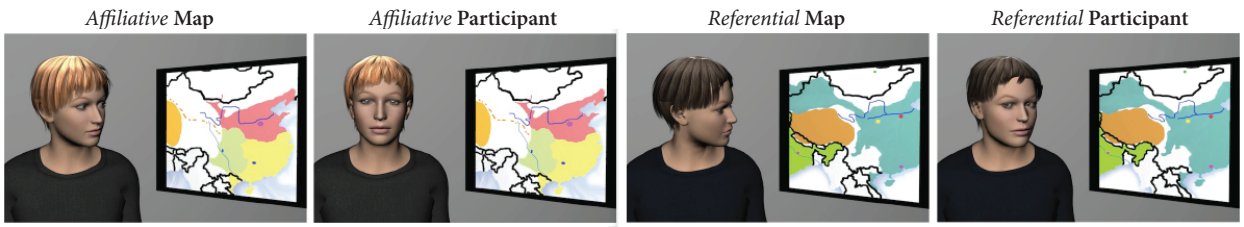

Figure 3.8: A visual depiction of an agent in different gaze conditions. From left-to-right: (1) Affiliative, looking at map; (2) Affiliative, looking at participant; (3) Referential, looking at map; (4) Referential, looking at participant.

is "more referential" (borrowing the term "referential" from linguistics) because the head maintains alignment with the information being referred to on the map, both when looking at the participant (out of the corner of the eyes) and when looking at the map (head and eyes fully aligned).

- **Both**: When gazing at the participant, the agent keeps its head fully aligned to the participant. When gazing at the map, the agent aligns its head fully with the map.

Figure 3.7 and Figure 3.8 illustrate the difference in head motion between the three conditions *affiliative*, *referential*, and *both*. It should be stressed that in all three of these conditions, the agent's eyes move the full distance from eye contact with the participant to each map location being referred to. The agent's eyes always converged on each map location while it was being referred to.

Each participant viewed four lectures given by four different virtual agents, each utilizing a different gaze condition. Thus, every participant was exposed to all four gaze conditions. All agents, three of which can be seen in Figures 3.6 and 3.8, were specifically designed to look and sound androgynous so as to eliminate gender biases as much as possible. Each agent always gave the same lecture, but pairings between agent and gaze condition were stratified across participants for balance. The order in which the lectures—and accordingly gaze conditions—were presented to the participants was completely randomized to offset any bias due to fatigue. Lectures were also kept informationally distinct in an attempt to decrease learning bias. Each lecture was carefully controlled to be near the same length—approximately three minutes.

During each lecture, the agent makes eleven gaze shifts to the map, each followed shortly by a gaze shift back to the participant. The agent fixates its gaze on different targets on the map as they are mentioned, e.g. while giving facts about a Chinese city visible on the map. Timing of this gaze shift is done according to Griffin (2001) and Meyer et al.

(1998), which indicate that a deictic gaze shift should occur 800 – 1000 ms before the object being gazed at is mentioned in speech.

Procedure

This experiment was conducted in a closed study room with no outside distraction. Participants entered the experiment room and were asked to sit at a table with a single computer monitor and mouse. The monitor was a 32-inch flat panel display, allowing the virtual agent representation (only head and shoulders) to be near life size. This setup can be seen in Figure 3.9. The experimenter gave the participant a brief description of what they would be asked to do in the experiment, and then asked them to review and sign a consent form. The experimenter told the participants that they were going to be listening to and quizzed on four short lectures from different virtual lecturers, each on a topic pertaining to ancient China.

After consenting, the experimenter told participants that they could press the start button on the screen after he left the room. Upon pressing the start button, the first lecturer (randomly chosen by the software) began introducing itself and giving its lecture. Upon completion of the lecture the screen went black, and participants filled out on paper a subjective evaluation of the lecturer they had just viewed, followed by a quiz on the material presented in the lecture. During the initial instructions, the experimenter made it clear that the quiz was to be filled out *after* the subjective evaluation had been completed. This allows the subjective evaluation to double as distractor task, strengthening any subsequent recall measures. All participants were monitored via closed-circuit camera by the experimenter to ensure that these instructions were followed, but no participants were observed to neglect the instructions.

After filling out the subjective evaluation and quiz, the participant could go back to the monitor and click the on-screen button to begin the next lecture. This process was repeated until all four lectures had been viewed, rated, and quizzed. At this point, the experimenter re-entered the room with a short questionnaire of demographic information. Following completion of the questionnaire, the experimenter debriefed and payed the participant. The total experiment took approximately 30 minutes, and participants were paid $5.

Measures

The experiment involved one independent manipulated variable, *type of gaze*, manipulated within participants. The dependent variables included objective measurements for evaluat-

Figure 3.9: The physical setup of the experiment.

ing participants' recall of the lecture material and subjective measurements for evaluating the participants' impressions of the virtual agent.

The objective measurement of recall involved quizzes taken by all participants following each lecture. Each quiz had ten short-answer questions. These questions were split into three categories (not visible to the participant). One category included three questions that asked about information not directly associated with information on the map. Thus, referential gaze should not have had an effect on the recall of this information. An example question in this category would be, "In what year did the Jin dynasty overtake control of China?" The Jin dynasty was not represented on the map during the associated lecture. The second category, consisting of four questions, asked about purely spatial information. For example, one question here was "Which of the *Three Kingdoms* dynasties extended farthest south?" This question is only answerable by having studied the map. The third category, including the remaining three questions of the quiz, relied on building associations between verbal lecture content and spatial map locations. For example, "Give one reason why Emperor Wen declared the city of Luoyang to be his capital city." Referential gaze was expected to make the biggest impact on questions from the latter two categories. Questions from all three categories were randomly permutated to create the final ten-question quiz.

The subjective measurements were split into six broad indicators. Each question within the indicators took the form of a seven-point rating scale. Item reliability (Cronbach's α) was acceptable or better for all except *skilled communicator*. The full questionnaire for this

study is provided in Appendix A.

1. *Likeability*: Four-item measure of how likeable the participant found the agent to be. Includes questions on perceived friendliness and helpfulness. (Cronbach's $\alpha = .78$)

2. *Rapport*: Six-item measure of how much the participant felt feelings of rapport in relation to the agent. Questions asked, e.g., how well the participant felt he or she connected with the agent and how willingly he or she would disclose personal information to the virtual agent following the lecture. (Cronbach's $\alpha = .84$)

3. *Trust*: Two-item measure of how trustworthy the participant perceived the agent to be. Includes ratings of trustworthiness and honesty. (Cronbach's $\alpha = .72$)

4. *Intelligence*: Three-item measure of how intelligent the participant perceived the agent to be. Includes ratings of competence and expertise. (Cronbach's $\alpha = .84$)

5. *Skilled Communicator*: Three-item measure of how effective at conveying lecture material the participant perceived the agent to be. Item reliability was questionable (Cronbach's $\alpha = .62$)

6. *Engagement*: Six-item measure of how engaged the participant felt during the lecture. Includes personal ratings of focus, attentiveness, and satisfaction. (Cronbach's $\alpha = .89$)

Manipulation Checks

Part of the post-lecture questionnaires included questions to check whether the manipulations were effective. The first test was whether the agents created for this evaluation were actually being perceived as androgynous. Two questions in the form of seven-point rating scales, anchored by *very feminine* (value = 7) and *very masculine* (value = 1), were asked. One referred to the agent's *appearance* and the other to the agent's *voice*.

To check that the *gaze type* manipulations were being noticed between the visible agent conditions, the experimenter asked participants to rate from 0% to 100% how much they felt the agent was paying attention to them and to the map. It was expected that participants would feel more attended to in the *affiliative* and *both* conditions than in the *referential* condition. Conversely, it was expected that participants would feel like the agent was attending to the map more in the *referential* and *both* gaze conditions than in the *affiliative* condition.

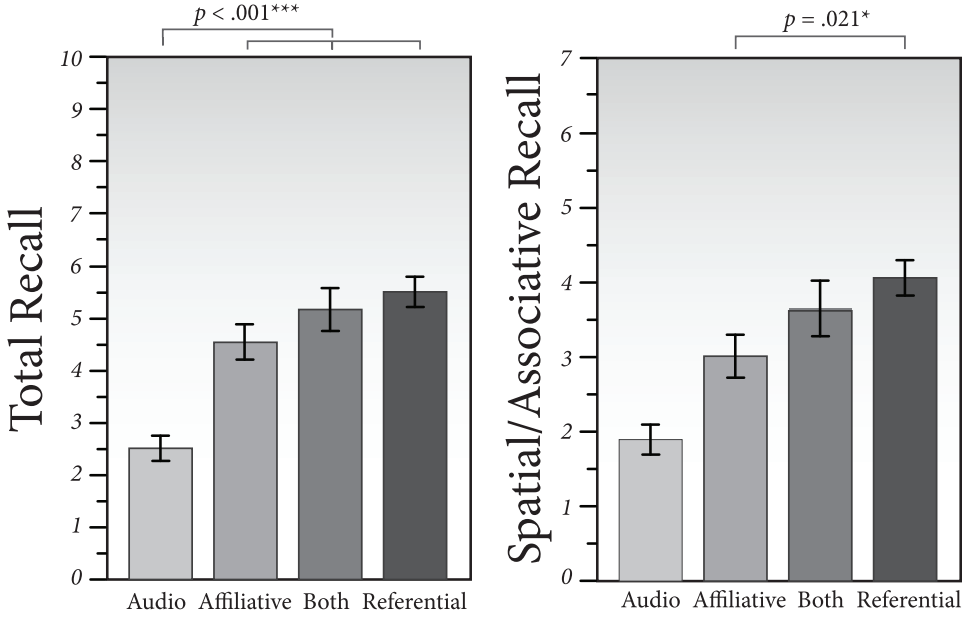

Figure 3.10: Objective measure (recall measured by post-lecture quiz). On the left is the total quiz performance, on the right is the quiz performance when only considering a subset of the questions: those dealing with spatial information and building associations.

Results

Analysis of the data was conducted using a repeated measures analysis of variance (ANOVA). Tukey-Kramer HSD was used to control the experiment-wise error rate in all post-hoc tests. The analysis started with the manipulation check for the perceived gender of the virtual agent and the *gaze type* manipulations. First, it was found that the agents were rated on average as being mostly androgynous (appearance: $M = 4.88, SD = 1.21$ voice: $M = 4.51, SD = 1.58$). Participants felt more attended to in the *affiliative* gaze condition versus the *referential* condition, $F(1, 69) = 12.53, p < .001$. They also felt more attended to in the *both* condition versus *referential*, $F(1, 69) = 4.37, p = .040$. Finally, participants felt that the agent attended to the map more in the *referential* condition versus *affiliative*, $F(1, 69) = 7.75, p = .007$. This difference was not found to be significant for the *both* condition versus *affiliative*, $F(1, 69) = 0.37, p = .55$, however participants rated the agent as attending more to the map in the *referential* condition over the *both* condition, $F(1, 69) = 4.75, p = .033$.

Next analyzed were the objective results in the form of recall quiz scores (Figure 3.10). In terms of overall score, the *audio* condition resulted in significantly lower recall than the other three visible agent conditions, including *affiliative* gaze, $F(1, 69) = 19.38, p < .001$, *referential* gaze, $F(1, 69) = 37.78, p < .001$, and *both*, $F = 31.91, p < .001$. When considering only

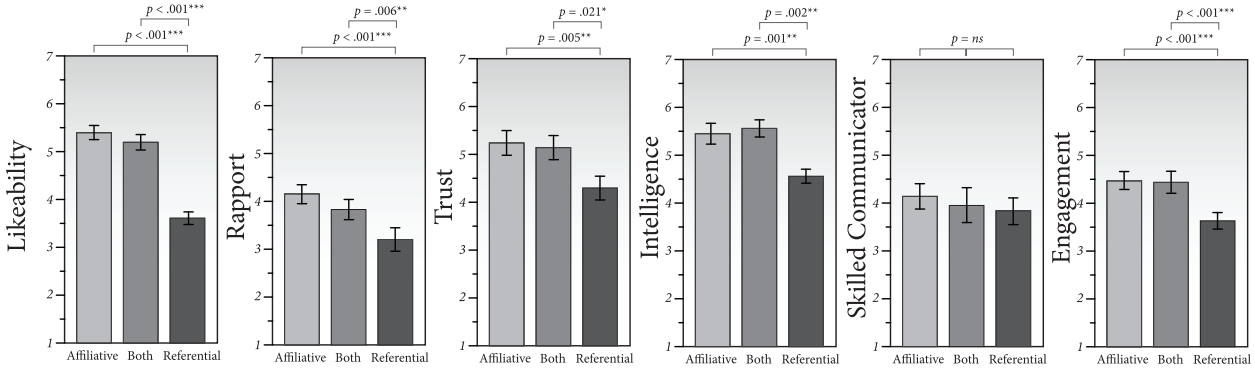

Figure 3.11: Results for subjective evaluations (likeability, rapport, trust, intelligence, skilled communicator, and engagement) based on gaze condition.

the seven (out of ten total) questions that dealt with purely spatial map information and building associations between verbal lecture content and locations on the map, it was found that the *referential* gaze condition resulted in significantly better recall performance than the *affiliative* gaze condition, $F(1, 69) = 5.62, p = .021$. The increase in recall performance from the *both* condition over *affiliative* does not quite reach significance, $F(1, 69) = 2.58, p = .11$. *Referential* and *both* were not significantly different, $F(1, 69) = 0.59, p = .45$.

Finally, the subjective measures were analyzed. On the *likeability* scale, the *referential* condition rated lower than both the *affiliative* condition, $F(1, 69) = 58.86, p < .001$, and *both* condition, $F(1, 69) = 52.65, p < .001$. On the *rapport* scale, the *referential* condition also rated lower than both *affiliative*, $F(1, 69) = 13.25, p < .001$, and *both*, $F(1, 69) = 7.95, p = .006$. The *trust* scale had similar results, with the *referential* condition rated lower than both *affiliative*, $F(1, 69) = 8.63, p = .005$, and *both*, $F(1, 69) = 5.55, p = .021$. The *intelligence* scale again shows the *referential* condition getting rated lower than *affiliative*, $F(1, 69) = 11.38, p = .001$, and *both*, $F(1, 69) = 10.99, p = .002$. The *skilled communicator* scale yielded no significant results between conditions, but the engagement scale showed the *referential* condition getting rated significantly lower than the *affiliative* condition, $F(1, 69) = 14.53, p < .001$, and *both* condition, $F(1, 69) = 17.16, p < .001$. These results are summarized in Figure 3.11.

Discussion

The purpose of the experiment was to demonstrate how subtle changes in gaze behavior can lead to significant high-level effects, with the goal of improving learning and positive feelings of affiliation. To do this, the evaluation manipulated the *head alignment* parameter in the gaze shift model implemented on various virtual lecturing agents. The study showed that by manipulating just this parameter for a virtual agent shifting its gaze from the

participant to an object in the scene (i.e., the map of China) and back again, the agent could achieve very different subjective and objective effects, including the participant's feelings toward the agent and information recall respectively.

Providing support for the first hypothesis, the mere presence of an agent resulted in better recall performance than audio alone, no matter which gaze condition was being used. For five out of six subjective scales, the *affiliative* and *both* gaze conditions achieved better ratings than the *referential* condition. This lends support for the second hypothesis, showing that human listeners prefer when the virtual agent speaker aligns its head fully with the participant while speaking, rather than looking out of the corners of its eyes with its head aligned towards something else. The *skilled communicator* scale did not exhibit this effect, which could mean that participants thought the agent was doing a similarly decent job of communicating the lecture content no matter which gaze condition was used. The evaluation also yielded strong support for the third hypothesis, where it was shown that *referential* gaze results in better participant recall than *affiliative* gaze. By keeping its head aligned with the map as much as possible, the agent compelled the participant to concentrate more on the map and learn the spatial locations better, while building associations between verbal lecture content and those same locations. As a reminder, the same amount of gaze was used by the agent in both the *affiliative* and *referential* gaze conditions. Hence in both conditions the participant was able to benefit from gaze as an arousal stimulus to learn the lecture material better. The difference lies in the way each gaze shift was performed, proving that not all ways of gazing are equal and subtle changes can create significant outcomes. Overall, this study showed that the current model was capable of achieving targeted outcomes.

3.5 Chapter Summary

Embodied agents' potential in computer interfaces is increasing with the ever-growing popularity of interactive games and computer-based tools for learning, motivation, and rehabilitation. However, agent designers need controllable models of interactive behavior and evidence that these models effectively improve outcomes such as learning and rapport. This chapter provided one such example in the context of gaze.

This chapter presented a parametric, computational model of head-eye coordination that can be used in the animation of directed gaze shifts for embodied agents. The model is based on research in human neurophysiology. It incorporates control parameters that allow for adapting gaze shifts to the characteristics of the environment, the gaze targets, and the idiosyncratic behavioral attributes of the virtual character. A validation study

confirmed that the model communicates gaze targets as effectively as real humans do, while being preferred subjectively to state-of-the-art models.

This chapter also demonstrated that it is possible to implement in virtual agents the very subtle gaze cues that humans use to great effect in social situations, and that manipulating these cues can achieve significant objective and subjective high-level effects in a human interacting with the agent. Designers of virtual agents can use these subtle gaze behaviors, such as head alignment, to reach different desired outcomes. If the agent designer wants human interlocutors to pay more attention to specific objects in the environment, possibly to learn more about them, the agent could be programmed to use high head alignment when gazing to those objects. Similarly, if the agent designer wants the agent to build a stronger relationship with the human interlocutor, increasing feelings of, e.g., rapport and trust, the agent should be programmed to use high head alignment when gazing towards the human. This model of gaze behavior offers a simple and effective means to control the low-level gaze parameters found in physiological research. Virtual agent designers can use and build off of this model to create rich, compelling gaze behaviors that accomplish the high-level effects they wish to achieve.

The gaze shift model is a powerful tool for creating a rich variety of natural gaze motions for agents, but although it specifies *how* to gaze, it does not handle *when* to gaze. In basic conversational interactions, an artificial agent will by default make continuous eye contact with the user. This issue is addressed in the next chapter.

4 GAZE AVERSION

Engaging in mutual gaze with others has long been recognized as an important component of successful social interactions. People who exhibit high amounts of mutual gaze are perceived as competent, attentive, and powerful (Argyle and Cook, 1976). In the same way, embodied agents that use eye contact to exhibit some degree of mutual attentiveness have been shown to achieve a number of positive social and conversational functions, including building rapport with people (Wang and Gratch, 2010) and increasing positive perceptions of affiliation (as demonstrated in the previous chapter).

However, periodically engaging in gaze *aversion* in conversation also serves a number of communicative functions. Gaze aversion is defined as the intentional redirection of gaze away from the face of an interlocutor, and it is used in conversations to achieve three primary functions: signaling cognition, intimacy modulation, and floor management. First is the *cognitive* function: speakers spend much more time averting their gaze than listeners in order to better attend to the planning and delivery of their utterances while limiting external distraction (Argyle and Cook, 1976). Second is the *intimacy-modulating* function: periodic gaze aversions while speaking or listening can serve to modulate the overall level of intimacy in the conversation (Abele, 1986). Third is the *floor management* function: looking away while pausing during speech is used to indicate that the conversational floor is being held and the speaker should not be interrupted (Kendon, 1967). These three functions—cognitive, intimacy, and floor management—correspond with three social contexts identified in previous work on social gaze in HRI: projecting mental state, establishing agency, and regulating the interaction process respectively (Srinivasan and Murphy, 2011).

While the social science literature has highlighted the positive functions of gaze aversion, it does not provide the precise spatial and temporal measurements required to synthesize a model of gaze aversion for artificial agents that could similarly achieve these functions. This chapter extends existing social science knowledge through the collection and analysis of a corpus of human-human conversations in order to obtain precise spatial and temporal parameters of gaze aversion movements in relation to speech and conversational function. This data collection and analysis enabled the design of gaze aversion behaviors for embodied agents.

From the human-human data collection, a gaze controller was designed that can generate appropriately timed gaze aversion behaviors for embodied agents. This controller was first implemented for virtual agents, and an evaluation was conducted to demonstrate its effectiveness in achieving intended conversational functions. The agents' new gaze aversion

behaviors were evaluated for their ability to achieve positive conversational functions in a laboratory experiment with 24 participants. In this experiment, human participants interacted with four different virtual agents in four conversational tasks, each of which was designed to test a different conversational function of gaze aversion (Figure 4.5). Results show that virtual agents employing gaze aversion are perceived as thinking, are able to elicit more disclosure from human interlocutors, and are able to regulate conversational turn-taking.

Next, this model was extended to apply to conversational robots, enabling them to also achieve these functions in conversations with people. This chapter presents a system that addresses the challenges of adapting human gaze aversion movements to a robot with very different affordances, such as a lack of articulated eyes. This system, implemented on the NAO platform (Figure 4.8), autonomously generates and combines three distinct types of robot head movements with different purposes: face-tracking movements to engage in mutual gaze, idle head motion to increase lifelikeness, and purposeful gaze aversions to achieve conversational functions. The results of a human-robot interaction study with 30 participants show that gaze aversions implemented with this approach are perceived as intentional, and robots can use gaze aversions to appear more thoughtful and effectively manage the conversational floor.

The next section reviews the social-scientific literature on gaze aversion as well as previous work on modeling gaze for embodied agents. Next presented is the analysis of a video corpus of human dyadic conversations from which were obtained temporal parameters of gaze aversion. Following that, an implementation and evaluation of gaze aversion behaviors are presented for first virtual agents and then humanlike robots.[1]

Research Questions

- When, why, and in which direction do people avert their gaze from one another in open conversation?
- How can we enable embodied agents to appropriately avert their gaze in conversations with people?
- What are the positive interaction outcomes achievable by having an agent avert its gaze in a humanlike way?

[1]Portions of this chapter were published in Andrist et al. (2013b) and Andrist et al. (2014). I would like to acknowledge Xiang Zhi Tan, an undergraduate student collaborator, for his contributions in implementing the gaze aversion mechanism on the Nao robot platform.

4.1 Related Work

In order to design gaze aversion behaviors for virtual agents, it is necessary to first examine how humans use gaze aversion in conversation. This section presents an overview of social and cognitive science research on human gaze aversion in conversations, in which three primary functions of gaze aversion have been identified: facilitation of cognitive processing, intimacy modulation, and floor management.

Conversational Gaze Aversion in Humans

Previous social science research has identified a number of underlying mechanisms to explain human gaze aversion and the social functions it achieves. One such mechanism is the "cognitive interference hypothesis" (Beattie, 1981; Doherty-Sneddon and Phelps, 2005; Ehrlichman and Micic, 2012; Glenberg et al., 1998). This hypothesis posits that gaze aversions facilitate cognitive activity by disengaging the speaker from the environment and limiting visual inputs. The frequency of speaker gaze aversions in conversation has been shown to be related to the difficulty of *cognitive processing* (Glenberg et al., 1998). When constant mutual gaze is required from someone speaking spontaneously in social interaction, that person's speech becomes significantly impaired (Beattie, 1981).

Averting gaze has the practical benefit of improving cognitive performance (Glenberg et al., 1998). This is because an interlocutor's face is rich in social information, and is a cognitively demanding visual target. When people avert their gaze from their interlocutor, they are able to deploy additional cognitive resources to the task of thinking or remembering information. Research also shows that forcing oneself to look away from a conversational partner while retrieving information from long-term memory or when planning a response to a challenging question significantly improves performance (Glenberg et al., 1998; Phelps et al., 2006). With these aversions, a speaker signals to the listener that cognitive processing is occurring, creating the impression that deep thought or creativity is being undertaken in formulating their speech (Argyle and Cook, 1976).

Previous research has also shown that eye contact is a significant contributor to the intimacy level of an interaction, such that reducing eye contact can decrease the perceived intimacy of a conversation (Argyle and Cook, 1976). For example, people generally engage in less eye contact while responding to embarrassing questions than while responding to less objectionable questions (Exline et al., 1965). Research has shown that children avert their gaze more when answering questions in face-to-face rather than video-mediated conversations (Doherty-Sneddon and Phelps, 2005). In general, people frequently avert their gaze to alleviate feelings of self-consciousness and, while listening, to make speakers

more comfortable and to reduce negative perceptions associated with staring (Abele, 1986). Other work has examined how topic intimacy and eye contact interact over the course of a conversation (Abele, 1986).

Another primary function of gaze aversion is to facilitate turn-taking. When exchanging speaking turns in conversation, a pattern is frequently observed in which the first speaker finishes speaking, looks toward their interlocutor and engages in momentary mutual gaze, and then the second speaker averts their gaze and begins their speaking turn (Kendon, 1967; Novick et al., 1996). By looking away at the beginning of an utterance, the speaker strengthens his or her claim over the speaking turn. Looking away during a pause in speech is also used to indicate that the conversational floor is being held and that the speaker should not be interrupted (Kendon, 1967).

This chapter groups the social-scientific findings discussed above into three broad conversational functions: the *cognitive*, *intimacy-modulating*, and *turn-taking* functions of gaze aversion. These groupings informed an empirical investigation to develop a more computational understanding of how gaze aversions are temporally and spatially employed in conversation.

4.2 Modeling Gaze Aversion

As outlined above, research in the social sciences has identified a number of conversational functions of gaze aversion. To extend this knowledge to include temporal patterns that will be directly implemented on embodied agent systems, video data was collected from 24 dyadic conversations. Statistical parameters were derived for the length, timing, and frequency of gaze aversions in relation to speech and conversational functions. Three primary conversational functions of gaze aversion were addressed in this analysis, which are defined and described below.

Cognitive – These gaze aversions serve to disengage a speaker's attention from the face of their interlocutor in order to facilitate thinking and remembering (Glenberg et al., 1998). With these aversions, people signal that cognitive processing is occurring while creating an impression that deep thought or creativity is being undertaken (Argyle and Cook, 1976).

Intimacy-modulating – Gaze aversions also serve to moderate the overall intimacy level of the conversation. Periodic gaze aversions while listening can serve to make speakers more comfortable and reduce negative perceptions associated with staring (Abele, 1986).

Turn-taking – These gaze aversions serve to regulate conversational turn-taking. By looking away at the beginning of an utterance, the speaker strengthens his or her claim over

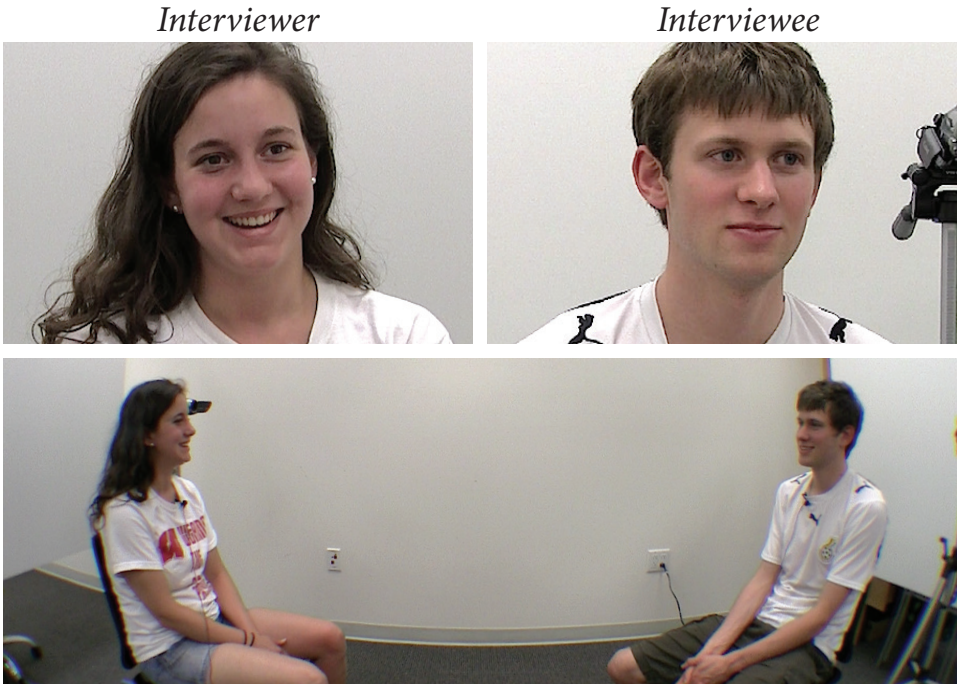

Figure 4.1: A participant dyad from the data collection. The participants were designated as the *interviewer* and the *interviewee*. The interviewers were instructed to ask interviewees about their movie preferences.

the speaking turn. Looking away during a pause in speech indicates that the conversational turn is being held and that the speaker should not be interrupted (Kendon, 1967).

Data Collection & Analysis

The data collection study included 24 females and 24 males, aged 18 to 28 and previously unacquainted (Figure 4.1). Each dyad engaged in a structured conversation for approximately five minutes. One participant was instructed to learn about the other participant's taste in movies, with the goal of making a movie recommendation. All conversations were counterbalanced for both gender—female and male—and conversational role—recommender and recommendee. The study also counterbalanced gender concordance—there were equal numbers of gender-matched and gender-mismatched dyads.

VCode was used to analyze the recorded videos of the participants' gaze and speech.[2] Video coding was carried out by two independent coders with partial overlap. Sequences of time spent speaking and averting gaze were annotated. Cognitive events were marked

[2]http://social.cs.uiuc.edu/projects/vcode.html

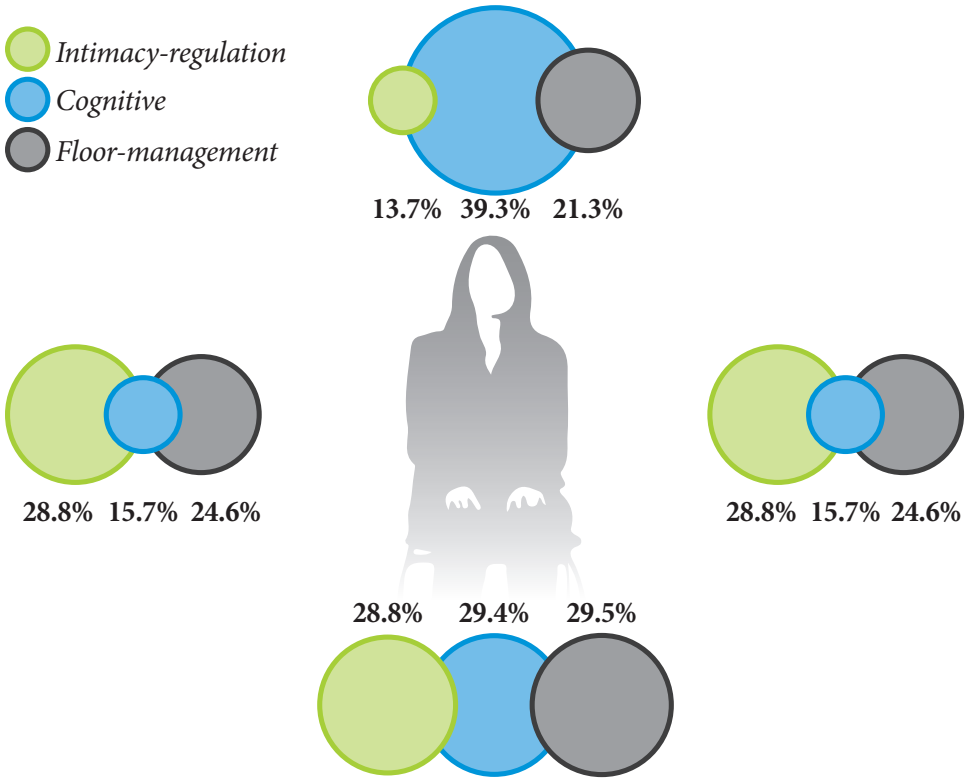

Figure 4.2: Percentages of gaze aversions directed up, down, and to the side, split by conversational function. Intimacy-regulating and floor-managing gaze aversions are more likely to be directed sideways, while cognitive gaze aversions are more likely to be directed upwards.

as discrete points in time where the participants appeared to be thinking or remembering, commonly occurring at the beginning of responses to questions.

Gaze aversions were coded for the conversational function that they were perceived to be supporting: cognitive, intimacy-modulating, or floor management. This coding took place in three passes. In the first pass, the coder marked gaze aversions as cognitive if they occurred near perceived cognitive events, e.g., when a participant appeared to be thinking of a response to a question. In the second pass, gaze aversions were marked as floor management if they occurred near the beginning of a speaking turn or during a pause in speech. In the third pass, all remaining gaze aversions were labeled as intimacy-modulating. An inter-rater reliability analysis showed substantial agreement on the identification of gaze aversions and their conversational function (Cohen's $\kappa = .747$).

This analysis yielded timing statistics for different kinds of gaze aversions, including the frequency, length, and temporal placement of these gaze aversions relative to speech (Table 4.1). Each of these parameters is modeled as a Gaussian distribution with means and

standard deviations derived from the data. For cognitive gaze shifts, the length ($M = 3.54s$, $SD = 1.26s$), start time in relation to cognitive events ($M = -1.32s$, $SD = 0.47s$), and end time after cognitive events ($M = 2.23s$, $SD = 0.63s$) are modeled. A cognitive event is any point in time at which the agent should display a state of deep cognitive processing, e.g., at the beginning of a response to a user's question. For intimacy-modulating gaze aversions, the length and time between consecutive gaze aversions while speaking (length: $M = 1.96s$, $SD = 0.32s$; time between: $M = 4.75s$, $SD = 1.39s$), and while listening (length: $M = 1.14s$, $SD = 0.27s$; time between: $M = 7.21s$, $SD = 1.88s$) are modeled. For floor management gaze aversions, the length ($M = 2.30s$, $SD = 1.10s$), start time in relation to the start of the next utterance ($M = -1.03s$, $SD = 0.39s$), and the end time in relation to that same utterance ($M = 1.27s$, $SD = 0.51s$) are modeled. The model also captures the time before the end of a floor-passing utterance at which point mutual gaze is engaged with the interlocutor and no more gaze aversions are generated during that utterance ($M = -2.41s$, $SD = 0.56s$). Figure 4.3 illustrates each type of gaze aversion.

Each gaze aversion was also labeled for its direction as *up*, *down*, and *side* (Figure 4.2) revealing that intimacy-regulating and floor-managing gaze aversions were more likely to be directed sideways, while cognitive gaze aversions were more likely to be directed upwards.

4.3 Gaze Aversion in Virtual Agents

Implementation

Findings from the data analysis were first synthesized into a gaze controller for virtual agents that automatically plans and performs gaze aversions to accomplish the conversational functions previously discussed. This controller takes as inputs the current conversational state, the start time and length of upcoming planned utterances, and the time of upcoming cognitive events, and then supplies as outputs the start and end times of planned gaze aversions to be executed by the agent. The exact timings of the gaze aversions are drawn from the parameter distributions shown in Table 4.1.

Source of Inputs

Recognized speech from the user is passed to a dialogue manager that associates a semantic tag with the utterance and plans the agent's speech accordingly. For example, if the dialogue manager receives a recognized question, it will produce the associated answer. The dialogue manager sends upcoming cognitive events, speech events, and the current conversational

Table 4.1: Gaze aversion parameters in relation to conversational functions and coordinated with (before, after, or within) speech and cognitive events.

Conversational Function	Coordinated With	Parameter	Value
Cognitive	Cognitive Event	Length (sec)	3.54 (SD = 1.26)
		Start time (sec)	1.32 before (SD = 0.47)
		End time (sec)	2.23 after (SD = 0.63)
Intimacy	Speaking	Length (sec)	1.96 (SD = 0.32)
		Between (sec)	4.75 (SD = 1.39)
	Listening	Length (sec)	1.14 (SD = 0.27)
		Between (sec)	7.21 (SD = 1.88)
Turn-taking	Utterance Start	Frequency (%)	73.1
		Length (sec)	2.30 (SD = 1.10)
		Start time (sec)	1.03 before (SD = 0.39)
		End time (sec)	1.27 after (SD = 0.51)
	Utterance End	End time (sec)	2.41 before (SD = 0.56)

state to the gaze controller. Cognitive events could alternatively be passed to the gaze controller from a dedicated cognitive architecture, but in the current implementation, cognitive events were created by labeling some of the agent's utterances as "cognitively difficult" and generating a cognitive event at the beginning of those utterances.

Gaze Controller

Cognitive events are represented with a single timestamp, t_c. Planned speech events are represented as a vector containing start and end times, $[t_s, t_e]$. Conversational state, CS, indicates that the agent is currently in either *speaking* or *listening* mode. As the gaze controller receives these inputs from the dialogue manager, it continuously plans future gaze aversions in real-time. The first priority is to plan gaze aversions around upcoming cognitive events, t_c. The start and end times of the gaze aversion, $[GA_s, GA_e]$, are computed by drawing from the cognitive parameter distributions shown in Table 4.1. The controller next looks for upcoming speech events and calculates first if a turn-taking gaze aversion will be performed. If a gaze aversion will be performed, the controller then calculates $[GA_s, GA_e]$ around the start of the utterance, t_s, by drawing from the turn-taking parameter distributions provided in Table 4.1. Finally, the controller calculates the next intimacy gaze

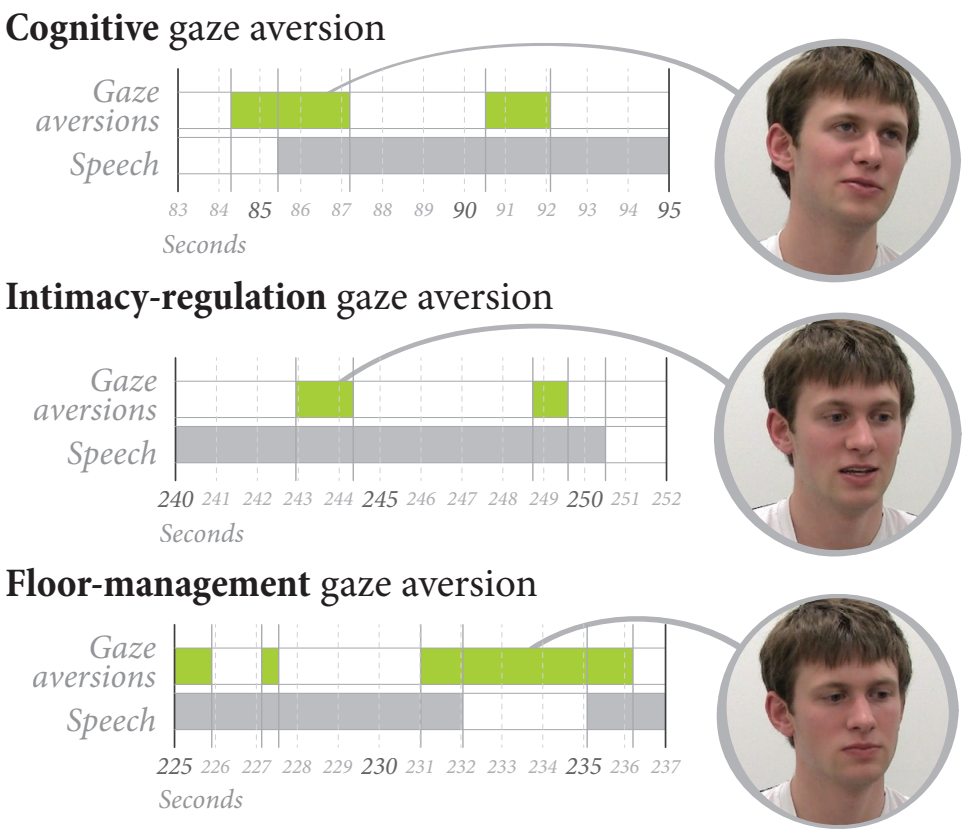

Figure 4.3: Three examples of gaze aversions from the human-human conversational data. Top: an upward *cognitive* gaze aversion at the beginning of a question response. Middle: a short *intimacy-regulation* gaze aversion while speaking. Bottom: a *floor-management* gaze aversion during a pause.

aversion according to CS. These gaze aversions are only planned for times when cognitive and turn-taking aversions are not already planned. Also, intimacy gaze aversions are prohibited near the end of utterances, t_e, so that virtual agents can appropriately pass the floor by maintaining mutual gaze.

Example Simulation

Figure 4.4 illustrates a simulation of the gaze aversion behaviors produced by our controller. In this example, two agents, A1 and A2, are having a conversation. Both are using the gaze aversion controller. A1 asks a question constructed from two utterance parts with a pause in between. A turn-taking gaze aversion is planned and executed around the start of the second utterance in order to hold the conversational floor. While A1 is listening, it occasionally looks away to regulate the intimacy of the conversation. Upon recognizing

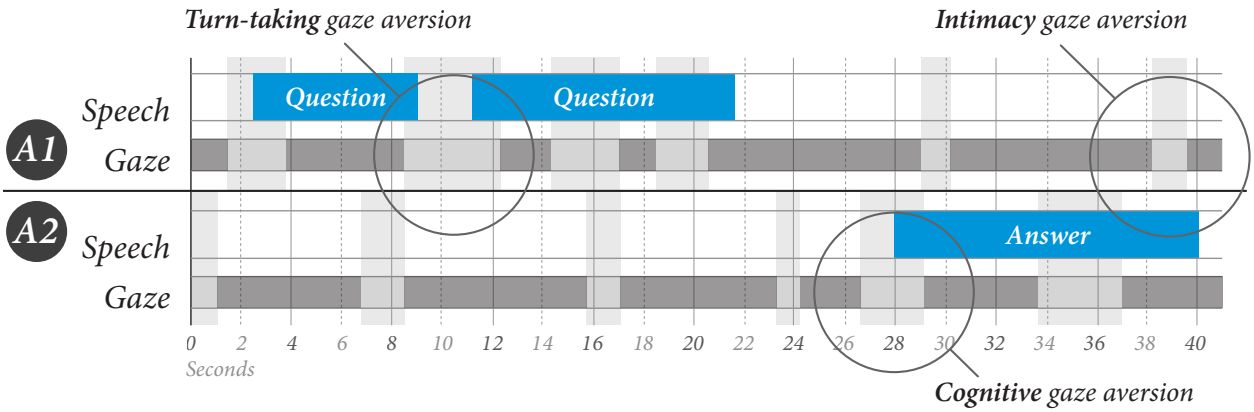

Figure 4.4: Gaze aversions created by the controller for two agents in conversation. Dark gray intervals on the gaze stream indicate periods of gazing toward the interlocutor, and light gray intervals indicate gaze aversions.

A1's question, A2 plans to give its response, which has been tagged with a cognitive "thinking" event at its beginning. The gaze controller plans and executes a cognitive gaze aversion around the beginning of the utterance to express this thinking. All other gaze aversions in the example have been similarly produced by the controller to achieve one of the three conversational functions.

Evaluation

In order to test the effectiveness of gaze aversion behaviors generated by the controller, this section presents the design and results of a laboratory experiment in which virtual agents conversed with human participants. In this experiment, participants interacted with four different virtual agents in four conversational tasks, each of which was designed to test a different conversational function of gaze aversion.

All gaze aversions in this study were executed using the gaze shift model described in the previous chapter, with a moderate amount of head movement. Head alignment was high as the agent oriented its gaze back to the interlocutor, in accordance with the previous chapter's finding that high head alignment increases people's feelings of affiliation with agents.

Hypotheses

Four hypotheses were developed to test how agents might use the gaze aversion behaviors generated by the aversion controller to achieve conversational functions. The first two

hypotheses focus on the cognitive function, the third on the intimacy-modulation function, and the fourth on the turn-taking function.

Hypothesis 1 — A virtual agent averting its gaze while not currently speaking will be perceived as *thinking*, whereas an agent that does not avert its gaze will not elicit this impression.

This hypothesis is derived from research which has shown how gaze aversion of human speakers is recognized by listeners as the speaker's attempt to disengage attention from the listener's face in order to put cognitive effort into organizing a new utterance (Argyle and Cook, 1976; Doherty-Sneddon and Phelps, 2005).

Hypothesis 2 — Virtual agents that display gaze aversions at the start of utterances will be rated as being more *thoughtful* and creative than virtual agents that do not display gaze aversions.

People will not only recognize that the agent is thinking when it is looking away, but also that it is averting its gaze in order to better construct a creative and thoughtful response.

Hypothesis 3 — Virtual agents that display periodic gaze aversions while listening will increase a human interlocutor's comfort and elicit more *disclosure* than agents that do not display gaze aversions.

Eye contact is one of the key factors in feelings of intimacy between people and too much of it results in highly uncomfortable levels of intimacy (Argyle and Cook, 1976). This hypothesis posits that using gaze aversions appropriately will mediate this potentially negative outcome.

Hypothesis 4 — Virtual agents that display gaze aversions during pauses will be perceived as *holding the floor* and will be interrupted less than agents that do not display gaze aversions.

Gaze has been previously shown to be important in regulating conversational turn-taking (Kendon, 1967). By averting its gaze at a pause in speech, this hypothesis posits that the virtual agent will be able to hold the conversational floor, whereas making eye contact during this pause will result in the agent being interrupted.

Participants

Twenty-four participants were recruited for this study (12 females and 12 males), aged between 18 and 45 ($M = 23$, $SD = 6.82$). All participants were native English speakers and were recruited from the University of Wisconsin–Madison campus through postings on physical and online bulletin boards and through canvassing techniques on a University campus.

Figure 4.5: The four agents used in the virtual agent evaluation: Norman, Jasmin, Lily, and Ivy. Norman, Jasmin, and Lily are performing gaze aversions in different directions, while Ivy is maintaining mutual gaze with her interlocutor.

Study Design

The experiment involved a single independent variable, *gaze aversion condition*, with three conditions varying between participants. One condition involved the virtual agents using gaze aversions generated by the controller described in the previous section, referred to as the *good timing* condition. The other two conditions were baselines for comparison. The first baseline was a *static gaze* condition in which the virtual agents did not employ any gaze aversions. The second baseline was a *bad timing* condition in which the virtual agent employed just as many gaze aversions as in the *good timing* condition but with reverse timings. When the gaze controller indicated that a gaze aversion should be made, the *bad timing* model engaged a mutual gaze shift, and vice versa. This third condition was included as a baseline to show that both the presence and the timing of gaze aversions are important for achieving positive social outcomes.

Separate tasks were designed to test each hypothesis, each using a different virtual agent (Figure 4.5). Participants were randomly assigned to one of the three gaze aversion conditions, which was held constant for all four tasks (eight participants per condition). Tasks were presented in random order.

Task 1

The first task was designed to test Hypothesis 1, the hypothesis that virtual agents that display gaze aversions would be perceived as "thinking." The participant was told that the virtual agent, Norman, was training to work at a help desk in a campus library. The

participants were given five library-related questions to ask Norman. They were instructed to ask each question and listen to the response. Norman would pause for 4 to 10 seconds (randomly determined) before answering each question. Participants were instructed to ask a question again if they thought Norman did not understand or was not going to answer.

The primary measure was the time participants waited for Norman to respond to questions before interrupting him to ask the question again. Specifically, the measurement was the time from the end of each question until either the start of the question being asked again (if interrupting the agent) or the start of the agent's response (if not interrupting the agent).

This task deliberately included an agent with an abstract design that minimally elicits attributions of intent or thought in order to ensure that the agent's gaze aversions were solely responsible for the impression of thinking, unconfounded from any other animation variables.

Task 2

The second task was designed to test Hypothesis 2, the hypothesis that virtual agents that avert their gaze before they start speaking would be seen as more thoughtful than if they did not use gaze aversions. For this task, participants were instructed to ask the agent, Jasmin, a series of five common job interview questions. Jasmin was programmed to respond with answers taken from real-world job interviews.

Participants rated each response immediately after it was given on four seven-point rating scales, capturing participants' subjective impressions of the agent's responses in the job interview. These scales measured the perceived thoughtfulness, creativity, disclosure, and naturalness of each response. In the analysis, scales were combined into a single broad indicator of *thoughtfulness*. Internal consistency was excellent for this measure (Cronbach's $\alpha = .903$).

Task 3

The third task was designed to test Hypothesis 3, the hypothesis that participants would disclose more to a virtual agent that averted its gaze at regular intervals while listening to regulate intimacy. In this task, participants spoke to an agent named Lily, who was introduced as training to be a therapist's aide who would conduct preliminary interviews with incoming clients. Lily asked the participant a series of five questions of increasing intimacy, and participants were instructed to respond with as much or as little detail as

they wished. Questions ranged in intimacy, from "What do you like to do in your free time?" to "What is something you would like to accomplish before dying?"

The primary measure for the third task was the *degree of self-disclosure*, specifically the breadth of disclosure. Breadth of disclosure was obtained using a word count of participants' responses to Lily's questions. Word count has been validated as an appropriate measure of disclosure in previous research on how computers can be used to elicit self-disclosure from people (Moon, 2000).

Task 4

The fourth task was designed to test Hypothesis 4, the hypothesis that virtual agents that display gaze aversions during pauses in speech will be able to effectively hold the conversational floor. Participants were provided with a list of five questions to ask a virtual agent named Ivy, with the goal of getting to know each other. Participants were instructed to ask each question, listen to Ivy's response, and then reciprocate with their own response to the same question. Ivy's responses had two parts, separated by a pause between 2 and 4 seconds in length (randomly determined). If participants started speaking during the pause, Ivy refrained from giving the second part of her response.

The primary measure of the fourth task was the time participants waited for Ivy to be silent during a pause in her speech before taking a speaking turn of their own. The measurement was the time between the end of the first part of the virtual agent's utterance and either the start of the participant's speech (if interrupting the agent) or the start of the agent's second utterance part (if not interrupting the agent).

Setup & Procedure

After giving informed consent, the experimenter led each participant into the study room and gave a brief introduction to the experiment. The participant sat in a chair approximately six feet away from a large screen on which the life-size virtual agent was projected (Figure 4.6). A wireless touchpad was used as a button to begin each conversational task, and a Kinect microphone was used for capturing speech. The Microsoft Speech Platform[3] was used for speech recognition in combination with a custom dialogue manager specific to each task. After completing all four tasks, the participant responded to a survey of demographic characteristics and was debriefed. The study took approximately 30 minutes, and each participant was given $5 as compensation.

[3]http://msdn.microsoft.com

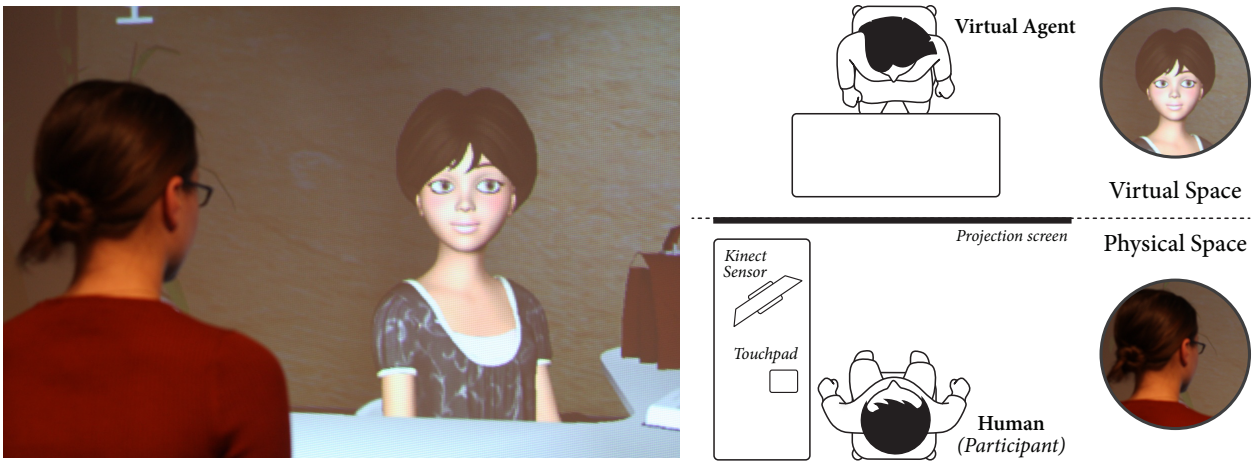

Figure 4.6: An experimenter demonstrating the interaction with the virtual agent on a life-size projected display (left) and the physical setup of the experiment (right).

Results

A mixed-design analysis of covariance (ANCOVA) was performed to assess how agent gaze aversion behaviors affected the dependent variable for each task. Participant gender was included as a covariate to control for gender differences. Question ID was included as a covariate to control for learning effects. Planned comparisons were carried out as *apriori* contrast tests using Scheffé's method.

Hypothesis 1 – The evaluation supported this hypothesis. The time given to the virtual agent before interrupting was significantly higher when the agent used proper gaze aversion with good timing rather than bad timing, $F(1, 110) = 5.06, p = .027$, or with no gaze aversion at all, $F(1, 110) = 12.71, p < .001$.

Hypothesis 2 – The evaluation did not support this hypothesis. Participants' ratings did not differ for virtual agents using proper gaze aversion over agents using gaze aversion with bad timing, $F(1, 110) = 0.0004, p = .98$, or with no gaze aversion at all, $F(1, 110) = 0.002, p = .97$.

Hypothesis 3 – The evaluation supported this hypothesis. Virtual agents using gaze aversions with good timing elicited significantly more disclosure from participants than when their gaze aversions were badly timed, $F(1, 110) = 4.48, p = .037$, or when they used no gaze aversion, $F(1, 110) = 4.25, p = .042$.

Hypothesis 4 – The evaluation partially supported this hypothesis. The time given to the virtual agent during its pause before interrupting was marginally higher when the agent used properly-timed gaze aversion than when its gaze aversions were badly timed, $F(1, 110) = 3.64, p = .059$, and significantly higher than when it did not use gaze aversion

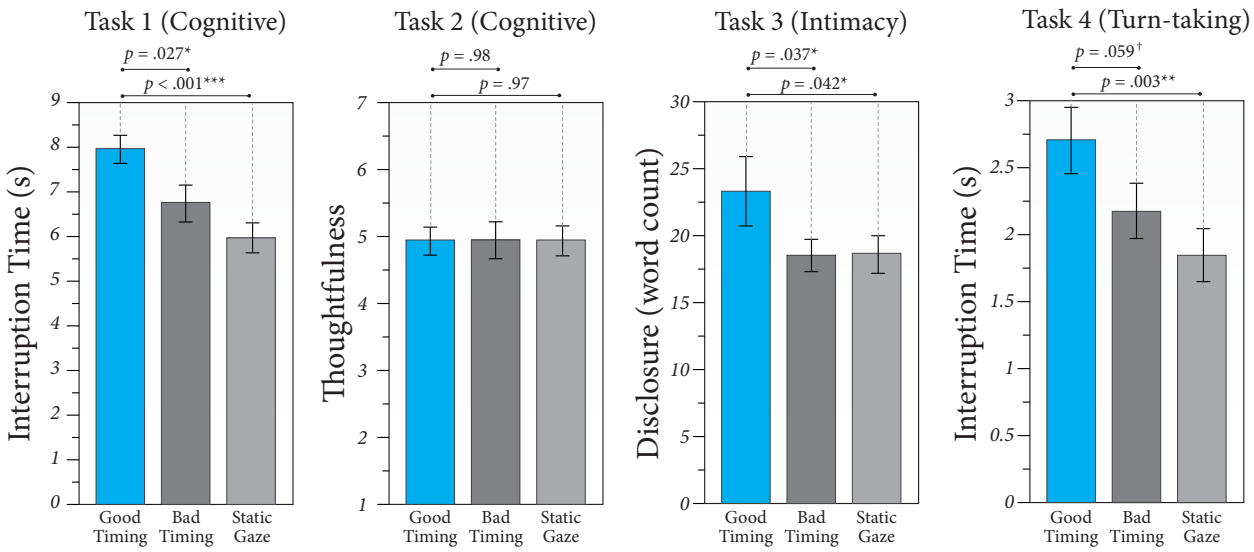

Figure 4.7: The results of the evaluation. Virtual agents that displayed gaze aversions with appropriate timings successfully conveyed the impression that they were "thinking," elicited more disclosure from participants, and were better able to hold the conversational floor during breaks in speech. (†), (*), (**), and (***) denote $p < .10$, $p < .050$, $p < .010$, and $p < .001$, respectively.

at all, $F(1, 110) = 9.48, p = .003$. All of the primary results are illustrated in Figure 4.7.

Discussion

The goal of this experiment was to show that virtual agents can use gaze aversions to achieve different conversational functions, such as indicating cognitive processing occurrences, modulating intimacy, and facilitating conversational turn-taking.

Virtual agents that displayed gaze aversion behaviors generated by the controller described in this chapter were partially successful in achieving the cognitive conversational function of gaze aversions. As shown in Task 1, virtual agents successfully used gaze aversion to indicate that they were engaged in a form of cognitive processing with a response forthcoming and thus delayed interruptions by a human interlocutor. However, as shown in Task 2, using gaze aversions before responses did not affect how thoughtful participants thought those responses were. A possible explanation for this result is that while participants responded *behaviorally* to the agent using gaze aversion to achieve conversational functions, these cues failed to elicit explicit attributions of thought when participants were asked to reflect on the interaction afterwards.

Virtual agents displaying gaze aversion behaviors generated by the current controller were successful in eliciting more disclosure from participants. Measurements of the breadth

of participants' responses in Task 3 show that participants disclosed more when the virtual agent periodically looked away from them with appropriate timings than when the agent did not look away or looked away at inappropriate times.

Finally, virtual agents displaying gaze aversion behaviors generated by the controller were successfully able to regulate conversational turn-taking. By averting their gaze at the appropriate time in Task 4, virtual agents more effectively held the conversational floor than when they used gaze aversion at inappropriate times or not at all.

Although the gaze aversion mechanisms modeled and implemented on virtual agents were found to be quite effective, it is not a trivial matter to implement these behaviors on a physical robot with very different affordances. The next section presents efforts to extend the gaze aversion implementation to a robot platform, followed by an evaluation of that robot's performance in similar conversational tasks as those explored for virtual agents above.

4.4 Gaze Aversion in Robots

Implementation

Previous sections presented a human-human data collection leading to a gaze aversion model consisting of precise spatial and temporal parameters—including length, timing, and frequency—in relation to conversational functions and speech. This model generated appropriately timed gaze aversions for virtual agents during conversations with people. However, the above work did not consider the challenges of applying the model to an autonomous physical robot, nor did it provide evidence that gaze aversions can also be effectively used by robots. This section presents the design of gaze aversion motions for humanlike robots without articulated eyes, secondary head motions for achieving mutual gaze and lifelikeness, techniques for combining the different head motions, and the overall system implementation for expressing conversational gaze aversion on the Nao robot platform.

A fundamental difference between the virtual agents utilized in the previous section and the robots used in this section is the difference in physical affordances available for carrying out gaze motions. Overall differences in geometry mean that gaze motions and control laws need to be adapted (Pejsa et al., 2013). Even more critically, the robots lack articulated eyes and must rely solely on head motions to convey gaze motions, making it unclear if they are capable of eliciting the same positive conversational outcomes found in the virtual agent work. Previous research points to the possibility of such capabilities, including work

which has shown that people are capable of recognizing a robot's gaze according to its head orientation (Imai et al., 2002) and that robots can use head motions alone to gaze effectively in a storytelling scenario (Mutlu et al., 2006). The physically co-located robot must also move its head to track users' faces, motions that are not required of a virtual agent due to the Mona Lisa gaze effect (Al Moubayed et al., 2012).

This section presents the development of an autonomous system to appropriately display gaze aversion behaviors on the Nao platform. This system includes methods to overcome three key technical challenges in adapting the gaze aversion model to a head controller for conversational robots: adapting the human movements to a robot with non-human affordances, making the movements appear lifelike and intentional, and integrating the gaze aversion movements with other head movements. The system generates three distinct types of head movements with different purposes: face-tracking movements to engage in mutual gaze, idle head motion to increase lifelikeness, and purposeful gaze aversions to achieve conversational functions. To adapt gaze movements to a robot platform without articulated eyes, head movement is substituted for combined eye and head rotations. To make the movements appear lifelike and intentional, aversion control is combined with face tracking and structured random movements to create idle motion. To realize these behaviors in an autonomous system, they are implemented in a predictive filtering framework that affords graceful combination of multiple goals and effective reaction to external events.

Aversion Movement Design

A robot without articulated eyes must rely on head motion alone to carry out gaze aversion motions. There are two considerations when designing these motions: the magnitude and the dynamics.

The appropriate magnitudes for gaze-averting head motions were determined in an iterative process to achieve a natural subjective appearance. The goal was to generate head movements that are not too extreme yet clearly serve to avert the robot's gaze away from the user. The system generates vertical gaze-averting head movements with a magnitude of approximately 20 degrees, with horizontal and downward head movements at 28 and 22 degrees respectively. For intimacy-modulating gaze aversions, these angles are scaled by a factor of 0.4, because these gaze aversions occur quite frequently and were observed in the human-human data to be employed with more subtle eyes-only motions.

For the generated gaze aversions, the system recreates the head velocity profile identified in neurophysiological research on human gaze shifts (Kim et al., 2007). This profile resembles standard ease-in/ease-out curves found in animation, in which the head rotation

starts slow, speeds up during the bulk of the shift, then slows down before coming to a halt. The velocity profile was implemented using programmatically-defined bezier curves. Care was taken to respect both the upper and effective lower limits of the motors' rotational speeds, ensuring smooth motions throughout the robot's gaze shifts.

Achieving Mutual & Lifelike Gaze

In order to interact effectively with humans and accomplish a general feeling of lifelikeness, the robot must use its head for not only gaze aversion, but also mutual gaze. To engage in mutual gaze, the robot must track the user's face. For this system, a face-tracking algorithm was employed that utilizes a Microsoft Kinect situated behind the robot. The robot continuously adjusts its head rotation based on the results of the face tracking algorithm, polled every $200ms$.

At this point the robot is capable of executing gaze aversions and face tracking, however a problem remains in that the transition between statically gazing at the user and engaging in a gaze aversion is still quite abrupt. When the user is sitting still and the robot is not currently engaged in a gaze aversion motion, the robot's head becomes motionless and loses its sense of liveliness. To address this problem, the system generates a small amount of idle head motion for the robot to execute at all times. This idle motion was implemented by adding a small amount of structured noise, generated by a Perlin noise function (Perlin, 1995), to the results of the face tracking algorithm. A Perlin noise function generates band-limited, pseudorandom signals that are useful for emulating biological forms and motions.

Combining Behaviors

The robot uses its head for three different types of gaze motions: gaze aversions, face tracking, and idle motion. At any point in time, the robot has multiple distinct goals for its target head rotation. These goals must be combined in a natural manner that maintains an ability to be reactive to new goals. To solve this problem, the system utilizes a Kalman filter (Kalman, 1960), a linear predictive filter used for estimating the state of a system given past states and target goals. The filter predicts the appropriate motion by blending estimated trajectories generated from the current state and current goals, and corrects these estimates as the goals change. The filter gains were chosen empirically to provide an appropriate balance between smoothness and reactivity.

Target head rotations from the three types of motions are streamed through this filter and combined into a single head rotation signal, allowing for graceful transitions between

motions. Without this filter, head motions would have to be serially generated and completed, e.g., a head tracking motion to the face of the user would have to be completed before a gaze aversion motion could be executed. The Kalman filter solves this problem by blending estimated trajectories over time, resulting in smooth, interruptible motions.

System Integration

Gaze aversion behaviors were implemented on the Nao using two controllers: a high-level *gaze controller* to plan the timing and direction of gaze aversions and a *head controller* to physically execute the gaze shifts.

The head controller is situated at the end of an overall system pipeline that also includes a speech recognition system and a dialogue manager. A Microsoft Kinect with the Microsoft Speech Platform[4] is utilized for speech recognition. Recognized speech from the user is passed to a dialogue manager that associates a semantic tag with the utterance and plans the agent's speech accordingly. For example, if the dialogue manager receives a recognized question, it will produce the associated answer. Robot speech is generated in the system by playing pre-recorded audio files, but the Nao's built-in text-to-speech system can also be used.

The dialogue manager sends upcoming speech events and the current conversational state to the gaze controller. As the gaze controller receives these inputs, it continuously plans gaze aversion motions to be executed by the head controller. The exact timings of the gaze aversions are drawn from the parameter distributions reported in Table 4.1 and visually depicted in Figure 4.3. The direction of each gaze aversion is similarly generated according to its function and the directional likelihoods presented in Figure 4.2.

Also included in the overall system are a wireless touchpad and the Nao's chest light. The wireless touchpad is used by the user to signal that the NAO can begin the next phase of the interaction, depending on the context. For example, after the user is done giving an open-ended response to a question posed by the Nao, the user would touch the touchpad to signal to the Nao that it can move on to its next question. The Nao's chest light is used to signal to users the beginning (green) and end (red) of the interaction as well as when the Nao is ready to begin a new interaction phase (blue). The chest light was not used during core interaction sequences while gaze aversions were being displayed (i.e., while asking a question, answering a question, or listening to a user's utterance).

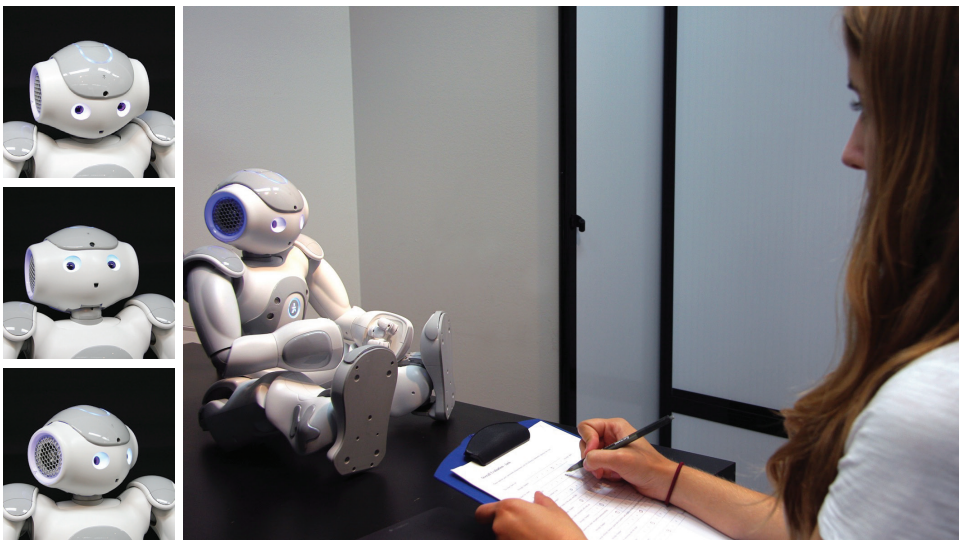

Figure 4.8: A human conversational partner interacting with the NAO robot. Three example gaze aversion directions implemented for the NAO are shown: down, up, and to the side.

Evaluation

This section presents an experimental evaluation to test user perceptions of the generated gaze aversions and their effectiveness in enabling robots to achieve positive conversational functions in human-robot conversations. In this study, participants interacted in multiple conversational tasks with two Nao robots. Each task involved the participant either asking questions to the Nao or responding to questions posed by the Nao.

The study was designed to test two primary hypotheses: first, that the designed gaze motions would be perceived by participants as intentional gaze aversions rather than meaningless motions, and second, that a robot using gaze aversions appropriately could accomplish each of the three conversational functions of gaze aversion, including the cognitive, intimacy, and floor management functions. In the experiment, human participants interacted with two robots in four conversational tasks (Figure 4.8). The first task was designed to test the first hypothesis, and the other three were designed to test the second hypothesis, one for each conversational function. This evaluation demonstrates how the gaze aversions employed by the robot were perceived as intentional, served to make the robot appear more thoughtful, and helped the robot manage the conversational floor.

[4]http://msdn.microsoft.com

Study Design

The experiment involved a single independent variable, *gaze aversion behavior*, with three conditions varying between participants. One condition involved the robots using gaze aversions generated by the controllers described in the previous section, called the *good timing* condition. The other two conditions served as baselines for comparison. The first baseline was a *static gaze* condition in which the robots did not employ any gaze aversions. This baseline was included to demonstrate the importance of generating gaze aversion motions regardless of the timing. The second baseline was a *bad timing* condition in which the robot employed just as many gaze aversions as in the *good timing* condition but with reverse timings. More precisely, the *bad timing* condition produced a gaze shift toward the participant—to engage in mutual gaze—every time the *good timing* condition would have triggered a gaze aversion. Similarly, a gaze aversion is triggered every time the *good timing* condition would have produced a shift toward mutual gaze. This third condition was included as a baseline to show that not only the presence, but also the timing of gaze aversions is important for achieving positive social outcomes. Participants were randomly assigned to one of the three gaze aversion conditions, which was then held constant for all tasks (ten participants per condition). Regardless of condition, the robot always tracked the participant's face and utilized a small amount of idle head motion.

Participants

Thirty participants were recruited for this study (15 females and 15 males)—aged between 20 and 38 ($M = 22.90$, $SD = 4.27$)—from the University of Wisconsin–Madison campus. Participants were primarily University students with a range of fields of study, including biology, computer science, economics, and communication.

Hypotheses, Tasks, & Measures

This evaluation involved two hypotheses related to participants' perceptions of gaze aversions carried out by a robot and how robots might use gaze aversion behaviors to achieve each of the three conversational functions. The second hypothesis is split into three sub-hypotheses, one for each conversational function. Separate tasks and measures were created to test each hypothesis and sub-hypothesis. For clarity, each hypothesis is presented together with the task designed to test that hypothesis as well as the primary measure in that task.

Hypothesis 1 — Well-timed robot gaze aversions will be seen as *intentional* motions used to engage in some cognitive processing, rather than randomly generated motions

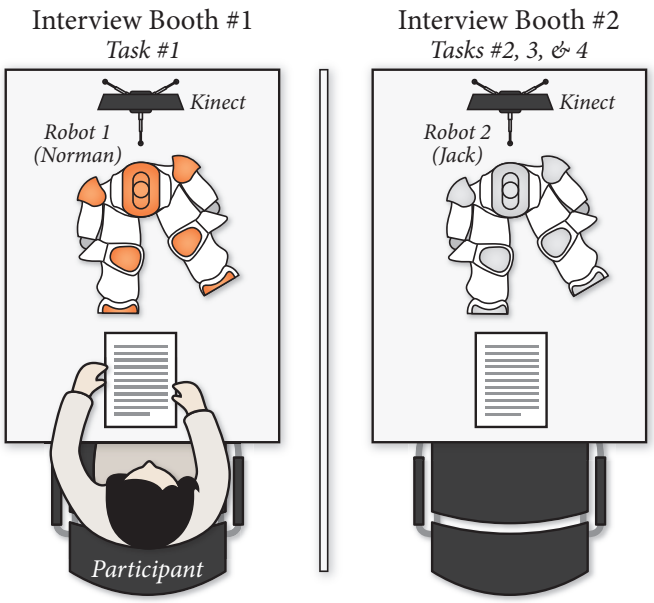

Figure 4.9: A diagram of the physical setup of the human-robot interaction experiment. Participants interacted with Norman for the first task, and Jack for the other three tasks.

without meaning.

This hypothesis is derived from research that has shown how speaker gaze aversion is recognized by listeners as the speaker's attempt to disengage attention from the listener's face in order to put cognitive effort into organizing a new utterance (Doherty-Sneddon and Phelps, 2005; Glenberg et al., 1998). Finding support for this hypothesis supports the premise that the robot's head movements, which were designed to convey gaze aversion, are indeed perceived as gaze aversions and not as random movements.

Task 1 — The participants were told that the robot was training to work at a library help desk and given five library-related questions to ask the robot. They were instructed to ask each question and listen to the robot's response. The robot paused for 7 to 10 seconds (randomly determined to decrease predictability) before answering each question. Participants were instructed to ask a question again if they thought that the robot did not understand.

Measure 1 — The primary measure was the time participants waited for the robot to respond to questions before interrupting it to ask the question again. Based on the hypothesis, it was expected that participants would give the robot the most time to respond when it was using gaze aversions in the *good timing* condition, implying that the gaze aversion during its long pause was perceived as an intentional motion to formulate responses to the participants' questions.

Hypothesis 2 — Gaze aversions generated by the model and employed by the robot will enable it to achieve the positive conversational functions observed in human-human interaction. This hypothesis has three sub-hypotheses, one each for the cognitive, intimacy-modulating, and floor management functions of gaze aversion.

Hypothesis 2a — When a robot utilizes cognitive gaze aversions at the start of answers to questions, its answers will be rated as being more *thoughtful* and creative than when it does not display gaze aversions or displays gaze aversions with inappropriate timings.

This hypothesis is derived from research which has shown that people use gaze aversions to signal that cognitive processing is occurring and to create an impression that deep thought or creativity is being undertaken in formulating their utterance (Argyle and Cook, 1976).

Task 2a — Participants engaged in a mock job interview with the robot, in which the participant was the interviewer and the robot was the interviewee. Participants were instructed to ask a series of five common job interview questions, and the robot was programmed to respond with answers taken from real-world job interviews.

Measure 2a — Participants rated each response immediately after it was given on four seven-point rating items. A single scale of *thoughtfulness* was constructed from the four items, including ratings on perceived thoughtfulness, creativity, disclosure, and naturalness of each response. Internal consistency was good for the items in this scale (Cronbach's $\alpha = .852$). The highest overall ratings were expected to be given by participants when the robot used cognitive gaze aversions at the start of its responses according to the model.

Hypothesis 2b — Robots that display periodic gaze aversions while listening will increase a human interlocutor's comfort and elicit more *disclosure* than robots that do not display gaze aversions or display gaze aversions with inappropriate timings.

Eye contact is one of the factors that shape feelings of intimacy between people. Too much of it results in uncomfortable levels of intimacy (Argyle and Cook, 1976). This hypothesis posits that using gaze aversions appropriately will alleviate this potentially negative outcome.

Task 2b — The robot in this task was introduced to the participant as training to be a therapist's aide that would conduct preliminary interviews with new clients. During the task, the robot asked the participant a series of five questions with increasing levels of intimacy, ranging from "What do you like to do in your free time?" to "What is something you don't like about yourself?," and participants were instructed to respond with as much or as little detail as they wished.

Measure 2b — The primary measure for this task was the *breadth of self-disclosure*, which was obtained using a word count of participants' responses to the robot's questions. It was

Figure 4.10: An experimenter demonstrating the conversational interaction with Jack, one of the NAO robots.

expected that participants would disclose more to a robot that used appropriately timed intimacy-modulating gaze aversions while listening.

Hypothesis 2c — Robots that display gaze aversions during pauses will be perceived as *holding the floor* and will be interrupted less than robots that do not display gaze aversions or display gaze aversions with inappropriate timings.

Gaze has been previously shown to be important in regulating conversational turn-taking (Kendon, 1967). By averting its gaze at a pause in speech, this hypothesis posits that a robot will be able to hold the conversational floor, whereas making eye contact during this pause will result in the robot being interrupted.

Task 2c — Participants were provided with a list of five questions to ask the robot, with the goal of getting to know each other. Participants were instructed to ask each question, listen to the robot's response, and then reciprocate with their own response to the same question. The robot's responses had two parts, separated by a pause between 2 and 4 seconds in length (randomly determined to decrease predictability). If participants started speaking during the pause, the robot refrained from giving the second part of its response.

Measure 2c — The primary measure of this task was the time participants waited for the robot to be silent during the pause in its speech before interrupting or before the robot successfully produced the second part of its utterance. It was expected that participants would wait longer before interrupting the robot—if interrupting at all—when it used appropriate turn-taking gaze aversions during its speech pauses, as specified by the model.

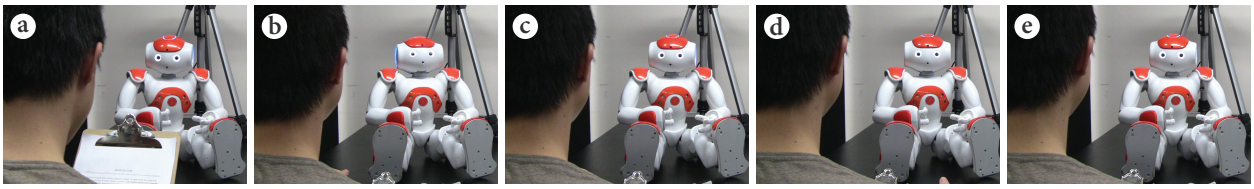

Figure 4.11: Example of a single question-answer sequence. (a) The participant reads a question from his list in the preparation phase. (b) The participant looks toward the robot, and the robot engages an upward cognitive gaze aversion at the start of its answer. (c) The robot looks back toward the participant during its utterance. (d) The robot engages in a sideways intimacy-modulating gaze aversion. (e) The robot looks back toward the participant to complete its utterance.

Setup & Procedure

An orange NAO, given the name Norman, was used for the first task that tested the perceived intentionality of generated gaze shifts. A gray NAO, given the name Jack, was used for the other three tasks. A separate robot was used for the first task due to the unnaturally long pauses present in that task. It would be undesirable to have negative perceptions associated with these long response times to become associated with the robot during the other three tasks in which conversations proceeded more naturally. Each robot had a unique voice, implemented as pre-recorded audio files from separate male voice actors modulated in pitch to better fit the design of the robot. Participants sat in a chair approximately three feet away from the robot they were currently interacting with. The robot's position was carefully chosen to be at approximately eye-level with the participant and at a comfortable social distance. A black dividing wall was placed between the two robots so that participants could only see a single robot at a time. The setup resembled interview booths commonly used in job fairs (Figure 4.9 and Figure 4.10).

Each question-answer interaction sequence, illustrated in Figure 4.11, began with a *preparation phase*—a common element of human-human interviews—in which the participant looked down and read a question from the list. During this phase, the robot displayed idle gaze movements while staying focused on the participant. Toward the end of the question, the participant redirected their attention toward the robot, at which point the robot began its response, displaying gaze aversions based on the experimental condition. Between gaze aversions, the robot displayed subtle random motions to increase lifelikeness. Upon completing its utterance, the robot always focused back on the participant, displaying subtle idle motion. At this point, the participant looked down to the list, beginning the preparation phase for the next question-answer sequence.

After obtaining informed consent, the experimenter led each participant into the study room and gave a brief introduction to the experiment. The participant first completed the single task with Norman and was then relocated to sit in front of Jack for the other three tasks, which were completed in random order. After completing all four tasks, the participant responded to a survey of demographic characteristics and was debriefed. The study took approximately 30 minutes, and each participant received $5.

Results

A mixed-design analysis of covariance (ANCOVA) was employed to assess how robot gaze aversion behaviors affected the dependent variable for each task. Participant gender was included as a covariate to control for gender differences. Question ID—five in each task—was included as a covariate to control for learning effects. Question ID was nested within participant ID and modeled as a random effect. This resulted in five observations per participant and 150 total observations in each measure. Planned pairwise comparisons described in our hypotheses were carried out using Tukey's HSD test. A summary of the primary results is presented in Figure 4.12.

Hypothesis 1 predicted that gaze aversions employed by the robot would be seen as *intentional* motions used to engage in some cognitive processing, rather than randomly generated motions without meaning. The study supported this hypothesis. The time given to the robot before interrupting was significantly higher when the robot used proper gaze aversion with good timing rather than bad timing, $F(1, 142) = 7.72, p = .017$, or no gaze aversion at all, $F(1, 142) = 5.99, p = .041$.

Hypothesis 2a predicted that when a robot utilized cognitive gaze aversions at the start of answers to participant-provided questions, those answers would be rated as being more *thoughtful* and creative than when the robot did not display gaze aversions or displayed gaze aversions with inappropriate timings. The study supported this hypothesis. Participants assigned higher ratings of thoughtfulness to utterances produced by robots using well-timed gaze aversions over those produced by robots using poorly-timed gaze aversion, $F(1, 142) = 27.97, p < .001$, and over those produced by robots using static gaze, $F(1, 142) = 10.19, p = .005$.

Hypothesis 2b predicted that robots that displayed periodic gaze aversions while listening would increase a human interlocutor's comfort and elicit more *disclosure* than robots that did not display gaze aversions or displayed gaze aversions with inappropriate timings. The study did not support this hypothesis. The robot using gaze aversions with good timing elicited no more disclosure—measured as word count per response—from participants

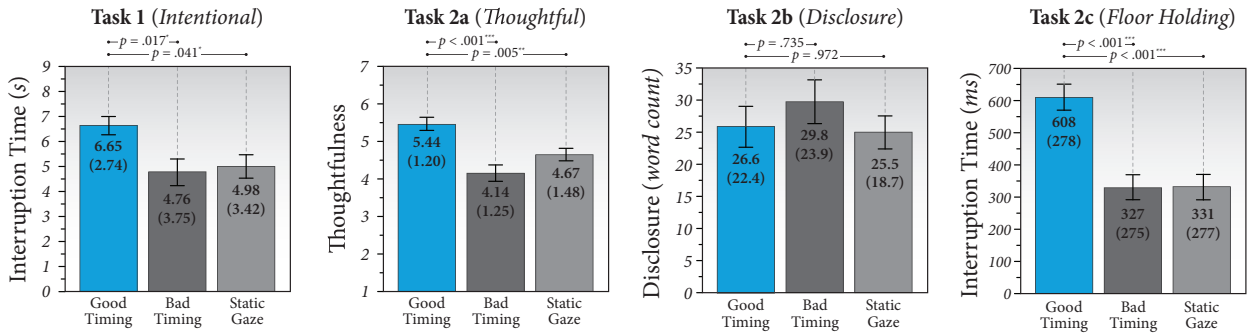

Figure 4.12: The results of the evaluation. Robot gaze aversions generated by the model were perceived as intentional and enabled the robot to appear more thoughtful and effectively manage the conversational floor. (*), (**), and (***) denote $p < .050$, $p < .010$, and $p < .001$, respectively. Means and standard deviations (in parentheses) are provided inside each bar.

than when its gaze aversions were badly timed, $F(1, 142) = 0.56, p = .735$, or when it used no gaze aversion at all, $F(1, 142) = 0.05, p = .972$.

Hypothesis 2c predicted that robots that displayed gaze aversions during pauses would be perceived as *holding the floor* and would be interrupted less than robots that did not display gaze aversions or displayed gaze aversions with inappropriate timings. The study supported this hypothesis. The time given to the robot during its pause before interrupting was significantly higher when the robot used properly-timed gaze aversion than when its gaze aversions were badly timed, $F(1, 142) = 25.53, p < .001$, or when it did not use gaze aversion at all, $F(1, 142) = 24.93, p < .001$.

Discussion

The goal of the evaluation was to show that gaze aversions generated by robots are perceived as intentional and meaningful motions, and that robots can use gaze aversions to achieve three conversational functions: signaling that cognitive effort is being spent in producing thoughtful utterances, modulating the overall intimacy level of the conversation, and facilitating floor management.

As shown in the first task, the robot's gaze aversions were perceived as intentional motions to engage in some sort of processing. Participants gave the robot significantly more time to formulate its response to a question when the robot used gaze aversion appropriately. Supporting the first hypothesis was important for interpreting the rest of the results, as it can be concluded that the robot's head movements, which were designed to convey gaze aversion, were indeed perceived as gaze aversions and not as random

movements.

As shown in the second task, in which a robot produced responses to a participant's interview-style questions, a robot that uses gaze aversions appropriately is capable of achieving the cognitive conversational function of producing utterances that are perceived as thoughtful and creative. Participants subjectively rated the robot's responses more highly when it used appropriate gaze aversion behaviors.

Robots using well-timed gaze aversions were not successful in eliciting more disclosure from participants, as shown in the third task. A possible explanation is that the intimacy-modulating function of gaze aversion is really about reducing eye-contact and staring and that the precise timing for this type of gaze aversion is not important. Furthermore, the idle Perlin noise motions generated by the robot system in all conditions may have been perceived as small gaze shifts that increased participant comfort without needing more overt gaze aversions. In general, the lack of a significant result is consistent with previous HRI work that also did not find a difference in participant disclosure based on the gaze behavior of a robot (Mumm and Mutlu, 2011b).

Finally, robots using well-timed gaze aversion behaviors were more successful in managing conversational floor. When the robot used gaze aversions appropriately, participants waited longer during the robot's pause before interrupting it to claim the floor. It should be noted, however, that the robot was never able to hold the conversational floor throughout the entirety of its pause in speech (planned to be 2–4 seconds in length) without being interrupted. When using properly timed gaze aversions, the mean time before interruption was 608ms, compared with 329ms and 327ms in the *static gaze* and *bad timing* conditions, respectively. This is still a noteworthy result, as most gaps between turns in human-human conversations are shorter than 400ms, about one-third of them less than 200ms (Wilson and Zimmerman, 1986). Adding other floor-holding behaviors, such as conversational fillers, might enable the robot to be even more effective in holding the floor during its pause. Previous work in HRI has already demonstrated the usefulness of conversational fillers in alleviating users' negative perceptions to long system response times (Shiwa et al., 2008).

This work has a number of design implications for social robots. In general, this work demonstrates the value of gaze aversion as a useful nonverbal conversational cue for designing applications such as an information booth robot that gives suggestions and should appear thoughtful and creative in the production of its utterances or a conversational robot that needs to effectively manage turn-taking with people. That said, gaze aversion motions alone are not sufficient for an autonomous robot system to appear natural and lifelike in its movements. This chapter demonstrated how to combine aversion control with face tracking and structured random movements to create lifelike idle motion. A predictive

filtering framework was very effective for gracefully combining multiple gaze motions with different goals.

4.5 General Discussion

The robot system evaluation demonstrated that robots that use gaze aversions appropriately are capable of achieving the cognitive conversational function of producing utterances that are perceived as thoughtful and creative. This result is especially interesting as it was not observed when evaluating the previous implementation of gaze aversion behaviors for virtual agents. A potential explanation might be found in the increased social presence of robots over virtual agents, making the generated cognitive gaze aversions more salient cues in this evaluation. That said, a follow-up evaluation that directly compares virtual agents and robots in this task would be required to better explain the observed difference in results.

These results have a number of implications for the designers of embodied agent interactions. Designers must consider gaze aversion as more than "lack of eye contact" and instead as a powerful cue that can achieve conversational goals. Agents can use gaze aversions to make their utterances appear thoughtful and creative. If the goal is to elicit disclosure from a human, the agent should use gaze aversion to regulate the intimacy of the conversation. When agents need to pause in their speech, e.g., to process information or plan their next utterance, gaze aversion is an effective strategy to hold the conversational floor and indicate to the human that a new utterance is forthcoming.

The previous section demonstrated how a robot can use gaze aversions effectively in conversations with people, similarly to what was shown for virtual agents earlier in the chapter. However, there are a number of important differences between virtual agents and robots that must be carefully considered when designing behaviors across both modalities, as evidenced by previous work and the difference in results across systems discussed here. It will take careful thought and creativity to develop precise representations of future agent behaviors that achieve targeted conversational functions while still being sufficiently generalizable across virtual and physical platforms.

4.6 Chapter Summary

While previous research has explored how embodied agents can use gaze to achieve positive social outcomes (Chapter 2), a precise account of when agents should avert their gaze from human conversational partners and what social functions these aversions might achieve was

missing. This chapter addressed this knowledge gap from both theoretical and empirical perspectives through the application of existing social-scientific knowledge and a study of human dyadic conversations to design gaze aversion behaviors for embodied agents.

Gaze aversions are commonly associated with negative social outcomes, including discomfort, inattention, and deceit, but in reality they serve a number of important positive conversational functions, including cognitive, intimacy-modulating, and turn-taking functions. This chapter first presented an analysis of human dyadic conversations that informed the development of a gaze aversion controller that can automatically plan and execute appropriately timed gaze aversions for embodied agents. The analysis identified precise spatial and temporal parameters of gaze aversion from a video corpus of human-human interactions.

This computational understanding was then implemented into a gaze controller on a virtual agent platform. A user study was conducted to evaluate the gaze aversion behaviors generated by the controller for their effectiveness in achieving positive conversational functions. The experiment demonstrated that virtual agents using gaze aversions generated by the controller were perceived as thinking, elicited more disclosure from human interlocutors, and effectively managed turn-taking.

This chapter also presented work to implement these behaviors on a humanlike robot system. This implementation was designed to overcome the inherent challenges of adapting a human-based gaze model for a robot without articulated eyes. This chapter presented a head controller for the Nao platform that generates and combines head motions with three purposes: purposeful gaze aversions to achieve conversational functions, face tracking for engaging in mutual gaze with a user, and noisy idle head motions to increase lifelikeness. An evaluation of the designed gaze aversion behaviors generated by this system demonstrated that they are perceived as intentional when expressed by a humanlike robot, and that robots can use gaze aversions to appear more thoughtful and effectively manage the conversational floor.

These results have important implications for designers of human-robot and human-agent interactions. Gaze aversions should be considered as an important and powerful cue for developing natural and effective conversational interactions between humans and embodied agents. However, this model of conversational gaze aversions is not appropriate for more complex interactions involving physical collaboration over a shared task space, in which the agent must be able to effectively distribute its gaze across relevant task objects in addition to its human collaborator. This situation is considered in the next chapter.

5 GAZE COORDINATION

The key to successful social interaction lies in the coordination and alignment of both verbal and nonverbal cues (Clark, 2005; Richardson et al., 2007). This coordination is especially critical in collaborative physical tasks, where coordination must occur across several communicative channels simultaneously. Humans are remarkably adept at this kind of coordination, which arises naturally and unconsciously in interaction. However, artificial embodied agents lack models and mechanisms that would allow them to engage in similar coordination when interacting with people. If an agent could coordinate with users in a humanlike way, the outcomes achievable range from giving users positive perceptions of the quality of the interaction and agent, to practical task outcomes such as higher efficiency and fewer breakdowns. Unfortunately, exactly how gaze coordination arises in everyday human-human social interaction is still not well understood.

Coordination in human-human conversation enables participants to manage speaking turns (Sacks et al., 1974) and to draw each other's attention toward objects of mutual interest using actions such as pointing, placing, gesturing, and gazing (Clark, 2003; Clark and Krych, 2004; Clark, 2005). Through the course of an interaction, interlocutors mimic each other's syntactic structures (Branigan et al., 2000) and accents (Giles and Powesland, 1997), and their bodies even begin to sway in synchrony (Condon and Ogston, 1971; Shockley et al., 2003). These acts of coordination are critical to ensuring that *joint activities*, including conversation and collaboration, flow easily and intelligibly (Clark, 1996; Garrod and Pickering, 2004).

Of particular importance to successful interaction is the coordination of gaze and attention across a shared visual space (Brown-Schmidt et al., 2005; Clark, 1996; Clark and Brennan, 1991; Schober, 1993). *Gaze coordination* has been succinctly defined as a coupling of gaze patterns (Richardson et al., 2009). This coupling does not result from interlocutors explicitly aiming to synchronize their gaze movements, but instead gaze patterns become aligned over time due to the need for coordination in joint activities. Coordinated gaze allows conversational participants to monitor their interlocutor for understanding, regulate the amount of mutual gaze and averted gaze, quickly pass and receive the conversational floor, disambiguate verbal references early in their production, and so on. Key elements of gaze coordination include mutual gaze and joint attention, which serve as primary instruments of prelinguistic learning between infants and caregivers (Baldwin, 1995) and play a crucial role throughout life in coordinating conversations (Bavelas et al., 2002).

Although a large number of studies over the past several decades have investigated gaze behavior and the crucial role it plays in communication, how tightly coordinated gaze behaviors unfold over the course of an interaction is not well understood. For example,

previous work has examined the timings of when people look toward referents—objects to which they or their interlocutors verbally refer (Griffin, 2004; Meyer et al., 2004; Tanenhaus et al., 1995). However, these investigations are generally one-sided, looking at each person's gaze in isolation, and do not capture the intricate coordinative patterns in which partners' referential gaze behaviors interact. Previous work has also investigated gaze alignment, exploring the extent to which conversational partners gaze toward the same targets at various time offsets (Richardson and Dale, 2005; Bard et al., 2009). However, existing research still lacks a more nuanced description of how gaze alignment changes over the different phases of the interaction.

This chapter firstly presents work to develop a deeper understanding of coordinated referential gazing in collaborating dyads. Of particular interest is how the gaze behaviors of two collaborating participants unfold throughout a *reference-action sequence* in which one participant makes a verbal reference to an object in the shared workspace that the other participant is expected to act upon in some way. Data was collected from 13 dyads outfitted with mobile eye-tracking glasses in a sandwich-making task; one participant (the instructor) made verbal references to visible ingredients they would like added to their sandwich while the other participant (the worker) was responsible for assembling those ingredients into the final sandwich (Figure 5.2). This task was chosen to represent collaborative interactions that contain a large number of reference-action sequences. Because these behavior sequences are common and frequent across many kinds of interactions, the results of the analyses discussed in this chapter should generalize beyond the specific sandwich-making task to any interactions that involve reference-action sequences.

Due to the highly dynamic and interdependent nature of the collected data, this project utilized a relatively new analysis technique–*epistemic network analysis (ENA)*—to analyze and visualize the gaze targets of both participants as a complex and dynamic network of relationships. This analysis was shaped by three questions: (1) How do a collaborating dyad's gaze behaviors *unfold* over the course of a reference-action sequence? (2) How does the *alignment* of gaze behaviors shift throughout the different phases of a reference-action sequence? (3) How do coordinated gaze behaviors differ in sequences which include breakdowns and/or *repairs*?

To answer these three research questions, ENA was utilized to conduct three separate analyses of the dyadic gaze data. In the first analysis, ENA was used to characterize different phases of a reference-action sequence, discovering clear differences in gaze behavior at each phase. This analysis also revealed a consistent pattern of gaze behavior that progresses in an orderly and predictable fashion throughout a reference-action sequence. The second analysis explored the progression of gaze alignment between the interacting participants

throughout a reference-action sequence. In general, a common rise and fall in the amount of aligned gaze throughout a sequence was discovered, as well as a back-and-forth pattern of which participant's gaze "led" the other's. The third analysis explored the difference in gaze behaviors arising during sequences with repairs—verbal clarifications made in response to confusion or requests for clarification—versus sequences without such repairs. ENA revealed detectably different patterns of gaze behavior for these two types of sequences, even at very early phases of the sequences before any verbal repair occurs.

In order to produce similar patterns and achieve positive collaborative outcomes, embodied agents will need to similarly coordinate their gaze movements with those of their human users. They can use their embodiments to express social gaze cues, and through the use of gaze tracking technologies, agents have the opportunity to detect and interpret social gaze cues employed by the user. This level of gaze coordination requires a detailed model specifying how an agent should temporally employ its own gaze behaviors contingent on real-time interpretations of the user's gaze. Gaze cues must be simultaneously expressed by the agent and interpreted from the human user in an interactive, dynamic process.

Coordination is particularly important for collaborative virtual agents embedded in immersive virtual reality (VR) applications. However, most VR systems do not yet integrate eye tracking, instead tracking only the user's head position and orientation. Therefore, if gaze coordination is to be practical for VR applications, it is important to consider how the user's head pose might act as a proxy for full gaze tracking, as well as to test whether the desired interaction outcomes are still achievable in VR under this limitation.

The next section (Section 5.1) reviews the relevant background on shared gaze in collaborative interactions between people, as well as in human-agent interactions. Section 5.2 presents a description of the data collection in the sandwich-making task and Section 5.3 presents network analysis, specifically epistemic network analysis (ENA), as an alternative to previous analysis techniques with a number of desirable properties for studying shared gaze in dyads. Section 5.4 covers three analyses of the collected data conducted in ENA.

In Section 5.5, mechanisms of gaze coordination are presented by which an embodied agent can tightly coordinate the expression of its own gaze cues with those being tracked from a human user. These mechanisms take the form of a stochastic finite-state machine model synthesized from the data collected from human-human interactions. A lab evaluation with 32 participants (Section 5.6) demonstrates how these coordinated gaze behaviors can lead to positive collaborative outcomes in error rate, completion time, and the ability of the agent to produce fast but ambiguous references without the need for follow-up verbal repairs. The implementation was also extended to utilize head pose tracking as a low-cost proxy for gaze tracking, and demonstrate in a second study (Section 5.7) that such

an implementation achieves comparable results. Finally, the gaze coordination model was implemented in a head-mounted virtual reality application and demonstrate in a third study (Section 5.8) the usefulness of these techniques for collaborative agents in VR.[1]

Research Questions

- When there are task-relevant objects in the environment, how should an agent distribute its gaze to these objects?
- How do a collaborating dyad's gaze movements—to each other and over a shared task space—coordinate with each other over the course of an interaction?
- How can we design agent gaze behaviors that similarly coordinate with a human collaborator's gaze behaviors?
- What might an agent achieve by coordinating its gaze with user gaze in a humanlike way?

5.1 Related Work

Previous research on gaze behavior has characterized it as a key mechanism for communication and coordination in human interactions and a potential resource for creating natural and effective interactive experiences involving artificial characters. This section reviews related work on gaze coordination from human-human collaboration, existing approaches to analyzing such interactions, and previous work on coordinating the gaze and other social behaviors of artificial agents with human users.

Gaze in Human Coordination and Collaboration

Collaboration requires individuals involved to draw on social interaction to coordinate their actions toward changing their environment (Clark, 1996, 2005; Sebanz et al., 2006). These interactions involve a continuous process of *conversational grounding* whereby communicators interactively come to a state of mutual understanding (Clark and Wilkes-Gibbs, 1986; Clark, 1996). They continuously monitor visual information available on their environment in order to establish *situation awareness*, an ongoing awareness of the environment and the activities taking place within it (Endsley, 1995; Clark and Krych, 2004; Gergle et al.,

[1]Portions of this chapter were published in Andrist et al. (2015). I would like to acknowledge Wesley Collier, a graduate student collaborator, for his contributions in the development of ENA (Section 5.3) and the analyses presented in Section 5.4.

2013). They pay particular attention to the actions and language use of their partners in order to establish perceptual and procedural *common ground* (Sebanz et al., 2006). This awareness enables them to predict breakdowns in coordination and engage in *repair* (Hirst et al., 1994), the use of language to re-establish common ground, such as a teacher seeing that a student is performing an incorrect operation and offering clarification to correct the student's misunderstanding.

Gaze cues facilitate both the process of establishing common ground and the process of engaging in repair. Conversational partners monitor each others' gaze and engage in shared gaze to indicate attention to and understanding of references to objects in their environment (Clark and Krych, 2004; Bard et al., 2009; Gergle and Clark, 2011; Brennan et al., 2012). Breakdowns in understanding or need for more information by listeners can be judged based on whether or not their attention is directed toward referents (Bard et al., 2009). When breakdowns do occur, gaze cues of the speaker serve to rapidly disambiguate references (Hanna and Brennan, 2007). The gaze patterns of a partner can also help predict ensuing task actions (Yi and Ballard, 2009) as well as cognitive processes such as language comprehension (Tanenhaus et al., 1995). When gaze patterns are unavailable or uninformative, e.g., when spoken language contains all the necessary information for establishing common ground and gaze patterns lack any additional information, partners quickly learn not to seek these patterns (Macdonald and Tatler, 2013). For example, collaborators pay significantly less attention to the faces of their partners than other visual information, such as the work area, task objects, tools, and task actions performed by the helper, signaling the relative importance of these visual resources for the collaboration (Fussell et al., 2003).

This continuous process of grounding and repair is facilitated by *gaze coordination*, the emergent coupling of gaze patterns between conversational partners (Richardson et al., 2009). This coordination signals how well speakers and listeners achieve visual common ground (Richardson and Dale, 2005; Bard et al., 2009) and predicts conversational outcomes such as listener comprehension (Richardson et al., 2007). Gaze coordination is considered to be an efficient way to facilitate collaboration, reducing the cost of language production for coordination and repair (Clark and Wilkes-Gibbs, 1986; Brennan et al., 2008), particularly in situations that require rapid communication of spatial information (Neider et al., 2010).

Teams of people generally use eye gaze as a subtle, non-intrusive channel of communication (Shah and Breazeal, 2010). When partners refer to objects or locations in the environment, people use their partner's gaze to predict their partner's next verbal object reference, and can more quickly respond to that reference (Boucher et al., 2012). In contrast, when access to a partner's eye gaze is restricted, people are slower at responding to their partner's referential communication (Boucher et al., 2012).

When collaborating over a shared workspace, conversational partners use each others' gaze to indicate attention toward and understanding of verbal references to objects in the shared environment (Gergle and Clark, 2011). Partners show increased shared gaze toward referents while they speak about those objects (Bard et al., 2009). Referencing is often a multimodal process, with objects being evoked through a speaker's actions, movement, or other pragmatic contextual cues such as gestures or head nods (Gergle and Clark, 2011). Speakers often under-specify their referents, relying on the listener to seek clarification if more information is needed to uniquely identify a particular referent (Campana et al., 2001). Previous research has shown that speakers look toward their addressees in order to check their understanding of references to new entities (Nakano et al., 2003) and that addressees rely on the speaker's gaze as a cue for disambiguating references, often before the reference could be disambiguated linguistically (Hanna and Brennan, 2007). This use of gaze has the effect of minimizing the joint effort of the participants in an interaction by reducing the time speakers must spend specifying referents.

Analyzing Gaze Coordination

Previous work has examined the timings of when people look toward referents—objects to which they or their interlocutors verbally refer (Griffin, 2004; Meyer et al., 2004; Tanenhaus et al., 1995). However, these investigations are generally one-sided, looking at each person's gaze in isolation, and do not capture the intricate coordinative patterns in which partners' referential gaze behaviors interact. Most previous research on gaze in interaction makes a simplifying assumption of *pseudounilaterality*—the implicit assumption that a behavioral variable is unilaterally determined by the actions of the participant expressing that behavior (Duncan et al., 1984). This assumption results in erroneously interpreting data on a participant's actions as representing the unilateral conduct of that participant, overlooking the partner's contribution to those data. A primary cause of pseudounilaterality is the use of simple-rate variables—generated by counting or by timing the occurrence of an action during an interaction and dividing that number by some broader count or timing. These variables do not contain information on the sequences in which actions occur in interaction.

Mobile dual eye-tracking is a relatively recent approach to capturing gaze behaviors that allows researchers to overcome problems of pseudounilaterality and develop more nuanced and ecologically valid accounts of how interlocutors coordinate their gaze during natural, situated conversations (Clark and Gergle, 2011). They have provided great opportunities for researchers to better understand the role of gaze as a coordination mechanism in conversation. Dual eye-tracking methods can be used to better understand the role gaze

plays as a conversational resource during reference—how people specify the person, object, or entity that they are talking about (Clark and Gergle, 2012).

Cross-recurrence analysis is a commonly used technique for analyzing gaze data captured from participant dyads, as it permits the visualization and quantification of recurrent patterns of states between two time series, such as the gaze patterns of two conversational participants (Zbilut et al., 1998) (Figure 5.1). This analysis approach can reveal the temporal dynamics of a dataset without making assumptions about its statistical nature. The horizontal and vertical axes of a cross-recurrence plot specify the gaze of each of the two partners in interaction. Each diagonal on the plot (lower-left to upper-right) corresponds to an alignment of the participants' gaze with a particular time lag between them. A point is plotted on the diagonal whenever the gaze is *recurrent*—their eyes are fixating at the same object at the given time. The longest diagonal, from bottom-left to top-right of the plot, represents the gaze alignment at a lag of `0`. Diagonals above and below that line represent alignments with positive and negative offsets, shifting one of the participants' time-series gaze data in relation to the other participant.

Previous research utilizing cross-recurrence analysis has successfully expanded knowledge on gaze coordination. For example, research has shown that a listener's eye movements most closely match a speaker's eye movements at a delay of two seconds (Richardson and Dale, 2005) (Figure 5.1). In fact, the more closely a listener's eye movements are coupled with a speaker's, the better the listener does on a comprehension test. These results were later extended to find that eye movement coupling is sensitive to the knowledge that participants bring to their conversations (Richardson et al., 2007). The presence of the visual scene and beliefs about its perception by others also influence language use and gaze coordination in remote collaborations (Richardson et al., 2009). Gaze is not always well aligned; when speakers' referring expressions ignore listeners' needs, dyads show poor coordination of visual attention (Bard et al., 2009). Dyads whose members more effectively produce referring expressions better coordinate their attention better and in a way linked to the elaboration of the referring expressions.

Although cross-recurrence analysis has yielded some success in studying gaze coordination, it is best suited for examining data from short time windows and one pair at a time. Cross-recurrence plots do not support aggregating data from numerous dyads over long time spans in order to abstract away individual differences and discover generalizable patterns of interaction. These plots can also be difficult to interpret visually and lack the sophistication to represent the complex, dynamic relationships that characterize coordinated gaze over a shared physical workspace. This chapter utilizes a particular instantiation of network analysis—epistemic network analysis—as an alternative analytical tool that

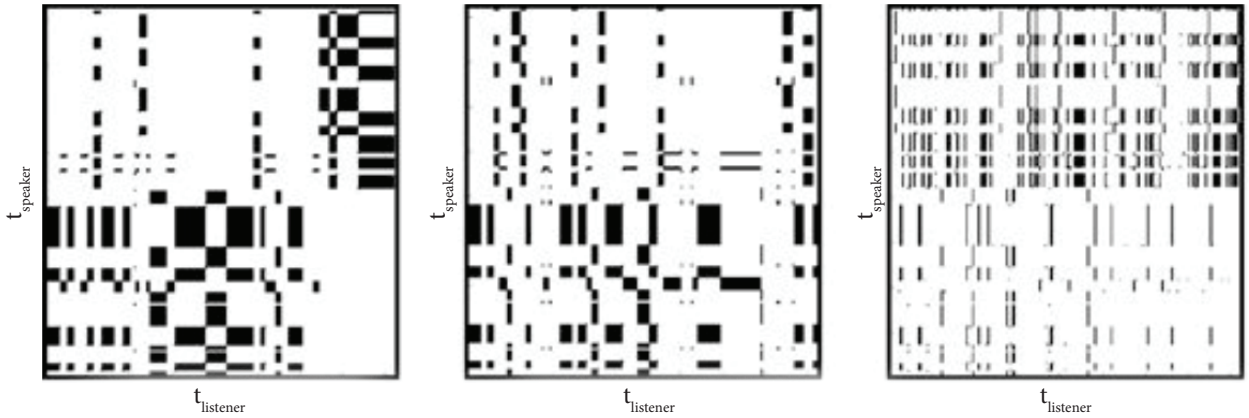

Figure 5.1: Cross-recurrence plots adapted from work by Richardson and Dale (2005). Horizontal and vertical axes specify the gaze of a speaker and a listener in discrete time windows. Diagonal slices (lower-left to upper-right) correspond to an alignment of the participants' gaze with a particular time lag between them. A point is plotted on the diagonal whenever the gaze is recurrent. These plots visually compare a "good" listener (well aligned with the speaker's gaze) to a "bad" listener (not as well aligned). They also show the poor alignment of random gaze with a speaker's gaze.

overcomes these issues.

Gaze in Coordination with Embodied Agents

Research in human-computer and human-robot interaction includes a rich body of work that explored how both virtual agents and social robots can utilize gaze cues for signaling attention (Peters et al., 2010; Pejsa et al., 2015) spatial referencing (Brooks and Breazeal, 2006), and action coordination (Boucher et al., 2012; Moon et al., 2014). Only a few studies in this body of work investigated how embodied agents can utilize gaze as a dynamic, interactive resource for coordination. These studies explored how embodied agents can monitor the gaze of their users and seek to establish shared or mutual attention by aligning their gaze with those of their users (Yoshikawa et al., 2006), by selectively using gaze shifts, head motion, and voice depending on user orientation (Hoque et al., 2012), and by dynamically engaging in mutual and averted gaze depending on user gaze (Bee et al., 2010). Yoshikawa et al. (2006) found that a robot that monitors the gaze of its user and aligns its gaze to that of its user in order to establish shared attention improves user feelings of being looked at. Similarly, Bee et al. (2010) showed that maintaining mutual attention by dynamically engaging in mutual and averted gaze improves user rapport with and perceptions of the social presence of a virtual storyteller.

Another line of work among these studies involves enabling embodied agents to monitor user gaze for signals of communication or task breakdowns and to offer repair. The earliest study to achieve this ability developed a spoken dialogue system that monitored user gaze patterns to determine the need for reference resolution (Campana et al., 2001). Sakita et al. (2004) developed a robot system that monitored user gaze as they performed an assembly task, predict opportune moments, such as a user deciding on what parts to use next in the assembly, and proactively provide task assistance, such as picking up and passing the desired part to the user. Torrey et al. (2007) enabled a robot to monitor user gaze and task actions, predict hesitation or breakdowns in understanding, and offer additional task information. They found that offering assistance based on breakdowns in task actions reduced the need for repair but assistance based on user gaze did not. A similar study explored how an intelligent tutoring system can use student behaviors, including gaze direction, to determine whether or not the student lost attention and prompt the student to pay attention to the tutoring (D'Mello et al., 2012). The prompt succeeded in regaining student gaze toward the tutor and resulted in partial learning gains.

A small number of studies in human-robot interaction have investigated gaze coordination between humans and robots. Pitsch et al. (2013) asked participants to demonstrate a set of actions to a robot and studied the coupling between the robot's gaze and participant demonstration behaviors. They observed that participants adapted their demonstrations to the speed in which the robot engaged in gaze following and delayed their demonstrations when the robot gazed away, suggesting a tight coupling between robot and user gaze. Skantze et al. (2014) compared user interactions with a robot that employed coordination mechanisms such as joint attention, turn-taking, and action monitoring to those lacking these mechanisms or those lacking gaze cues altogether and found that these mechanisms facilitated reference disambiguation and turn-taking.

Although previous work has extensively studied human gaze coordination and explored integrating some of its mechanisms into the design of embodied agents, humanlike gaze coordination with agents is still an unrealized goal. This chapter seeks to address this knowledge gap by presenting a detailed analysis of human gaze coordination, along with a computational model of gaze coordination that allows virtual characters to express gaze cues and react to user gaze cues in a coordinated fashion in order to create richer and more natural interactive experiences.

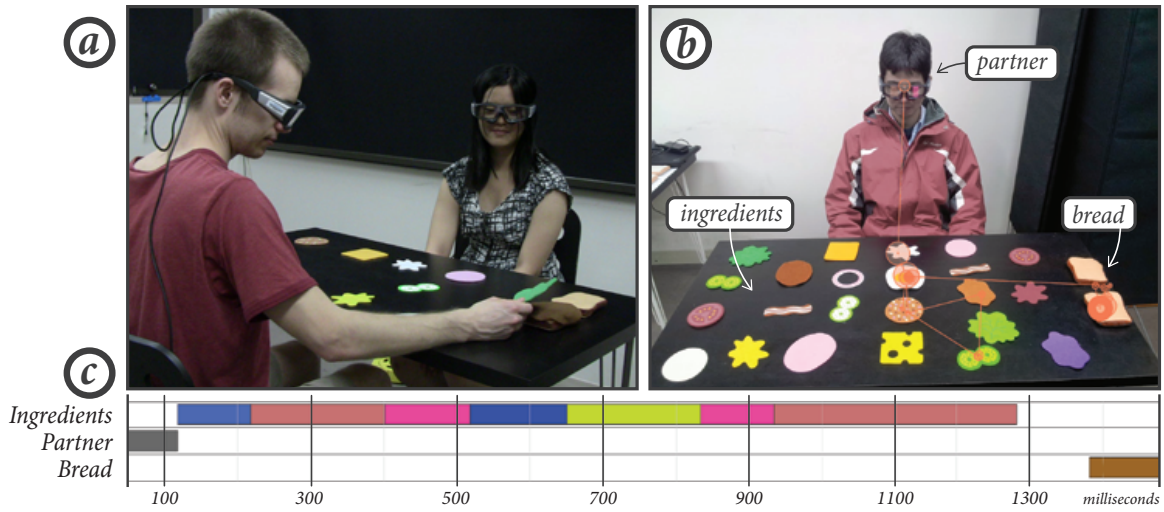

Figure 5.2: (a) The setup of the data collection experiment in the sandwich-making task. (b) A view from one participant's eye-tracking glasses, showing their scan path throughout a reference-action sequence. (c) A timeline view of the gaze fixations to ingredients, the partner, and the bread shown in the scan path in (b).

5.2 Data Collection

In order to gain a better understanding of how gaze coordination unfolds over reference-action sequences in dyadic collaborations, a data collection study was conducted in which pairs of participants engaged in a collaborative sandwich-making task. Thirteen previously unacquainted dyads of participants were recruited from the University of Wisconsin–Madison campus. This data collection study was approved by the Education and Social/Behavioral Science Institutional Review Board (IRB) of the University of Wisconsin–Madison and all participants granted their written informed consent at the beginning of the study procedure. Participants sat across from each other at a table on which were laid out a number of potential sandwich ingredients and two slices of bread (Figure 5.2). One participant was assigned the role of *instructor*, and the other was assigned the role of *worker*. The instructor acted as a customer at a deli counter, using verbal instructions to tell the worker what ingredients they wanted on their sandwich, and the worker carried out the actions of moving the desired ingredients to the bread.

Each dyad carried out the sandwich-making task twice so that each participant would have a turn as both instructor and worker, resulting in 26 dyadic interactions. The experimenter told the instructor to request any 15 ingredients for their sandwich from among 23 ingredients laid out on the table. The choice of ingredients was left to the instructor; no list was provided by the experimenter. The instructor was asked to only request a single

ingredient at a time and to refrain from pointing to or touching the ingredients directly. Upon completion of the first sandwich, an experimenter entered the study room to reset the ingredients back to their original locations on the table, and the participants switched roles for the second sandwich.

During the study, both participants wore mobile eye-tracking glasses developed by SMI.[2] 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 camera that was used to record both audio and video (24 fps). The gaze trackers were time-synchronized with each other so that the gaze data from both participants could be correlated.

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. Fixation identification minimizes the complexity of eye-tracking data while retaining its most essential characteristics for the purposes of understanding cognitive and visual processing behavior (Salvucci and Goldberg, 2000). BeGaze uses a dispersion-based spatial algorithm to compute fixations, emphasizing the spread distance of fixation points under the assumption that fixation points generally occur near one another. Eye fixations and saccades are computed in relation to a forward-facing camera located in the bridge of the eye-tracking glasses worn by the user. Thus, these fixations and saccades are defined within the coordinate frame of the user's head, and user head movements do not interfere with the detection of eye movements.

Gaze fixations are characterized by their duration and coordinates within the forward-facing camera view. Area-of-interest (AOI) analysis, which maps fixations to labeled target areas (AOIs) is a common method for adding semantic information to raw gaze fixations (Salvucci and Goldberg, 2000). In this work, all fixations were manually labeled for the target of the fixation. These labeled AOIs serve as the input data for ENA and the computational model for agents, rather than the raw gaze fixations. Possible target AOIs included the sandwich ingredients, the slices of bread, and the conversational partner's face and body. Around 80% of gaze fixations were mapped to these AOIs (79.47% for instructors, 81.65% for workers), and the remainder of gaze fixations were found to be directed elsewhere in space (e.g., to a spot on the table without a sandwich ingredient). Speech was also transcribed for each participant. Instructor requests for specific objects were tagged with the ID of the referenced object, and worker speech was labeled when it was either confirming a request or asking for clarification.

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

To make successful reference utterances, the speaker needs some form of feedback from the addressee. Despite the best efforts of speakers, there will inevitably be instances of breakdowns—misunderstandings or miscommunication—that can either impede ongoing progress of the interaction or lead to breakdowns in the future (Zahn, 1984). To correct breakdowns, humans engage in *repair*, a process that allows speakers to correct misunderstandings and helps ensure that the listener has the correct understanding of the relayed information (Hirst et al., 1994; Zahn, 1984). In the current data collection, if an instructor provided extra clarification for an initially inadequate reference, possibly prompted by the worker's request for clarification, that sequence was marked as containing a *repair*.

Following data collection, each interaction was divided into a set of *reference-action sequences*, such as a verbal request for bacon followed by the action of moving the bacon to the bread. Each sequence was further divided into five discrete phases: *pre-reference*, the time before any verbal reference has been made; *reference*, the time during the verbal request for a specific sandwich ingredient; *monitor*, the time directly after the verbal reference and up until the worker's action; *action*, the time during the worker's action of moving the ingredient to the target bread; and *post-action*, the time immediately following this action.

These phases are defined according to verbal and physical actions, not according to gaze behaviors, which are analyzed within each of these phases. The pre-reference phase (average length = 1.90s) ends at the onset of the verbal reference. The reference phase (average length = 1.32s) ends with the end of the utterance of the verbal reference. The end of the monitor phase (average length = 0.78s) is marked by the start of the physical action, which involves picking up the referent, particularly the moment it is first touched. The action phase (average length = 1.68s) ends with the end of the physical action, which involves moving the ingredient to the bread and is marked by the moment it is let go. Finally, the end of any feedback provided by the instructor or the beginning of some preparatory utterance for the next reference, e.g., "so, uh, next I'll have...," marked the end of the post-action phase (average length = 0.81s).

5.3 Epistemic Network Analysis

Studying gaze coordination and the temporal unfolding of collaborative gaze behaviors is difficult due to the highly dynamic and interdependent nature of the data. In order to explore this type of data, an approach was used that is similar to social network analysis, which provides a robust set of analytical tools to represent networks of relationships, including complex and dynamic relationships (Wasserman and Faust, 1994; Brandes and Erlebach, 2005). However, social network analysis was developed to investigate relation-

ships between people rather than relationships within discourse, gaze behaviors, or other indicators of cognitive processes.

Epistemic network analysis (ENA) is a relatively new analysis technique that is based in part on social network analytic models. ENA extends social network analysis by focusing on the patterns of relations among discourse elements, including the things people say and do. ENA networks are characterized by a relatively small number of nodes in contrast with the very large networks that techniques from social network analysis were designed to analyze, which often have hundreds, thousands, or even millions of nodes. In ENA networks, the weights of the connections between nodes (i.e., the association structures between elements) are particularly important, as are the dynamic changes in the weights and in the relative weighting of the links between different nodes.

ENA was designed to highlight connections among "actors," e.g., people, ideas, concepts, events, and behaviors, in a system. It was originally developed to measure relationships between elements of professional expertise by quantifying the co-occurrences of those elements in discourse and has been used for that purpose in a number of contexts (Orrill and Shaffer, 2012; Rupp et al., 2009, 2010; Shaffer et al., 2009). However, ENA is a promising method to effectively analyze datasets that capture the co-occurrence of any behaviors or actions in social interactions over time.

The data within ENA are represented in a dynamic network model that quantifies changes in the strength and composition of *epistemic frames* over time. An epistemic frame is composed of individual frame elements, f_i, where i represents a particular coded element in a specified window of time. For our purposes, "coded elements" of the epistemic frame are annotated *gaze targets* for each participant in the interaction, and these elements are represented as nodes in a network. For any dyad, p, in any given reference-action sequence, s, each segment of interaction discourse, $D^{p,s}$, provides evidence of which epistemic frame elements (gaze targets) were active (being gazed toward). For this work, each segment of interaction represents 50ms of time in the interaction.

Each segment of coded data is represented as a vector of `1`s or `0`s representing the presence or absence, respectively, of each of the codes. Links, or relations, between epistemic frame elements are defined as co-occurrences of codes within the same segment. To calculate these links, each coded vector is converted into an adjacency matrix, $A^{p,s}$, for dyad p. For present purposes, co-occurrence of two codes is equivalent to the recurrence of gaze to the gaze targets represented by the codes. For any two gaze codes, the strength of their association in a network is computed based on the frequency of their co-occurrence in the data.

$$A_{i,j}^{p,s} = 1 \text{ if } f_i \text{ and } f_j \text{ are both in } D^{p,s}$$

Each coded segment's adjacency matrix, $A_{i,j}^{p,s}$ is then converted into an adjacency vector and summed into a single cumulative adjacency vector for each dyad p for each unit of analysis.

$$U^{p,s} = \sum A^{p,s}$$

For each dyad, p, and each reference-action sequence, s, the cumulative adjacency vector, $U^{p,s}$, is used to define the location of the segments in a high dimensional vector space defined by the intersections of each of the codes. Cumulative adjacency vectors are then normalized to a unit hypersphere to control for the variation in vector length, representing frequencies of co-occurring code pairs, by dividing each value by the square root of the sum of squares of the vector.

$$nU^{p,s} = U^{p,s} / \sqrt{\sum (U^{p,s})^2}$$

A singular value decomposition (SVD) is then performed to explore the structure of the code co-occurrences in the dataset. The normalized cumulative adjacency vectors are first projected into a high dimensional space such that similar patterns of co-occurrences between coded elements would be positioned proximately. The SVD analysis then decomposes the structure of the data in this high dimensional space into a set of uncorrelated components, fewer in number than the number of dimensions that still account for as much of the variance in the data as possible, such that each accumulated adjacency vector, i, has a set of coordinates, P_i, on the reduced set of dimensions. The resulting networks are then visualized by locating the original frame elements, i.e., the network nodes, using an optimization routine that minimizes

$$\sum (P_i - C_i)^2$$

where P_i is the projection of the point under SVD, and C_i is the centroid of the network graph under the node positioning being tested. This operation produces a distribution of nodes in the network graph determined by the loading vectors that contain them in the space of adjacency vectors. Links are then constructed between the positioned network nodes according to the adjacency matrix.

The mean network for a group of networks can be calculated by computing the mean values of each edge weight in the networks. We can also conduct *t*-tests between groups of

networks to determine if one group's networks (group A) are statistically different from a second group's networks (group B). The *t*-test operates on the distribution of the centroids of each group on one dimension. For example, we can determine if group A is statistically different from group B on the x-axis by calculating the means of each group's centroid projected to the x-axis and then conducting a *t*-test with a standard alpha level of 0.05.

5.4 Analyzing Gaze Coordination with ENA

The first step in the analysis involved calculating common descriptive statistics for the gaze data. Unsurprisingly, very little mutual gaze was found during the reference-action sequences (0.92%) and a fairly large amount of simultaneous shared gaze toward the same target (31.16%). Instructors produced their verbal reference utterance on average 1.31s after first fixating on it, although they made on average 1.93 fixations to the reference object before verbalizing it. Workers fixated on the reference object on average 1.65s after the verbal reference. Previous research has found that referential gaze in speech typically precedes the corresponding linguistic reference by approximately 800–1000ms, and people look at what they hear after about 2000ms (Meyer et al., 1998; Griffin and Bock, 2000). The collected data seem to yield statistics close to these findings, and the slightly longer time offset between the gaze fixation and verbal reference among instructors may be due to occasionally having to search for an object, rather than having one already in mind at the beginning of the interaction.

Analysis 1

The entirety of the collected data was analyzed using ENA (Figure 5.3). The first analysis considered each dyad ($n = 26$; 13 dyads $\times$ two interactions each) and phase ($n = 5$; pre-reference, reference, monitor, action, or post-action) as the units of analysis. Each point in the central plot of Figure 5.3 represents the centroid of a network for a single dyad's interaction in one of the five phases, collapsed across all reference-action sequences that occurred in the interaction. Solid squares represent the centroid of the mean network for all dyads in each of the five phases. These mean network centroids are surrounded by squares representing the confidence interval along both dimensions. A clear separation between each of the five phases can be observed, indicating that the patterns of gaze coordination are significantly different in each of the five phases. We can also observe a clear cyclical pattern through the two-dimensional ENA space as we progress through each of the five phases in the reference-action sequence.

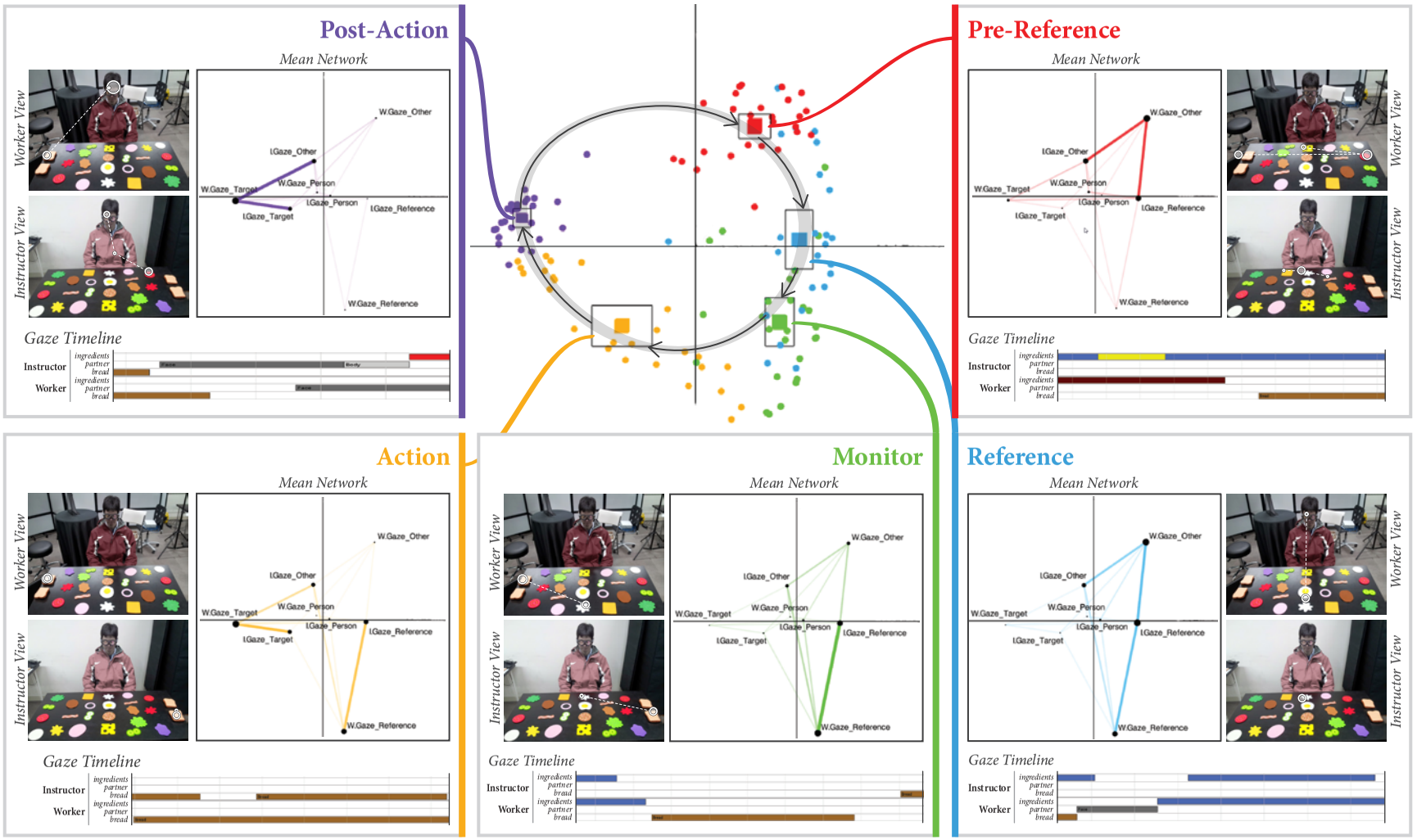

Figure 5.3: **Center**: Each circular point represents the centroid of a network for one dyad in a particular phase, collapsed across all reference-action sequences produced by that dyad. The centroid of the mean network for each phase is also plotted as a solid square surrounded by a larger square denoting the confidence interval. A cyclical relationship through the ENA space can be observed. **Boxes in periphery**: The mean network for each of the five sequences is fully plotted. A representative timeline of an example gaze sequence from the raw gaze data is shown beneath the mean networks to illustrate each phase. A view of the worker's and instructor's scan paths in that phase (same data as in the timeline) is also shown.

ENA Network Node Names and Meanings

Analysis 1, 2, 3	*I.Gaze_Reference*	Instructor gazing at reference ingredient
	I.Gaze_Other	Instructor gazing at non-reference ingredient
	I.Gaze_Target	Instructor gazing at target bread
	I.Gaze_Person	Instructor gazing at the worker
Analysis 1, 3	*W.Gaze_Reference*	Worker gazing at reference ingredient
	W.Gaze_Other	Worker gazing at non-reference ingredient
	W.Gaze_Target	Worker gazing at target bread
	W.Gaze_Person	Worker gazing at the instructor
Analysis 2	*W.Same*	Worker gazing at same object as instructor
	W.Different	Worker gazing at different object than instructor

Table 5.1: Naming convention and meanings of all network nodes used throughout the different analyses.

Figure 5.3 also plots the full mean networks for each of the five phases. As mentioned previously, nodes represent gaze targets, and edge weights represent the relative amount of recurrent gaze to those targets. There are four gaze target nodes for each participant: (1) the referent for the sequence, (2) the interaction partner, (3) the action target (the bread to which ingredients are moved), and (4) all other objects. In these networks, edges only connect instructor and worker gaze target nodes, as simultaneous gaze within one person toward different targets is not possible. The naming conventions and meanings of all network nodes are explained in Table 5.1.

By examining the placement of nodes in the mean networks, we can develop an intuitive sense of the meaning of each axis in ENA space. As can be observed in the mean networks shown in Figure 5.3, ENA keeps the node positions identical across all plots for a given analysis. Nodes placed at extreme edges of the space, far from the center, are the most informative for intuitively labeling axes. In this respect, three nodes stand out: *W.Gaze_Other*, *W.Gaze_Reference*, and *W.Gaze_Target*. We can therefore recognize that networks with centroids located high on the y-axis are most characterized by strong connections to *W.Gaze_Other*. In other words, these networks include more worker gaze toward non-referents. In general, moving from high to low along the y-axis seems to indicate a shift from worker gaze toward non-referents to worker gaze toward the referent. Similarly, moving from right to left along the x-axis seems to indicate a shift from worker gaze toward sandwich ingredients (referents or non-referents) to worker gaze toward the target bread.

In each of the mean networks plotted in Figure 5.3 for each of the five phases, the key differences to note are the shifting edge strengths between nodes. In the pre-reference phase, we can observe that the network—which has a centroid high along the y-axis in the central plot of Figure 5.3—has particularly strong connections between *W.Gaze_Other* and *I.Gaze_Other* and between *W.Gaze_Other* and *I.Gaze_Reference*. These connections tell us that the pre-reference phase is characterized mostly by the worker looking toward non-referents while the instructor scans the objects, including the object that they will verbally indicate as the referent in the next phase of the sequence. In the reference phase, we can observe a growing connection between *W.Gaze_Reference* and *I.Gaze_Reference*, pulling the network centroids lower along the y-axis. In the monitor phase, this connection is now strongest, and connections with *W.Gaze_Other* (the worker gazing to non-referents) have become much weaker, pulling these network centroids yet lower along the y-axis.

In the action phase, a strong connection between *W.Gaze_Target* and *I.Gaze_Target* appears, signaling simultaneous gaze toward the target, which, in this case, is the bread toward which the selected sandwich ingredient is being moved, pulling the network centroids left along the x-axis. Finally, the post-action phase retains the strong connection between *W.Gaze_Target* and *I.Gaze_Target*, with a new strong connection between *W.Gaze_Target* and *I.Gaze_Other*, indicating that the instructor has started to scan other objects in anticipation of the next reference-action sequence while the worker finishes gazing toward the target.

This first analysis provides an overall picture of the unfolding gaze patterns in dyadic collaborations throughout a reference-action sequence. A clear separation of shared gaze networks was observed between each of the five phases in the reference-action sequence, as well as an orderly cyclical pattern throughout the two-dimensional ENA space. It is important to reiterate that although the phases themselves are defined in terms of the temporal location of the reference speech and movement action, ENA is acting only upon the gaze data. Thus, patterns of shared gaze are uniquely different across the different phases of the sequence, e.g., before a verbal reference, during the reference, immediately after that reference, and so on. Furthermore, these patterns change and mutate in an orderly way through the abstract space defined by ENA. Theoretically, a mapping from the gaze networks back to the phases can be built. Given a segment of gaze, the phase of the reference-action sequence it came from could be predicted by computing the ENA network for that segment and plotting it in this space.

To validate and demonstrate the promise of the ENA analysis for prediction, a simple test was performed that involved computing the ENA network as described above, but leaving out data from one of the 13 dyads, which resulted in an ENA space very similar to

Predicting phase from segments of gaze data

Actual phase	Predicted phase (200 *ms* segments)					Predicted phase (1000 *ms* segments)				
	Pre-Reference	*Reference*	*Monitor*	*Action*	*Post-Action*	*Pre-Reference*	*Reference*	*Monitor*	*Action*	*Post-Action*
Pre-Reference	117	3	10	60	16	31	3	1	2	4
Reference	50	5	76	38	3	10	2	18	4	0
Monitor	6	0	31	6	0	0	2	7	0	0
Action	7	1	46	52	54	2	0	10	7	13
Post-Action	7	0	0	33	61	0	0	0	2	18

Table 5.2: To demonstrate how ENA analysis can be used for prediction, segments of gaze data were projected into the ENA space, and their phase was predicted according to the nearest centroid of phase networks. Rows are the actual phase that each segment of data is from, and columns are the predicted phase. Prediction appears to be fairly accurate except for some confusion in the shorter phases of *reference* and *action*.

that shown in Figure 5.3. From the left-out dyad, 200ms and 1000ms segments of gaze data were randomly selected. Each of these segments were then modeled as adjacency vectors and projected into the ENA space constructed from data from the other 12 dyads. The predicted phase for each of the projected segments was labeled according to the nearest centroid of phase segments in the ENA space. Table 5.2 illustrates the results from this analysis in the form of a confusion matrix. Rows are the actual phase that each segment of data is from, and columns are the predicted phase. As can be seen in the table, prediction appears to be fairly accurate except for some confusion in the shorter phases of *reference* and *action*. In realistic prediction scenarios, prediction accuracy can be improved by using more sophisticated methods than the one employed here for demonstrative purposes, such as dynamically updating phase predictions as segments of gaze data are collected over time or assigning confidence weights to predictions based on their distance from phase centroids.

Analysis 2

The second analysis was concerned with finding the optimal lag of gaze alignment within each of the five phases. In other words, which participant's gaze leads that of the other, and by how much, in each phase? For this analysis, two new ENA codes were created: *same*, which is active if the worker and instructor are gazing at the same target (person, referent, target, or other), and *different*, which is active otherwise. For each phase of the reference-action sequence, across all dyads, the instructor's gaze was shifted (in relation to the worker's gaze) from −2000ms to 2000ms in 50ms increments, computing at each

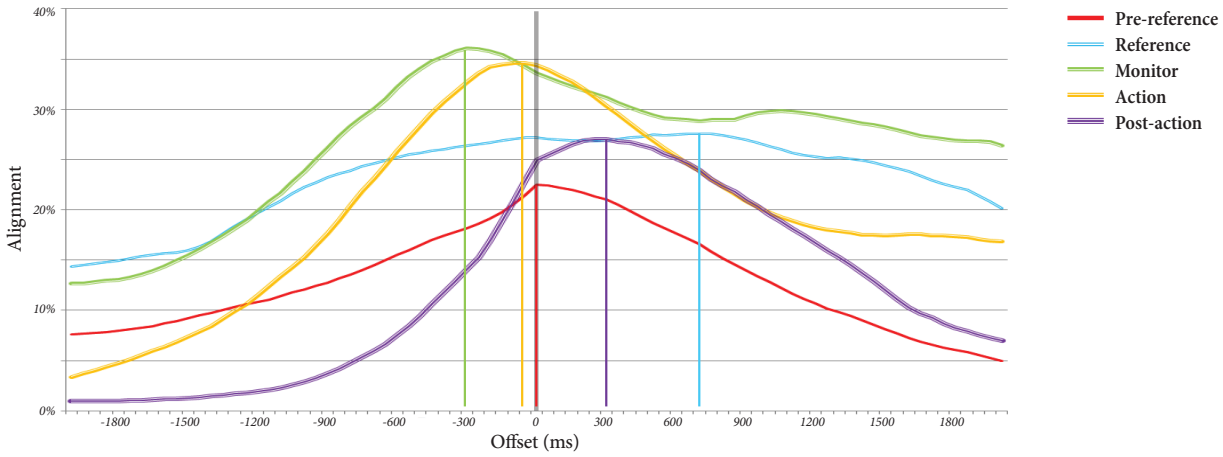

Figure 5.4: Percentage of gaze alignment between the instructor and worker at each of the five phases, plotted at offset lags from −2s to 2s. Positive lags indicate instructor lead, while negative lags put the worker ahead of the instructor.

increment the value for each of the new codes. To find the optimal overlap, the sum of the *same* code was divided by the total number of increments in order to find a measure of "alignment" at each time lag. These alignments for each of the five phases are plotted in Figure 5.4.

The peak of the line graph for each of the five phases represents the optimal time lag at that phase. These lags, as well as the amount of gaze alignment that occurs at those lags, are summarized in Table 5.3. Positive lags put the instructor ahead of the worker, indicating that the instructor is "driving" the gaze patterns, while negative lags indicate that the worker is driving the gaze patterns. As can be observed, the pre-reference phase is characterized by neither participant driving the gaze patterns (t = 0s) and a relatively low amount of gaze alignment ($alignment = 22.5\%$). However, during the reference phase, the instructor starts to lead the gaze patterns (t = 700ms), and the alignment increases ($alignment = 27.6\%$). In the monitor phase, the worker begins leading (t = −300ms), and the dyad is most aligned ($alignment = 36.1\%$). The action phase involves a slight lead by the worker (t = −50ms) and slight drop in alignment ($alignment = 34.6\%$). In the post-action phase, the instructor is once again leading (t = 300ms), and the alignment has dropped further ($alignment = 27.0\%$).

Next in the analysis, the gaze streams were shifted in each phase of the reference-action sequence by that phase's optimal time lag (Table 5.3). Analysis in ENA was conducted by modeling from the instructor's perspective (Figure 5.5). Four nodes represent the possible gaze targets for the instructor as before, but there are only two nodes for the worker: $W.Same$, signifying whether the worker is looking at the same target as the instructor, and

W.Different, indicating a different target than the instructor.

By examining the placement of nodes in the mean networks shown in Figure 5.5, we can again develop an intuitive sense of the meaning of each axis in this new ENA space. Along the x-axis, we can observe I.Gaze_Reference far to the left and I.Gaze_Target far to the right, indicating a progression from referent-directed gaze to target-directed gaze in this dimension, as the phases move from left to right along the x-axis.

For the y-axis, I.Gaze_Person is the lowest node, but the mean networks throughout the five phases in Figure 5.5 show only a few strong connections with I.Gaze_Person, indicating that the instructor's gaze is not directed toward the worker. Instead, connections with W.Same get stronger as the phases move from *pre-reference* to *reference* to *monitor* and then weaker again as they move to *action* and *post-action*. Strong connections with W.Same pull the network centroids lower along the y-axis in the central plot of Figure 5.5, suggesting an interpretation that this axis signifies "alignment." We can observe a rise and fall of alignment in the phases as their corresponding networks fall and rise respectively along the y-axis. This observation matches what we see in Table 5.3 where the alignment percentages rise and fall throughout the five phases.

Analysis 3

The third and final analysis was concerned with the differences between phases of reference-action sequences that included a repair—in which the instructor had to provide a clarification to their first verbal reference, possibly at the explicit verbal request of the worker—from phases that did not include such repairs. The purpose of this analysis was to answer the following questions. Do the patterns of coordinated gaze in ENA look different during typical sequences vs. those involving repair? More importantly, can the gaze patterns from early phases (pre-reference, reference, and monitor) be used to predict breakdowns later in the sequence, e.g., before the worker or the instructor offers repair or during repair?

This analysis included "repair" ($n = 2$; repair or no-repair) as another unit of analysis in addition to the "dyad" and "phase" units from before. As can be observed in Figure 5.6,

Optimal Lag & Alignment Percentage

	Pre-Reference	Reference	Monitor	Action	Post-Action
Optimal Lag (ms)	0	700	−300	−50	300
Alignment (%)	22.5	27.6	36.1	34.6	27.0

Table 5.3: Optimal time lags identified in Analysis 2 and the percentage of alignment at each offset.

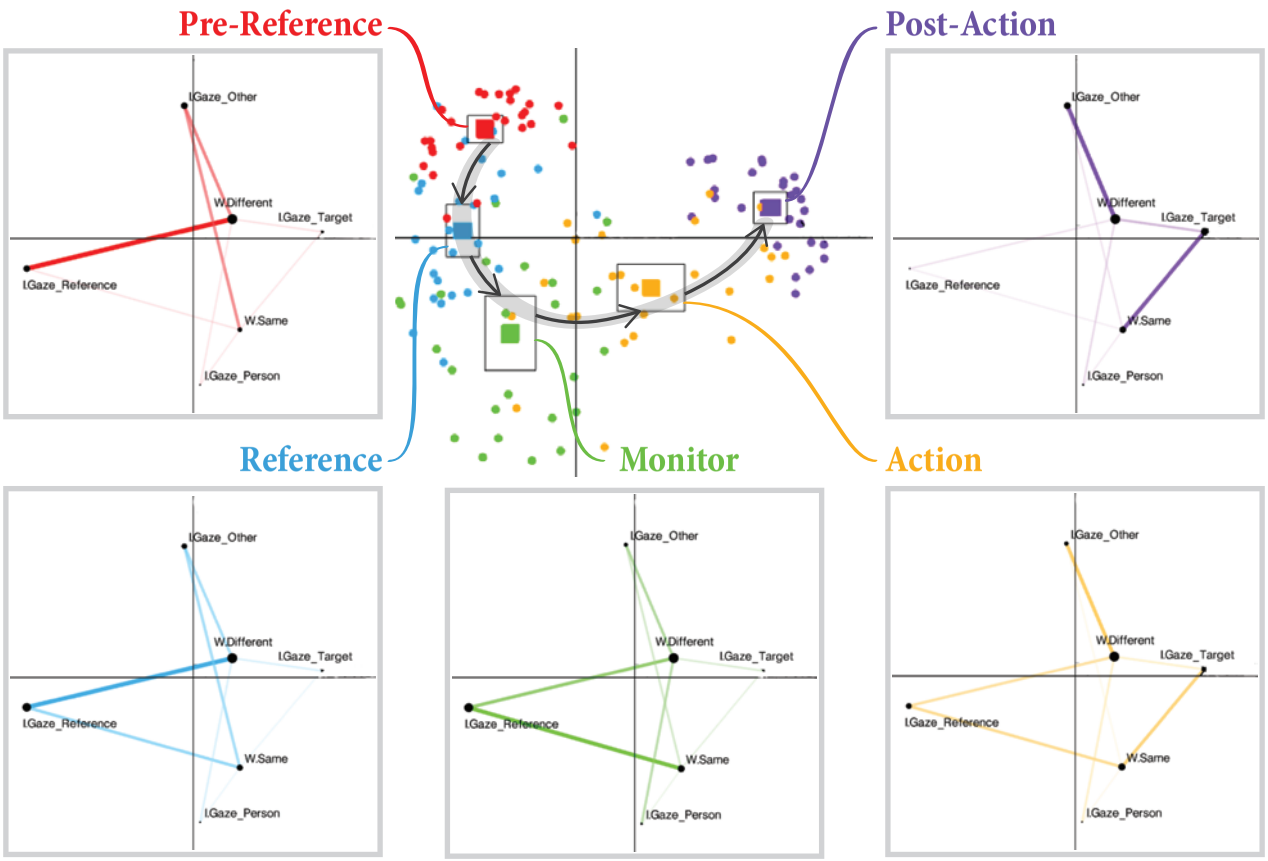

Figure 5.5: Centroids and mean networks from the ENA that used gaze data from each phase that was shifted by the optimal lag for that phase. The data is modeled from the perspective of the instructor. Four nodes represent the possible gaze targets for the instructor as before, but there are only two nodes for the worker, signifying whether the worker is looking at the same target or a different target. *W_Different* and *W_Same* are largely vertically separated. Networks that are low on the y-axis have strong connections to *W_Same*, while networks high on the axis have strong connections to *W_Different*. Thus, the y-axis can be interpreted as signifying "alignment," and we can observe a rise and fall of alignment in the phases as their corresponding networks fall and rise respectively in the ENA space.

gaze networks are significantly different between repair and no-repair along the y-axis for each of the first three phases in the reference-action sequence. The centroids of the mean networks (solid squares) for these phases are separated along the y-axis, and there is little vertical overlap in their confidence intervals. These phases, which occur before or during any possible repair, are thus potentially distinguishable along this dimension.

For the pre-reference phase, networks with repair are significantly higher on the y-axis than networks without repair, ($mean_{no-repair} = -0.46$, $mean_{repair} = -0.36$, $t = -2.17$, $p = .036$, Cohen's $d = -0.25$). Based on an inspection of the mean networks on the left side of Figure 5.6, this difference appears to be mostly due to the stronger connection between I.Gaze_Reference and W.Gaze_Target in the sequences with repair, which pulls the network centroids higher along the y-axis. This connection denotes a situation in which the worker is looking toward the target bread while the instructor is looking toward the referent. Here, the worker may still be cognitively engaged in the previous reference-action sequence, i.e., still looking toward the bread after moving the previous reference object there, while the instructor is already preparing their reference utterance for the current reference-action sequence, leading to an eventual breakdown in the interaction.

On the other hand, networks with repair are lower on the y-axis than networks without repair for the reference ($mean_{no-repair} = 0.057$, $mean_{repair} = -0.15$, $t = 2.12$, $p = .04$, Cohen's $d = 0.37$) and monitor ($mean_{no-repair} = 0.42$, $mean_{repair} = 0.18$, $t = 2.79$, $p = .008$, Cohen's $d = 0.45$) phases. These differences appear to be mostly due to stronger connections with W.Gaze_Other (situated very low on the y-axis) in the sequences with repairs, as shown in Figure 5.6. In other words, the worker is gazing more toward non-referents in these sequences. Also, the networks coming from sequences without repairs appear to have stronger connections between I.Gaze_Reference and W.Gaze_Reference, pulling these networks higher along the y-axis. This observation implies that, when both the instructor and worker are fixated on the reference object, repairs are less likely to happen.

This analysis revealed that the pattern of coordinated gaze identified in Analysis 1 shows both similarities and differences during sequences involving a repair. More interestingly, the gaze behaviors from phases early in the sequence, particularly the pre-reference and reference phases, are visibly different when a repair occurs later in the sequence than when a repair does not occur later in the sequence. Thus, the need for repair can theoretically be anticipated in advance by observing the pattern of gaze behaviors early in a reference-action sequence.

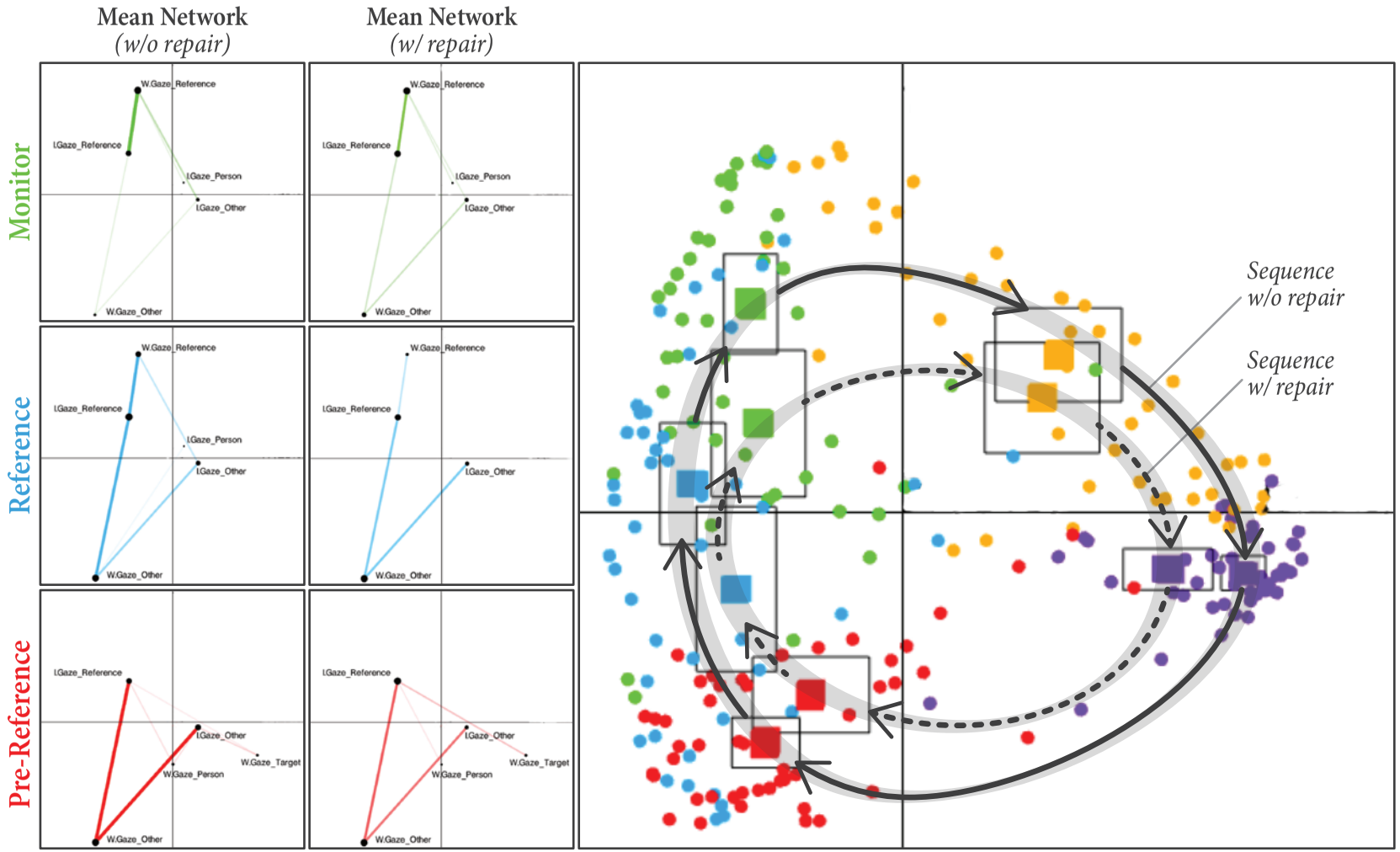

Figure 5.6: **Right**: Each circular point represents the centroid of a network for one dyad in a particular phase with or without a repair occurring in the reference-action sequence. The centroid of the mean network for each phase is also plotted as a solid square surrounded by a larger square denoting the confidence interval. **Left**: The difference in mean networks between repair and no-repair for each of the first three phases (pre-reference, reference, and monitor).

Discussion of Analyses

The overall goal of these analyses was to develop a more detailed and nuanced understanding of coordinated referential gaze patterns arising in physical dyadic collaborations. In particular, these analyses were conducted in search of answers to three questions: (1) How do a collaborating dyad's gaze behaviors *unfold* over the course of a reference-action sequence? (2) How does the *alignment* of gaze behaviors shift throughout the different phases of a reference-action sequence? (3) How do coordinated gaze behaviors differ in sequences that include breakdowns and/or *repairs*? Due to the highly complex, dynamic, and interdependent nature of coordinated two-party gaze behavior, a relatively new analysis technique was used in order to explore these questions. Epistemic network analysis is ideally suited for analyzing datasets that capture the co-occurrence of social cues, including the gaze behaviors of multiple participants.

Each of the three analyses revealed important properties and patterns of coordinated referential gaze behavior in relation to the three research questions. In the first analysis, ENA was able to characterize and separate the five phases of a reference-action sequence (pre-reference, reference, monitor, action, and post-action). Clear and significant differences were observed in shared gaze behavior across these phases. This analysis also revealed a consistent cyclical pattern of gaze behavior that progresses in an orderly and predictable fashion through the two-dimensional abstract space created by ENA. An important implication of this analysis is that the tracked gaze of a collaborating dyad could be used *in situ* to track their progression through a reference-action sequence. By continuously applying ENA to segments of shared gaze behavior, these segments could potentially be classified according to their location within the ENA space as visualized in Figure 5.3.

The second analysis explored the degree of alignment between gaze behaviors of interacting participants throughout a reference-action sequence. A general rise and fall in alignment was discovered throughout a sequence, as well as a back-and-forth pattern of which participant was leading the interaction in terms of their gaze behavior. The worker's gaze follows the instructor's gaze during the beginning and end of the sequence when the instructor is leading the interaction by producing the verbal reference or preparing for the next sequence. In contrast, the instructor's gaze follows the worker's gaze during the middle of the sequence (monitor and action phases) when the instructor appears to monitor the worker's behaviors as the worker attempts to fixate on the reference object and act on it appropriately.

The third analysis explored differences in gaze behavior between sequences with and without repairs. ENA revealed similar, but characteristically different, patterns of gaze behavior for these two types of sequences. An important implication of this analysis is that,

by tracking the shared gaze of a collaborative dyad, repairs can potentially be anticipated well in advance of their realization. By detecting when the sequence has entered the repair cycle, steps could be taken to quickly resolve any ambiguity or errors and move the interaction back to the non-repair cycle characterizing successful interactions.

There are a number of potential applications that could benefit from the properties and patterns of coordinated gaze discovered in these analyses. In particular, embodied agents could utilize this knowledge to better align their gaze with human interlocutors and improve coordination in collaborative interactions. This application would require a shift from the *descriptive* analyses that were carried out above to the development of *synthesizing* models that generate coordinative gaze behaviors. By synthesizing gaze behaviors appropriately in coordination with the detected gaze of a human interlocutor, the agent could attempt to produce gaze behaviors that follow the same cyclical pattern of natural humanlike gaze coordination as observed in Analysis 1. This idea is pursued further in the next section.

Analyses 2 and 3 similarly have specific implications for modeling and generating gaze behaviors for embodied artificial agents. Analysis 2 sheds light on the role of gaze in "mixed initiative" conversations (Novick et al., 1996). Specifically, the analysis suggests that the agent should shift between leading with its gaze (producing gaze behaviors to which the user is expected to respond) and following the user's gaze (gazing in response to the detected gaze behaviors of the user), as the interaction progresses through the phases of a reference-action sequence. Similarly, following the results of Analysis 3, an agent could recognize misunderstandings by the user before a repair is explicitly and verbally requested, potentially resulting in a more seamless interaction. Furthermore, the agent could make efforts to entirely avoid the patterns of gaze behavior that are characteristic of sequences involving disruptive breakdowns and repairs. These ideas are also pursued in designing gaze coordination mechanisms for embodied agents in the next section.

5.5 Gaze Coordination for Embodied Agents

In order to design and implement gaze coordination strategies for embodied agents, the analyses presented above were utilized and extended to create a computational model of gaze coordination. This section describes a stochastic finite-state-machine model of gaze coordination and a particular system implementation for virtual agents that integrates the gaze coordination model into a collaborative scenario. Reference-action sequences continue to be the fundamental interaction unit used to synthesize gaze coordination techniques. However, for simplicity, the post-action phase and pre-reference phases have been merged, resulting in four phases for each sequence (pre-reference, reference, monitor, action).

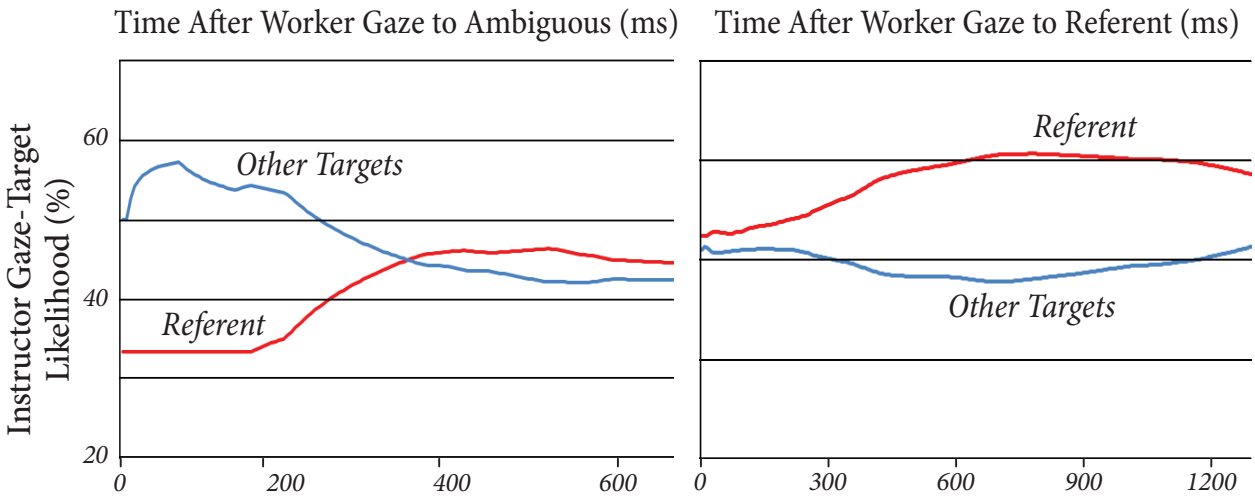

Figure 5.7: Gaze triggers informing the heuristic model component. **Left:** Likelihood of instructor gaze to the referent goes up over time following the worker's gaze to an ambiguous item. **Right:** Likelihood of instructor gaze to the referent goes up following the worker's gaze to the referent.

Verbal references in these sequences are often ambiguous, and gaze is a useful resource for disambiguating them and performing the task efficiently. Therefore, gaze coordination is particularly valuable as it provides a subtle cue for disambiguation without interrupting the flow of the interaction. Disambiguation from gaze is useful in situations where the agent may observe a user's confusion and provide correction. Gaze coordination also has the potential to provide more interaction clarity leading to higher efficiency through fewer errors, alongside the ability to enable agents to use shorter utterances that would otherwise be ambiguous.

Recall that the second analysis above revealed a general rise and fall in alignment throughout a sequence, as well as a back-and-forth pattern of which participant was leading the interaction in terms of their gaze behavior. A model of gaze coordination for agents should reflect this understanding. Early in an interaction sequence, the instructor agent should drive the interaction by finding and referencing an object that it wishes the human worker to find and act upon. Then the worker will take initiative by searching for and acting upon the referent while the agent monitors for potential breakdowns.

The above analyses were intentionally conducted from the perspective of the instructor, which continues to be the focus in this section such that a virtual character could effectively take on that role. In phases where the instructor seemed to be "following" the gaze of the worker, rather than leading, the above analyses were augmented by looking at potentially interesting events or triggers in the worker's gaze stream to see what the instructor is likely

to do in the seconds immediately following that event. For example, when the worker looks at an ambiguous object, what is the instructor likely to look at next? Figure 5.7 (left) depicts the likelihood, over all collected data, of what the instructor is looking at over time following the event of a worker gazing to a wrong ambiguous ingredient, e.g., the light green lettuce instead of the dark green spinach. As can be observed, the likelihood that the instructor gazes toward the correct referent goes up and the likelihood of gazing toward other ingredients goes down, likely as a means to drive attention toward the correct object. Similarly, in Figure 5.7 (right), when the worker does look at the correct referent, the likelihood of the instructor looking towards the referent again goes up over the next second. A variety of phenomena such as these in the data were examined in order to build intuitions, which are synthesized in the heuristic model presented next.

Model

The above analysis suggests that the agent should shift from leading with its gaze during pre-reference and reference phases (producing gaze behaviors to which the user is expected to respond) to following the user's gaze during the monitor and action phases (gazing in response to the detected gaze behaviors of the user). Thus, this model of gaze coordination has two major components: a stochastic component with statistical parameters of what to gaze at when, independent of the user, and a heuristic rule-based component on what to gaze at in response to the user during responsive phases.

For the purposes of modeling, the agent's world is broken down into five categories of gaze targets: the referent, objects that are ambiguous to the referent, the user, the action target (for this scenario, it is always the bread), and all other task-relevant objects.

At the highest level, the model traverses through the separate phases of a reference-action sequence according to the agent's speech and the user's actions. The pre-reference phases is entered at the conclusion of the user's action in the previous reference-action sequence. The reference phase is entered once the agent starts producing the verbal reference. Once the agent finishes speaking the reference, the monitor phase begins. This may involve responding to a user's request for clarification, either explicitly via speech recognition or implicitly from gaze via the heuristic part of the model described below. The action phase begins when the worker begins the relevant action, in this case grabbing the appropriate ingredient. Grabbing the wrong ingredient is considered an error, and the agent remains in the monitor phase, instructing the user to place that ingredient back and locate the correct one, providing additional description to help clarify. If the correct item is grabbed, the action phase lasts until the user brings it to the target bread location. Once

Mean Gaze Fixation Length (seconds)

Phase	Referent	Ambiguous	Other	User	Bread
Pre-reference	0.85	0.45	0.35	0.65	0
Reference	1.1	0.5	0.45	0.6	0
Monitor	1.2	0.6	0.47	1.7	0
Action	0	0	0.66	0.6	0.86

Gaze Shift Probabilities

Phase	to Referent	to Ambiguous	to Other	to User	to Bread
Pre-reference	0.4	0	0.57	0.03	0
Reference	0.48	0	0.41	0.11	0
Monitor	0.49	0.02	0.34	0.15	0
Action	0	0	0.65	0.11	0.24

Table 5.4: **Top:** Mean gaze lengths to targets within each phase. **Bottom:** Probabilities of gazing toward targets within each phase.

that motion is complete, the next reference-action phase begins.

Within each of the four phases of the reference-action sequence, a stochastic finite-state machine is employed to determine which targets to gaze at, and how long to gaze there. This state machine includes five states, one for each category of possible gaze targets. Each of these states is associated with a gaussian distribution (with mean and standard deviation derived from the data collection above) for the length of gaze to the associated target (top of Figure 5.4). These distributions are sampled from to determine the gaze length when producing a gaze shift to that target. Transitions among states (realized as a gaze shift to the target associated with the next state) are dictated by the probabilities of gazing toward each type of target within the current phase (bottom of Figure 5.4). When one gaze fixation is completed, the next target to gaze toward is determined through a weighted random sample of the probabilities of gazing at the other targets. For states containing multiple discrete instances of possible locations, e.g., the "other task-relevant objects" state, one of these instances is randomly selected as the next gaze target.

In general, the model reflects some intuitive trends derived from the data collection, such as longer fixations to the referent, as opposed to other objects, long gazes to the partner while monitoring, and gazes to the bread—the destination of the action—late in the sequence. In addition to the stochastic state-machine, which runs at all times during the interaction, the full model for responsive phases (monitor and action) also includes

heuristically defined triggers of what the agent should do in response to the gaze of the user. These heuristics override anything that is happening in the stochastic state-machine part of the model, which is active at all other times. The heuristics are captured in the following rules.

In the monitor phase (Figure 5.8):

1. *Joint attention following* — When the user gazes toward the referent, the agent also gazes toward the referent.
2. *Shifting joint attention* — When the user gazes toward an object that is ambiguous with the referent, the agent gazes toward the referent.

    a) If the user is gazing toward the ambiguous object for more than one second without fixating on the referent, the agent gazes toward the user and preemptively offers a verbal refinement.

3. *Mutual gaze* — When the user gazes toward the agent, the agent gazes back toward the user.

    a) If the user has not made the correct action within two seconds of gazing toward the agent, the agent preemptively offers a verbal refinement.

In the action phase:

1. *Tracking action intent* — When the user gazes toward the target bread, the agent also gazes toward the target bread.
2. *Mutual gaze and shifting joint attention* — When the user gazes toward the agent, the agent gazes back to the user and then to the target.
3. *Joint attention following* — When the user gazes toward any ingredient on the table, the agent gazes toward that ingredient.

To make successful reference utterances, the agent needs some form of feedback from the human addressee. Despite the best efforts of the agent, there will inevitably be instances of breakdowns—misunderstandings or miscommunication—that can either impede ongoing progress of the interaction or lead to breakdowns in the future (Zahn, 1984). The preemptive verbal refinements present in the heuristic model allow the agent to engage in repair in a natural and efficient way. In addition to these gaze-triggered refinements, the overall system (described below) enables the agent to respond to explicit verbal requests for refinement from the user.

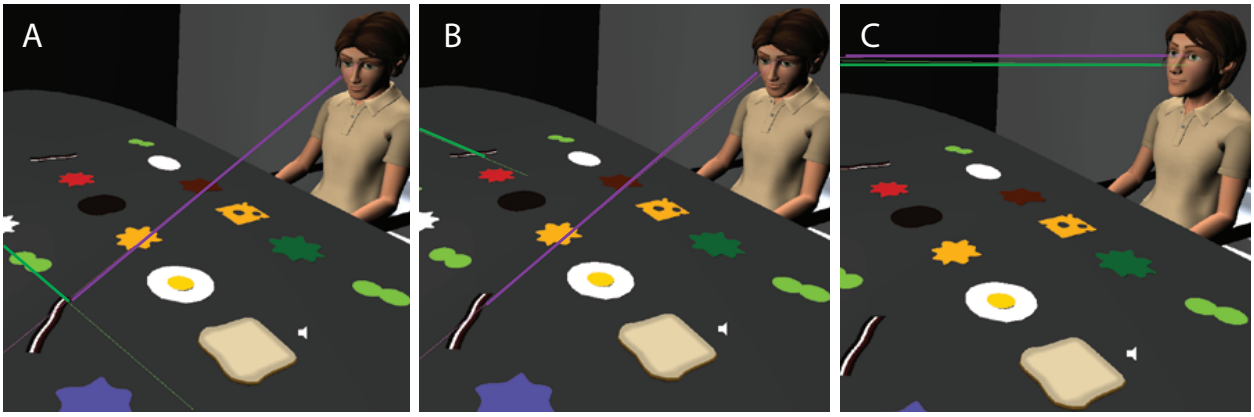

Figure 5.8: Heuristics in the monitor phase. User gaze is shown in green, agent gaze in purple. (A) Joint attention following to the referent. (B) Shifting joint attention from an ambiguous item to the referent. (C) Mutual gaze in response to agent-directed gaze.

System Design

In order to integrate the gaze coordination model into the virtual agent system (Section 1.3), some additional system components are required (Figure 5.9). First, a model of gaze motions is necessary to actually execute shifts in gaze from one target to another. The gaze model presented in Chapter 3 was utilized for this purpose. The agent utilizes full head alignment during gaze shifts toward the referent, target bread, and the user. It uses partial head alignment when gazing toward ingredients that are not the current referent. The gaze shifting of the agent was carefully designed such that gazes to specific real-world objects were perceived as such.

Throughout interaction, the agent requires real-time access to the human interlocutor's point-of-regard. In the first system implementation (Figure 5.10 left), the user wears the same mobile eye-tracking glasses as described in the data collection study to provide the agent with a constant stream of gaze points within the glasses' front-facing camera view. To classify these gaze points in terms of the actual object being looked at, a system of augmented reality (AR) tags are used to convert camera-space points into real-world points. The ArUco library (Garrido-Jurado et al., 2014) was used for the detection of tags, providing the camera-space corner points of any and all tags detected in the camera's view (10 fps). Given (1) these corner points, (2) a nearby gaze tracker point-of-regard, (3) the known real-world dimensions of the tag, and (4) the assumption that the gaze point falls on the same plane as the tag, the Jacobi method is used to solve the homography between camera-space and real-space. This produces the real-world coordinates of the gaze point-of-regard. Tags are arranged on a table to create a grid of 18 locations. Using

Figure 5.9: The system setup includes eye-tracking glasses for the user, AR tags to convert gaze fixations into semantically meaningful locations, speech recognition, task tracking, and the agent.

this system, the agent is provided with real-time access to the grid location being looked at by the human, which the agent can then associate with a particular item given its internal task representation. Two AR tags are also placed vertically around the agent to detect when the agent itself is being looked at by the human.

In follow-up system implementations, discussed below, the gaze tracking component is handled differently. The first follow-up implementation utilizes a Kinect 2 to track the head pose of the user. A virtual ray is extended forward from the point between the eyes and intersected with the task space. The gaze coordination model is relaxed to treat head pose as a gaze "cone" rather than a precise gaze point. For example, a reference gaze is detected when the user's head pose is directed toward the referent or to any objects within one grid cell of the referent in the task space. The second follow-up implementation similarly uses this relaxed version of the model for head-mounted virtual reality, utilizing the Oculus Rift's built-in head tracking as a proxy for gaze direction.

These systems also include speech recognition and task tracking modules. Speech recognition for verbal clarification requests is performed using a microphone and Microsoft Speech. Task tracking in the first two implementations utilizes a ceiling camera focused on the task space. As items are moved, AR tags become revealed, which communicates to the agent that an action has been performed. This action can be compared to the agent's current task model to check if the correct action has been performed, and if not, to issue a

Figure 5.10: **Left:** A user wears eye-tracking glasses to collaboratively assemble a sandwich with a virtual character. **Middle:** The virtual character produces gaze cues to relevant task objects. **Right:** A user interacting with the virtual character in head-mounted virtual reality.

verbal repair to correct the error. In VR, task tracking is accomplished by gazing toward an ingredient and tapping a button to indicate ingredient selections, whereupon the selected ingredient (if correct) is animated to move to the action destination.

5.6 Study 1

Next it was necessary to evaluate the ability of the model of gaze coordination to improve a virtual character's ability to interact with people in natural, engaging, and effective ways. The same interaction scenario that was previously used in data collection was chosen for the evaluation—learning how to make a sandwich.

Study Design

The user study followed a 2×2 within-participants design. The independent variables included whether or not the agent *produced* gaze motions and whether or not the agent *responded to* user gaze, resulting in a total of four unique conditions. Considering both variables separately enabled exploration of the relative effectiveness of a virtual character reacting to user gaze as input and/or producing gaze as output to facilitate interactive experiences. Each participant interacted with the virtual character system four times, one for each condition, in a randomized order.

Procedure

Following informed consent, the experimenter provided the participant with high-level instructions and calibrated the eye-tracking glasses. A virtual character named "Jason" was

introduced, which then instructed participants on how to assemble four sandwiches with different ingredients. The order of the sandwiches, as well as their assignment to condition, was randomized. Each sandwich required 12 ingredients, at least four of which were inherently ambiguous, e.g., the character asked for "cheese" when both swiss cheese and cheddar cheese were present. Following each sandwich, the participants completed a brief survey about the experience with that version of the agent and took a quiz that measured the participant's recall of the ingredients of the sandwich. Following the completion of all sandwiches, the participant filled out a demographic survey and received $5 USD for compensation.

Hypotheses

Our evaluation of the gaze coordination model was guided by three central hypotheses, focusing on the production of referential gaze behavior, the responsiveness to user gaze, and the benefits of doing both simultaneously.

Hypothesis 1—A virtual character that *produces* gaze according to the gaze coordination model will improve user interaction over one that statically gazes toward the user.

More specifically, this hypothesis predicts that producing gaze cues will enable the agent to utilize faster, ambiguous referencing without breakdowns or requests for repair (e.g., clarification), thus decreasing reference acquisition times for the user and increasing user perceptions of the character's competence as an interaction partner. These predictions are well-supported by previous work. In a human-human interaction experiment, a confederate that produced gaze cues resulted in participants more accurately selecting task targets (Macdonald and Tatler, 2013). Previous work has also shown that a virtual talking head mimicking the behaviors of a human speaker can successfully attract users' attention toward regions of interest, resulting in faster task completion and fewer errors (Bailly et al., 2010).

Hypothesis 2—A virtual character that *responds to* user gaze according to the gaze coordination model will improve user interaction over one that does not respond to user gaze.

By tracking and responding to user gaze, the virtual agent will be able to anticipate breakdowns before they occur or before a repair request is made, resulting in higher task efficiency and more rewarding interactions. This hypothesis is supported by previous work that has demonstrated that gaze behaviors precede physical actions (Land et al., 1999; Hayhoe and Ballard, 2005). Characters that respond to user gaze improve learning gains in tutoring scenarios (D'Mello et al., 2012), while robots that respond to user gaze elicit a

stronger sense of being looked at in their users (Yoshikawa et al., 2006).

Hypothesis 3—A virtual character that both *produces* gaze and *responds to* user gaze will further improve user interaction over only producing gaze or only responding to user gaze.

This hypothesis predicts that the effects of producing gaze and responding to user gaze will be collectively achieved when the character utilizes the complete gaze coordination model. Users will also perceive the character to be more responsive when it gazes reactively to objects and dynamically builds mutual gaze with its user. Previous work has shown that shared gaze between pairs of people improves performance in visual search tasks due to faster spatial referencing of the target (Neider et al., 2010). The degree of gaze coordination between a speaker and a listener also predicts their performance in answering comprehension questions (Richardson and Dale, 2005).

Measures

The study included objective, behavioral, and subjective measures to assess the quality and effectiveness of the interactive experiences resulting from a character's use of gaze coordination. The objective measures included task measures such as completion time and number of errors (i.e., incorrect actions taken) measured both within each reference-action sequence and across the making of the sandwich. The number of verbal requests the participant made for clarification was also counted. Finally, a quiz asked the participant to list the ingredients used for each sandwich to measure recall of the character's instructions.

Participant gaze was recorded in all conditions to extract behavioral measures. One such measure was the amount of shared gaze—percentage of time that the character and user looked toward the same object or toward each other—to determine whether or not gaze coordination resulted in more behavioral synchrony. The percentage of time that users looked toward ambiguous objects not relevant to the interaction was also included as a potential indicator of confusion. Finally, the time from the onset of the character's verbal reference to when the user's gaze fixated on the referent was also measured.

The study included a subjective questionnaire about the participant's impressions of the agent that they responded to after completing each sandwich. The full questionnaire is provided in Appendix A. Using an exploratory-confirmatory factor analysis, four scales were constructed from this questionnaire, each comprised of several seven-point-rating-scale questions of the form, "I believe that this version of Jason was...:"

1. *Task competence*: "engaged in the task," "dedicated to the task," "an active part of task," "helpful" (Cronbach's $\alpha = .88$);

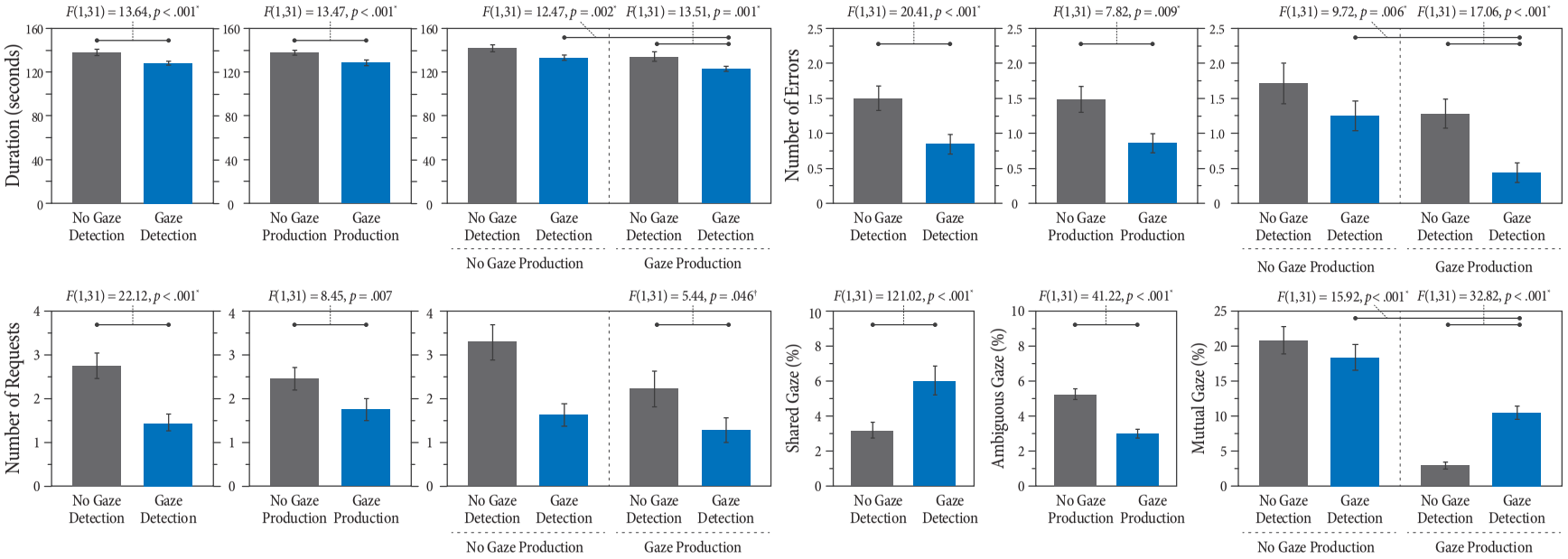

Figure 5.11: Results from the objective measures of task duration (seconds), number of errors, and number of clarification requests as well as behavioral measures of shared, ambiguous, and mutual gaze (%). Test details are provided only for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha levels for multiple comparisons ($\alpha = 0.05$ for H1 and H2 and $\alpha = 0.025$ for H3).

2. *Cognitive abilities*: "sensitive to my needs," "intelligent," "an expert" (Cronbach's α = .87);
3. *Expressiveness*: "lively," "expressive," "excited to help me," "humanlike in behavior" (Cronbach's α = .91);
4. *Visual attentiveness*: "watchful," "attentive," "observant" (Cronbach's α = .91).

Participants: Forty participants from the University of Wisconsin–Madison campus participated in the user study. Due to occasional system malfunctions, eight were excluded from analysis, resulting in 32 participants (14 females, 18 males). Participants were primarily university students with a range of fields of study, including nursing, engineering, film studies, computer science, and environmental science, and their ages ranged from 18 to 32 ($M = 22.4$, $SD = 4.0$). On a five-point scale from "not at all familiar" to "very familiar," participants reported moderate familiarity with interactive virtual characters ($M = 2.31$, $SD = 1.10$) and robots ($M = 2.63$, $SD = 0.99$).

Results

The data collected from the user study were analyzed with a repeated-measures analysis of variance (ANOVA). The statistical model included two independent variables: *gaze production* (on or off) and *gaze detection* (on or off). Both trial number and the particular sandwich type, e.g., "bacon special" or "turkey special," which varied randomly across conditions, were modeled as control variables. A Bonferroni correction was applied to control for Type I errors in multiple comparisons. All statistical test results are reported in Figures 5.11, 5.12, and 5.13 for ease of reading.

Results from Objective Measures

An analysis of the task duration measure supported all three hypotheses. Participants completed the task more quickly when the character detected and responded to user gaze and when it produced gaze cues compared to when it did not engage these mechanisms. The use of both mechanisms had an additive effect; participants completed the character's instructions more quickly when it used both mechanisms over when it engaged only one of these mechanisms. The data on the number of errors participants made in following the instructions similarly supported the hypotheses. Gaze detection and gaze production both reduced the number of errors participants made, and the use of both mechanisms further reduced errors over the use of one of these mechanisms (Figure 5.11).

The data provided support for H1 and H2 and partial support for H3 in the analysis of the number of requests participants made for refinement. While gaze detection and gaze

production both reduced the number of requests made, the use of both mechanisms further reduced requests only marginally over gaze production alone (Figure 5.11). The character's use of the gaze detection mechanism improved participant recall of the ingredients of the sandwich after the task, while its use of gaze production had no effect, and the combined use of both mechanisms did not further improve recall (Figure 5.12).

Results from Behavioral Measures

Analysis of the participants' gaze focused on three specific behavioral measurements: *shared gaze* toward objects of interest, shared gaze toward irrelevant objects (referred to as *ambiguous gaze*), and *mutual gaze* toward each other. This analysis found higher shared gaze when the agent used the gaze detection mechanism, indicating a higher degree of behavioral synchrony. There was also less ambiguous gaze when the character produced gaze, reducing the amount of time participants spent looking toward objects that were not relevant to the task. Finally, when the agent produced referential gaze, the analysis found higher mutual gaze when the agent also detected and responded to user gaze (Figure 5.11).

The character's production of gaze marginally reduced the amount of time it took participants to fixate on referent objects, indicating effective engagement of joint attention between the character and its user. The character's use of gaze detection did not affect participant fixation time. Reference ambiguity, i.e., whether or not multiple potential referents were present, delayed the ability of the participants to identify and fixate on the referent. When references were ambiguous, the character's gaze reduced participant fixation time, while it had no effect when the references were not ambiguous.

Results from Subjective Measures

The analysis next tested the hypotheses using data from the four subjective measures, including how competent, cognitively able, expressive, and attentive participant rated the character on a seven-point scale. This analysis found support for H1 in ratings of the character's cognitive ability and partial support in ratings of its competence; participants found the gaze-detecting character to be more cognitively able and marginally more competent. Support for H2 was provided by data from the perceived competence and perceived expressiveness measures, and partial support was provided by the perceived awareness measure. The gaze-producing character was rated as more competent and expressive and marginally more attentive. Finally, only data from the perceived expressiveness measure provided partial support for H3; the use of both the detection and production mechanisms

resulted in higher ratings of expressiveness over the use of gaze detection alone but not over the use of gaze production alone (Figure 5.13).

Qualitative Responses

After making each sandwich, participants were interviewed using a semi-structured set of questions to solicit open-ended feedback. Analysis of the interview transcripts indicated that, overall, participants had clear awareness of the character's gaze behaviors, interpreted them appropriately, and utilized the information it provided in their task when the character engaged the gaze production mechanism. The excerpts below illustrate participant descriptions of the character's gaze behavior.

- *"Jason was very attentive and seemed more human-like. He looked at which ingredient was to be placed on the sandwich and that helped me."*
- *"I liked that he always looked down when I placed something, it made me feel like he knew which one I selected and that it was right."*
- *"Jason's head moved and he actually looked at the food to which he was referring, that helped a lot."*

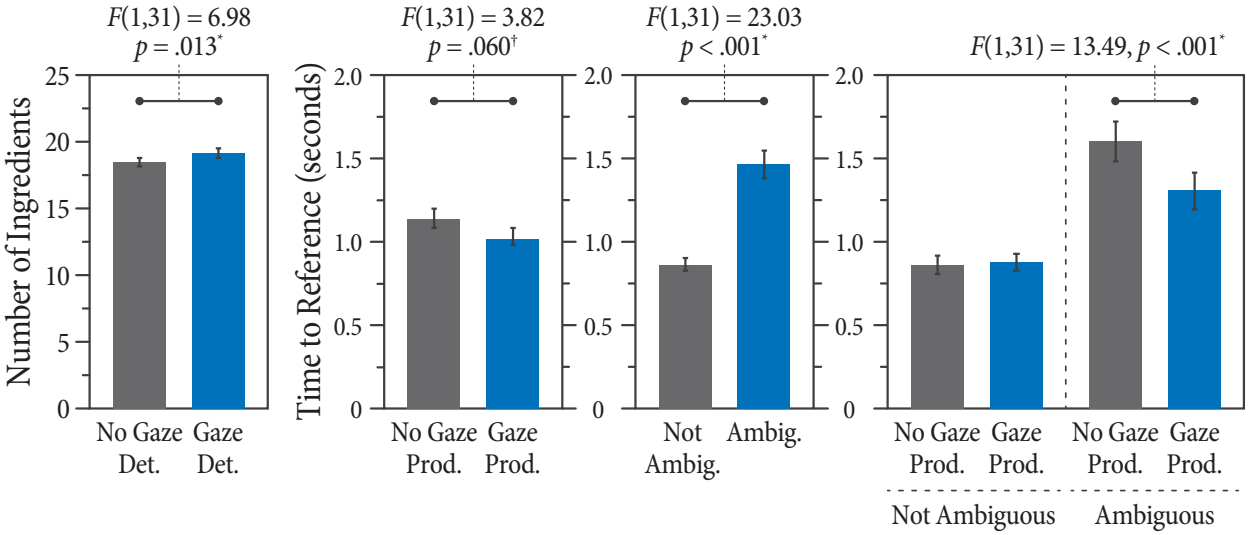

Figure 5.12: Results from measures of information recall and time it took participants to look toward the referent. Test details are provided only for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha levels.

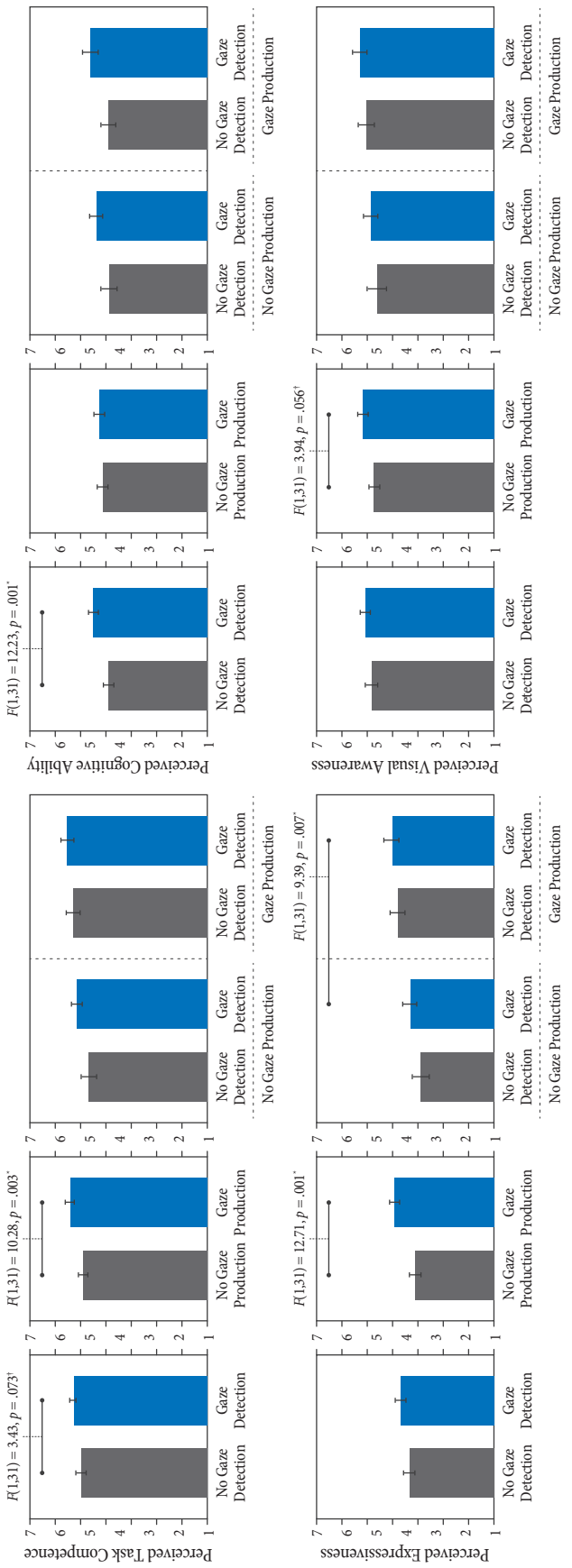

Figure 5.13: Results from subjective measures of how competent, cognitively able, expressive, and aware participants found the character to be. Test details are provided only for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha levels for multiple comparisons ($\alpha = 0.05$ for H1 and H2 and $\alpha = 0.025$ for H3).

- *"He was able to 'point with his eyes' at the different parts of the sandwich which helped clarify what he was talking about."*
- *"Clear instructions were given with good eye contact to direct what ingredients were needed for the sandwich"*

Participants also commented on the character's responses when it engaged the gaze detection mechanism, highlighting its attempts to establish joint attention to disambiguate references under high ambiguity or when the participant fixated on ingredients that were not relevant to the task. Below are excerpts that illustrated participant descriptions of the character's gaze-detection behaviors.

- *"I liked that he looked toward the ingredient I was going to need, since that helped me find it faster. I also liked that he would clarify and describe what the ingredient looked like in case I was unsure."*
- *"He could sense my uncertainty about certain ingredients without me speaking up and it made it easier for me to know what ingredients to put on."*
- *"He clarified the specific ingredient he was talking about before I needed to ask most of the time. He was looking around at the stuff I need to grab."*
- *"I liked that Jason clarified which ingredient he meant before I had to ask."*

Discussion

The first user study was designed to evaluate the ability of the gaze coordination model to improve a virtual character's ability to interact with people in natural, engaging, and effective ways. Conditions were tested in which the agent did or did not produce gaze cues to the task space, and in which the agent did or did not respond to the gaze cues of the user. This study demonstrated the benefits of gaze coordination in objective, subjective, and behavioral measures. Participants conducted the collaborative sandwich-task more efficiently when the agent used the full gaze coordination model, completing the task faster with fewer errors and less need to ask the agent for clarification. Participants also scored better in a recall quiz and engaged in higher amounts of mutual gaze and shared gaze with the agent, indicating a higher degree of coordination. Subjectively, participants felt that the agent was most expressive, most competent, and possessed greater cognitive abilities when using gaze coordination.

5.7 Study 2

One of the most promising applications for utilizing gaze coordination is in virtual reality systems. Virtual agents hold great promise to create compelling interactive experiences in VR, but it will be particularly important for these agents to coordinate their behaviors with users of these systems. Unfortunately, most VR systems do not include eye tracking, instead only tracking the user's head pose. This observation led to the following question: Does the gaze coordination model achieve comparable effects when only the head direction is known and without eye tracking?

To answer this question, a second user study was conducted to compare eye tracking and head tracking using an on-screen agent identical to the first study. Head pose detection was implemented using the Kinect 2. The gaze coordination model was relaxed to treat head pose as a gaze "cone" rather than a precise gaze point. For example, a reference gaze was detected when the user's head pose was directed toward the referent or to any objects within one grid cell of the referent in the task space. All other aspects of the model were identical to what was described above.

This study included three conditions: no gaze detection, gaze detection via eye tracking, and gaze approximation via head tracking. In all conditions, the agent produced gaze cues according to the stochastic finite-state-machine component of the model, but the responsive heuristic part of the model was only utilized in the latter two conditions. The overall procedure was similar to the previous user study, with participants assembling three sandwiches (one for each condition) in a random order. Fifteen participants were recruited, different from those who participated in the first study. After excluding three due to system difficulties, the study was left with 12 final participants for analysis (4 females, 8 males), aged 18–23 ($M = 21.2$, $SD = 1.57$). Participants were once again compensated $5 USD.

Results

The analysis utilized a repeated-measures analysis of variance (ANOVA) and pairwise comparisons with Bonferroni correction to establish benefits of the gaze-approximation method over no gaze detection and the equivalence of gaze approximation and gaze detection. To establish equivalence, guidelines suggested by Julnes and Mohr (1989) were followed, including a p-value larger .50. The analysis showed that participants took significantly shorter time to complete the sandwich, made significantly fewer errors, and made marginally fewer requests for clarification when the character used the gaze approximation method via head tracking compared to not utilizing gaze detection. On the other hand, gaze

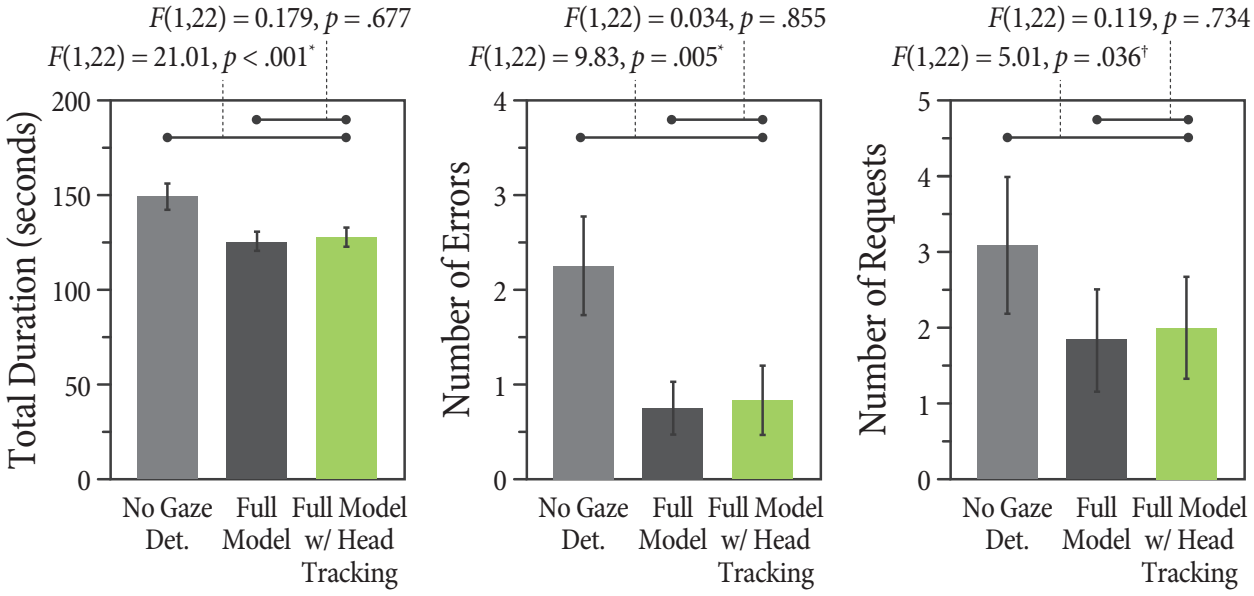

Figure 5.14: Results from measures of duration, number of errors, and requests for clarification. Test details are provided only for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha level for multiple comparisons ($\alpha = 0.025$).

detection based on approximation via head tracking and via eye tracking were equivalent across these measures (Figure 5.14).

Discussion

The second user study was designed to evaluate whether head tracking could serve as a sufficient proxy to more complex eye tracking in the gaze coordination model. The study confirmed this to be the case in terms of objective task performance. Participants completed the sandwich task more efficiently, faster with fewer errors and requests for clarification, when the agent reacted to the participant's head pose rather than when it performed gaze cues unreactively. The analysis yielded no significant differences between using gaze coordination with eye tracking in comparison with head tracking, but a future study with more participants might be able to tease out the small differences.

5.8 Study 3

Once it was confirmed in the second user study that head tracking could serve as an adequate proxy for eye tracking in the gaze coordination model, the following question was next asked: Does the gaze coordination model also work when interacting with agents

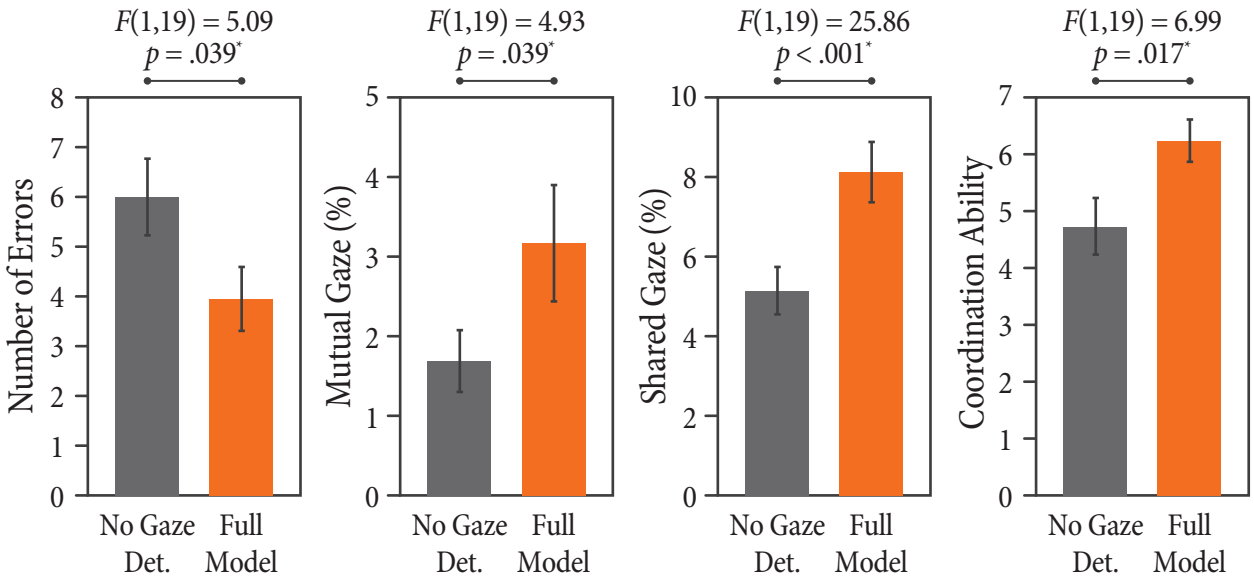

Figure 5.15: Results from measures of number of errors, mutual gaze, shared gaze, and perceived ability of the character for coordination. Data and test details are provided only for significant (*) differences.

in virtual reality? To answer this question, a third user study was designed and executed in which participants carried out the same sandwich-building task with the same virtual agent as in the previous studies, but this time while wearing the Oculus Rift DK2 headset (Figure 5.10 right). Actions were performed in this implementation by gazing toward a desired ingredient and pressing a button to select it. Everything else about the interaction was identical to the previous implementations.

This study tested two conditions: an agent gazing responsively using the full gaze coordination model vs an agent only producing gaze cues with no responsiveness. Participants tried both conditions in a random order and were compensated $3 USD. Twenty participants were recruited (8 females, 12 males), aged 18–47, ($M = 23.1$, $SD = 6.81$). A new subjective scale was utilized for this study to target users' perceptions of the agent's level of coordination with them. This scale included the following items on a seven-point scale: "How would you rate Jason's *responsiveness* to your attention and behaviors?" and "How *in sync* were you and Jason?"

Results

As in the analyses described above, data from the objective, behavioral, and subjective measures were analyzed using a repeated-measures analysis of variance (ANOVA). The

analysis showed that while the use of the full model reduced errors, it did not significantly improve the time it took them to complete the sandwich. Participants established more mutual and shared gaze with the character that used the full model compared to the character that only used the gaze production mechanism. Finally, the character that used the full model was rated by the participants as more effective in coordinating its behaviors with them (Figure 5.15).

Discussion

The third user study was designed to evaluate the effectiveness of gaze coordination for virtual characters in head-mounted virtual reality. This study revealed that an agent using the full gaze coordination model, rather than producing gaze in isolation, is able to help participants complete a collaborative task with fewer errors. Participants also made more mutual gaze and engaged in more shared gaze with the agent when it reacted to their gaze, indicating a higher degree of coordination. Participants explicitly indicated that they subjectively felt more coordinated with the agent when it was using gaze coordination than when it was not.

5.9 General Discussion

Study 1 served to demonstrate the effectiveness of gaze coordination, differentially examining both the gaze production and gaze detection components of the model. In general, this evaluation showed that gaze coordination mechanisms are effective in improving collaborative interaction with on-screen virtual characters providing instructions over a physical task space. Study 2 demonstrated that relaxing and extending the model to utilize head pose rather than full eye tracking retains much of that effectiveness. Study 3 demonstrated that the gaze coordination model utilizing head tracking is particularly useful in virtual reality collaborations with virtual characters.

People are able to use a range of subtle verbal and non-verbal communication mechanisms to collaborate and correct each other's mistakes and misunderstandings quickly and efficiently. Gaze coordination mechanisms have similarly proven to enable interactive virtual characters to preemptively offer refinements in a quick and unobtrusive way, responding to subtle nonverbal gaze of the user rather than relying on explicitly spoken questions. People also frequently produce ambiguous speech that they are able to easily resolve through context and the use of nonverbal cues, making speech faster and more efficient. Gaze coordination enables agents to tap into this power, making potentially

ambiguous but faster verbal references in conjunction with appropriate gaze production and detection.

A key aspect of this model is that it is grounded in observations of human-human interactions in order to ensure that it accurately captures human communication mechanisms of gaze coordination. Evaluations were driven by hypotheses derived from research on human-human interaction. This human-based modeling enabled targeting human-level competence while making specific predictions about how gaze coordination mechanisms should improve interactions objectively, subjectively, and behaviorally. The computational model was designed with separate modules for driving gaze and responding to gaze, allowing to activate them separately and test their respective contributions in the model evaluation.

5.10 Chapter Summary

When conversing and collaborating in everyday situations, people naturally and interactively align their behaviors with each other across various communication channels, including speech, gesture, posture, and gaze. Having access to a partner's referential gaze behavior has been shown to be particularly important in achieving collaborative outcomes, but the process in which people's gaze behaviors unfold over the course of an interaction and become tightly coordinated is not well understood. The first half of this chapter presented work to develop a deeper and more nuanced understanding of coordinated referential gaze in collaborating dyads. Thirteen dyads were recruited to participate in a collaborative sandwich-making task and used dual mobile eye tracking to synchronously record each participant's gaze behavior. A relatively new analysis technique—epistemic network analysis—was used to jointly model the gaze behaviors of both conversational participants. In this analysis, network nodes represent gaze targets for each participant, and edge strengths convey the likelihood of simultaneous gaze to the connected target nodes during a given time-slice. Collaborative task sequences were divided into discrete phases to examine how the networks of shared gaze evolved over longer time windows. This chapter presented three separate analyses of the data to reveal (1) properties and patterns of how gaze coordination unfolds throughout an interaction sequence, (2) optimal time lags of gaze alignment within a dyad at different phases of the interaction, and (3) differences in gaze coordination patterns for interaction sequences that lead to breakdowns and repairs.

The second half of the chapter explored how embodied agents stand to benefit from tracking user gaze and the knowledge of coordinated gaze patterns. Grounded in the data

collected from pairs of collaborating people, a number of subtle features of human gaze coordination were identified, including timings, spatial mappings, and repair strategies. These features were built into a model of gaze coordination that enables interactive virtual characters to interpret the gaze of their users and generate their own gaze to effectively communicate coordinated behaviors. This model enables virtual characters to achieve more efficient verbal referencing by signaling attention to the user and to items in the environment appropriately over time, and infer the user's current state and goals—such as confusion leading to an impending request for repair—from the user's gaze.

A user study demonstrated that gaze coordination mechanisms improve efficiency, subjective quality, and the overall level of behavioral coordination between people and interactive agents. In order to make the model more practical for users without access to eye tracking systems, techniques were introduced that allow head tracking to serve as a proxy for more precise eye tracking. A second user study demonstrated that the relaxed model retains much of the effectiveness of gaze coordination that was achieved by full eye tracking. Furthermore, the third user study demonstrated that using the model with the head tracking available in a head-mounted display provided sufficient fidelity to improve user experience and interaction with a character in a virtual reality setting. In general, gaze coordination is a powerful strategy that results in more immersive and fluent interactive experiences with virtual characters.

Overall, this chapter provided a new understanding of gaze coordination in dyadic physical collaborations, as well as an implementation and evaluation of gaze coordination behaviors for embodied agents. However, all of the gaze mechanisms presented thus far do not account for individual differences in user characteristics; they are applied the same way for all human users. This one-size-fits-all approach potentially limits the agent's ability to act most effectively for individual users. The next chapter addresses this issue and presents a mechanism of *gaze adaptivity* that demonstrates how the timing of an agent's gaze shifts can be manipulated in order to express extroversion or introversion, and how this personality expressed via gaze can be matched to a user's personality in order to improve motivation in a rehabilitation setting.

6 GAZE ADAPTIVITY

To be truly effective, embodied agents must be able to *adapt* their gaze behaviors in two ways. First, they must adapt to the unique characteristics of users, and second, they must adapt to changes in user needs and behaviors throughout an interaction and across multiple interactions.

This chapter investigates how agents might achieve the first form of adaptation by presenting the design and evaluation of gaze behaviors for *socially assistive robots* that enable the robot to match the personality of the user, thereby more effectively motivating users to repeatedly engage in a therapeutic task. This chapter focuses on the extroversion dimension of the Big Five personality model (John and Srivastava, 1999) as it is the most accurately observable dimension of personality expressed by nonverbal behaviors over short timescales (Lippa and Dietz, 2000). The evaluation also demonstrates the importance of taking the user's intrinsic motivation into account when attempting to motivate and increase compliance.[1]

Research Questions

- How should an agent adapt its gaze to the individual characteristics of its human user?
- How does a person's gaze motions reflect their underlying personality, specifically extroversion?
- How can we design gaze behaviors for agents that similarly signal a specific personality type?
- How do people with different personalities utilize gaze when attempting to motivate others to carry out tasks?
- How can we design gaze behaviors for therapeutic agents that are motivational for human users in rehabilitation with different personalities?

6.1 Related Work

Three threads of research inform this work, including previous work on socially assistive robots, previous work on adapting technologies to the characteristics—especially

[1] Portions of this chapter were published in Andrist et al. (2015).

personality—of users, and social-science research on the relationship between nonverbal behaviors, personality, and motivation.

Socially Assistive Robots

In the area of rehabilitation, social robots hold great promise for improving the quality of life of the elderly, individuals with physical impairment, and those with cognitive disorders. Social robots envisioned for use in these contexts are referred to as *socially assistive robots* (Feil-Seifer and Mataric, 2005). The purpose of these robots is to assist users by providing information, motivation, and feedback in order to increase compliance with an exercise regimen, take medication on a schedule, perform repetitive tasks for physical or cognitive therapy, and so on. Socially assistive robots are currently being developed to work as caregivers alongside doctors, nurses, and physical therapists (Kang et al., 2005); therapy aids for children with autism (Feil-Seifer and Matarić, 2009; Scassellati et al., 2012); and as companions in nursing homes (Fasola and Matarić, 2012; Marti et al., 2006).

Patient compliance with treatment for chronic diseases in the U.S. is often below 50%; over half of patients do not take their medication correctly (Osterberg and Blaschke, 2005). Following a stroke, one of the most effective rehabilitation methods is for patients to repeatedly exercise the affected limb(s), an activity patients find quite difficult to keep up without a therapist present (Eriksson et al., 2005). A powerful way of improving motivation and compliance in such settings is through nonverbal behaviors and adapting behaviors to the patient's characteristics. Socially assistive robots are particularly engaging in such scenarios due to their physical embodiment and ability to employ nonverbal cues.

Prior work has established that the mere presence of a robot can positively affect user compliance in an array of contexts. In previous work where participants were recruited for a weight-loss program, Kidd (2008) found that adherence to the program was shortest when participants tracked their progress with pen-and-paper, longer with a computer interface, and significantly longer when reporting their progress to a robot. Even a very simple robot—a Roomba vacuum cleaner robot augmented with a facial display—has been shown to be capable of helping people with medication compliance (Takacs and Hanak, 2008). Social facilitation theory (Mumm and Mutlu, 2011a) provides some explanation for the effectiveness of robots in these contexts; the presence of an embodied humanlike robot increases motivation in the same way that the presence of other people increases an individual's drive and enhances their performance in tasks in which the individual is skilled. The robot's physical embodiment and shared physical context create an opportunity for strong engagement between the robot and the user.

To improve effectiveness, the robot must relate to the user with praise and feedback on their actions. Previous research has examined the positive effects of relational discourse—including praise and feedback—in an exercise coach robot leading elderly participants in physical and cognitive exercises (Fasola and Matarić, 2012). Relational agents have also been found to be effective for delivering health communication and health behavior change interventions to older adults, especially those with low functional health, reading, or computer literacy (Bickmore et al., 2005).

Previous research has also investigated the ability of social robots to provide cognitive therapy to users, such as older adults with mild cognitive impairments that need help planning and executing everyday activities (Bruno et al., 2013). Robot therapists for the elderly have also been employed to play memory enhancing music games with their users (Tapus et al., 2009a). For individuals suffering from dementia and/or other cognitive impairments, socially assistive robots have been shown to improve, through social interaction, the cognitive abilities of their users, and thus their quality of life (Tapus et al., 2009b). Robots are also educationally useful interventions to improve social interactions for individuals with Autism Spectrum Disorders (ASD) (Nikolopoulos et al., 2011).

Socially assistive robots are capable of providing physical therapy for users without needing to make physical contact. Individuals who have recently suffered a stroke may benefit from the use of robots in this domain. Previous work has investigated the influence of robot coaching styles designed to enhance motivation and encouragement on post-stroke individuals during motor task practice (Wade et al., 2011). In other prior work, a robot asked healthy users to engage in physical exercises similar to those used during standard stroke rehabilitation, such as repeatedly lifting and moving books or pencils. Participants were asked to perform the tasks for as long as they wished. Researchers manipulated the interaction style of the robot and found that extroverted participants preferred and complied more with a robot that challenged them rather than one which focused on nurturing praise (Tapus and Matarić, 2008). In subsequent work, the robot adapted its behavior to match each participant's preferences in terms of therapy style, interaction distance, and movement speed (Tapus and Mataric, 2008). In that work, as well as in research presented in the next section, it is shown that adaptation is key to creating positive interactions, especially in assistive contexts. This chapter aims to further this research by presenting personality-expressing gaze behaviors and demonstrating that they must be targeted to the personality of users.

Adapting to Users

Previous research in HRI has demonstrated the benefit of a robot adapting to its users. In prior work in which a robot provided cooking help to participants in a kitchen, researchers found that adaptive dialogue—in which the robot adapted the content of its speech depending on the expertise of the user—improved information exchange and social relations, especially when users were under time pressure (Torrey et al., 2006). Previous research with children investigated the use of adaptive empathic behaviors (e.g., encouraging comments and offering help) for a chess-playing robot (Leite et al., 2012). Children responded positively to the robot when the empathic behaviors were employed adaptively, rather than randomly. Previous work has also investigated the positive effect in eliciting user compliance of matching a "playful" or a "serious" robot (conveyed through the robot's speech) to a playful or serious task (Goetz et al., 2003).

Adapting to user personality has been more widely studied in HCI. For example, previous research shows that computer interfaces can be manipulated to exhibit an extroverted or introverted personality through the use of language, pictures, and sounds, and that introverted users will perform tasks faster when using introverted software (Richter and Salvendy, 1995). This result follows from the theory of *similarity-attraction* which predicts that a person will be attracted more to others who match their personalities than to those who mismatch (Lee and Nass, 2003). In the same way, matching the personality of a synthesized voice, expressed through pitch, prosody, and so on, to user personality positively affects users' feelings of social presence, especially in extroverts (Lee and Nass, 2003). Emotion matching is also important in computer interfaces. In a study of emotional speech generation for car interfaces, it was found that when the emotion expressed by the car voice (energetic or subdued) matches the emotional state of the driver (happy or upset), drivers have fewer accidents, attend more to the road, and speak more to the car (Nass et al., 2005). This chapter parallels these efforts by following similarity-attraction theory to match the gaze behaviors of an embodied agent to the personality of the user.

Nonverbal Behaviors, Personality, and Motivation

Nonverbal behaviors, especially gaze, have long been recognized in social-sciences literature as useful tools in persuading others to comply with requests or demands. A number of theories have been proposed to explain this phenomenon, including speech accommodation theory, demand theory, and arousal intimacy theory (Segrin, 1993). According to *speech accommodation theory*, people may change their communication behaviors when interacting with others and convergence toward the style of the partner should produce a positive

attitude in the partner and thus lead to compliance. According to *demand theory*, nonverbal behaviors can function as demands (e.g., staring as a demand for a response), producing a level of arousal that targets can alleviate by complying with any implicit or explicit demands. In *arousal intimacy theory*, nonverbal behaviors are predicted to produce compliance because they produce greater perceptions of intimacy between the source and target, leading to compliance when the target experiences positive arousal. Each of these three theories predicts a strong relationship between nonverbal behaviors and compliance.

A large amount of previous research has empirically demonstrated the positive effect of gaze on compliance (Chapter 2). When a collector of money for charity engaged in mutual gaze with possible donors, rather than looking at the collecting tin, they were more successful in receiving donations (Bull and Gibson-Robinson, 1981). Nonverbal behavioral cues, such as gaze, gesture, and proxemics, are sometimes referred to as *immediacy* cues (Christophel, 1990). Students are more likely to comply when they perceive their teachers as moderately to highly immediate, and are more likely to choose to reject requests made by nonimmediate teachers (Burroughs, 2007). Similarly, attraction and dominance increase compliance and cues of such are often expressed through nonverbal cues like gaze (Peters, 2007). Gaze is also closely tied to personality, with extroverts commonly engaging in significantly more mutual gaze with their conversational partners than do introverts (Rutter et al., 1972).

In addition to nonverbal behaviors and personality, attempts to increase compliance in others must take into account a person's motivation, which can vary not only in magnitude but also in orientation (Ryan and Deci, 2000). The most basic distinction in motivation orientation is between intrinsic, which refers to doing something because it is inherently interesting or enjoyable, and extrinsic, which refers to doing something because it is accompanied by external pressure or control. In the current work, socially assistive robots are used to create extrinsic motivation in users, while also taking into consideration the intrinsic motivation that those users have to complete the task.

6.2 Designing Personality-Expressing Gaze

In order to develop adaptive gaze behaviors for a socially assistive robot to employ, it is necessary to first understand the relationship between gaze and personality, focusing particularly on extroversion, in a motivational context. This section presents a human-human data collection study conducted to inform the design of personality-expressive gaze behaviors. These behaviors were then evaluated in an online validation study to confirm that the designed gaze behaviors indeed express the intended personality.

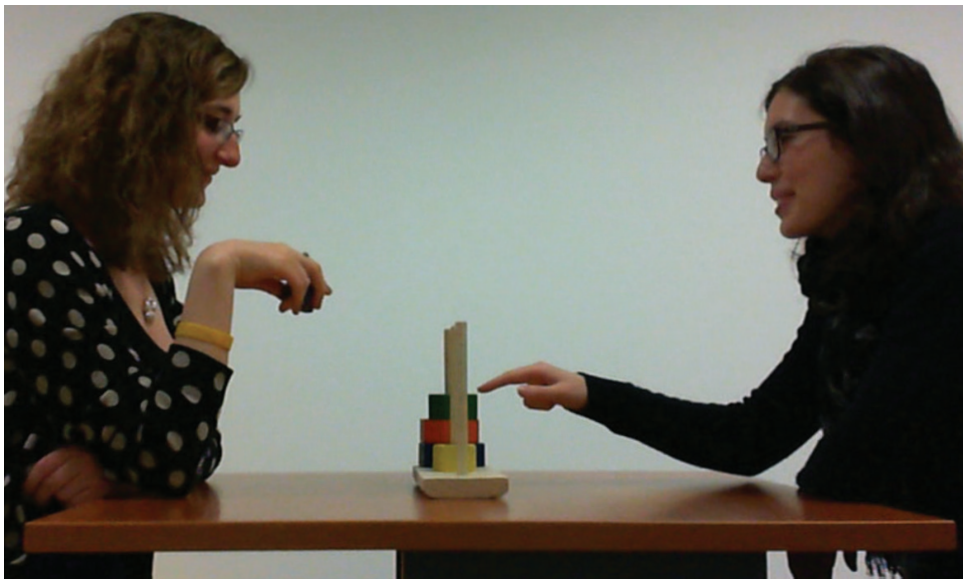

Figure 6.1: Setup of the human-human data collection study. The participant on the right (instructor) is providing extrinsic motivation to the participant on the left (worker) to complete the puzzle.

Data Collection and Modeling

The following questions are relevant to understanding the relationship between gaze and personality in a motivational task. How do people use their gaze when attempting to motivate others and increase compliance? What is the relationship between their use of gaze and their respective personalities? To answer these questions, a human-human data collection study was conducted with four participant dyads, obtaining more precise measurements of gaze behavior than what is traditionally presented in the social-sciences literature. The Tower of Hanoi puzzle was chosen for participants to complete collaboratively. The goal of this puzzle is to move a number of colored blocks from one location to another while following some simple rules. This task was chosen for its mix of cognitive (solving the puzzle) and physical (actually moving the pieces around) elements, mapping well to tasks commonly used in physical and cognitive rehabilitation. The task can also be broken down into two repeating phases common to rehabilitation activities: (1) the actual execution of the task, referred to here as the *in-task phase*, and (2) the time between tasks when the therapist must provide encouragement to persist with the task, referred to here as the *between-task phase*.

Participants filled out the Big Five inventory prior to participation to determine their position on the extroversion-introversion spectrum (John and Srivastava, 1999). The Big Five questionnaire contains 44 items on a five-point rating scale that ask the participant

to rate their agreement or disagreement with statements about their own personality and activities. Eight of these items contribute to the extroversion dimension of the participant's overall personality score. These items have good internal reliability (Cronbach's α = .88). Participants scoring lower than 2.5 on the extroversion dimension were labeled as introverted, and those above 2.5 were labeled as extroverted.

In each dyad, one participant was assigned to be the *instructor* and the other the *worker*. Each of the four dyads covered one of the four possible combinations of participant personality and role. The experimenter first explained the puzzle to the instructor without the worker present. Next, the instructor practiced solving the puzzle with the experimenter. The instructor was required to successfully solve the puzzle a number of times in front of the experimenter to prove that they were comfortable with the task. Then, the instructor was asked to carry out the following procedure: (1) explain the task to the worker, (2) monitor the worker as they complete the task, (3) provide encouraging feedback during the puzzle solving, (4) motivate the worker to keep working between puzzle tasks, and (5) correct workers when they make a mistake. The worker was told by the experimenter that everything would be explained by the instructor, but that they were welcome to work on the task for as long as they wished and that it was up to them to decide when they would like to stop.

The two participants sat at a table facing each other, with the puzzle between them (Figure 6.1). Over-the-shoulder view video cameras recorded the gaze of each participant, and a side camera with wide-angle view was used for recording the entire task. The

Worker Compliance & Mutual Gaze

Outcome	Compliance			Instructor Gaze toward Partner (%)		Worker Gaze toward Partner (%)	
Measure	Time (s)	Puzzles Done	Blocks Moved	In-Task Phase	Between-Task Phase	In-Task Phase	Between-Task Phase
Dyad (Instructor–Worker)							
Extroverted–Extroverted	1369	18	541	7.66	14.74	5.20	11.91
Introverted–Introverted	703	14	352	3.79	8.57	2.74	8.68
Extroverted–Introverted	320	9	227	5.31	10.53	1.77	8.97
Introverted–Extroverted	295	7	121	1.91	4.39	4.94	14.74

Table 6.1: Results from the human-human data collection on worker compliance in each dyad, as well as the amount of partner-directed gaze for all participants.

Gaze Lengths (Mean (Standard Deviation))

Personality	Extroverted		Introverted	
Phase	**In-Task**	**Between-Task**	**In-Task**	**Between-Task**
Partner	2.66 (0.80)	3.91 (1.22)	0.57 (0.19)	1.59 (0.39)
Puzzle	4.04 (2.12)	1.01 (1.26)	11.65 (11.17)	6.21 (8.14)

Table 6.2: Means and standard deviations of gaze fixations (in seconds) to the partner and to the puzzle for extroverted and introverted participants, divided into in-task and between-task phases of the interaction.

compliance of the worker to the instructor was measured in three ways: (1) total time spent solving puzzles, (2) total number of puzzles completed, and (3) total number of puzzle pieces moved. Worker compliance in each dyad, as well as the percentage of each participant's attempt to engage in mutual gaze with their partner, is presented in Table 6.1.

All videos were coded for participant gaze behavior. Participant gazes were recorded and labeled for two targets: the other participant and the shared workspace. The mean and standard deviation of these gaze lengths are presented in Table 6.2.

Three trends can be observed from the data. First, extroverts seem to be attempting to engage in more mutual gaze with their partner than do introverts. This relationship between extroversion and gaze behavior has been similarly demonstrated in previous research, including a study of dyadic interviews in which it was found that extroverts gaze at their interviewer more than introverts do (Iizuka, 1992). Second, there is more mutual gaze between puzzle phases when the instructor is attempting to motivate the worker to solve more puzzles and much less mutual gaze during the actual puzzle completion. Third, there is some preliminary indication that personality matching is an effective strategy for increasing compliance, as personality-matching dyads exhibited longer time-on-task than the mismatching personality dyads. This point is explored further in the experimental evaluation presented later in this chapter.

The results presented in Table 6.2 are used to generate two models of gaze behavior for robots, one to express an extroverted personality and the other to express an introverted personality. The presented means and standard deviations are used to create normal distributions of gaze lengths that the robot draws from when planning and executing gazes toward the user and toward the task space. In an expression of the extroverted model, the

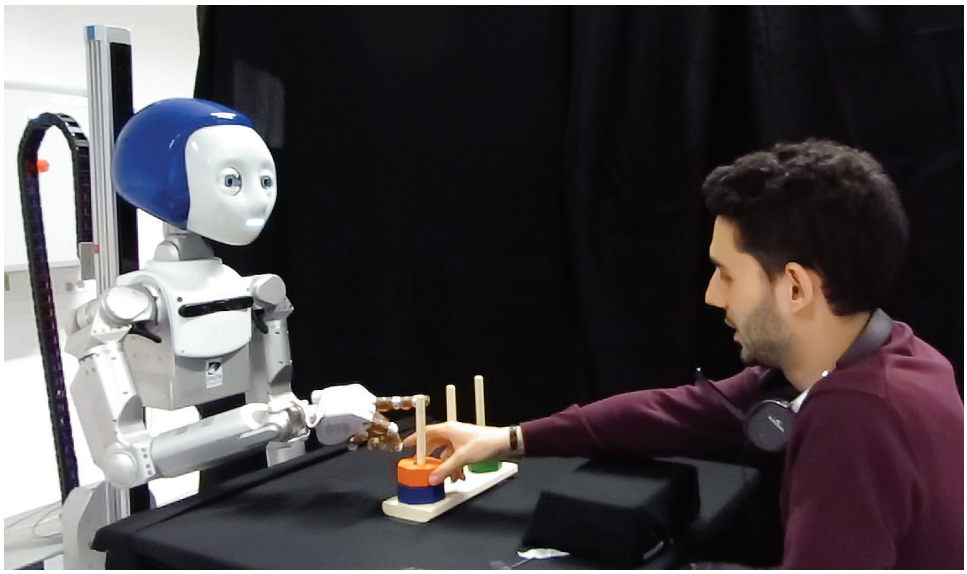

Figure 6.2: The socially assistive robot, Meka, guiding a user through the puzzle-solving task.

robot gazes into the face of the user more, while the introverted model generates more gaze toward the task space. In both models, more gaze is generated toward the user in the motivational between-task phase than in the in-task phase, which involves monitoring the user's actions. For example, an extroverted robot drawing from the distributions in Table 6.2 might generate a four-second gaze toward the user in a between-task phase, followed by a one-second gaze toward the task space. This sequence of long user fixations and short task fixations—randomly generated according to the distributions—would repeat until the start of the next in-task phase. At this point, the extroverted robot might generate a four-second gaze toward the task space followed by a two-and-a-half-second gaze to the user. This cycle of gaze shifts would repeat throughout the in-task phase.

Implementation

Following the data collection and analysis from human dyads was the design and implementation of a system to allow a robot to take on the instructor role in the same puzzle completion scenario. The system was implemented on the Meka robot platform (Figure 6.2). The Robot Operating System (ROS) was used to handle the execution and communication amongst each of the system components described below (Figure 6.3).

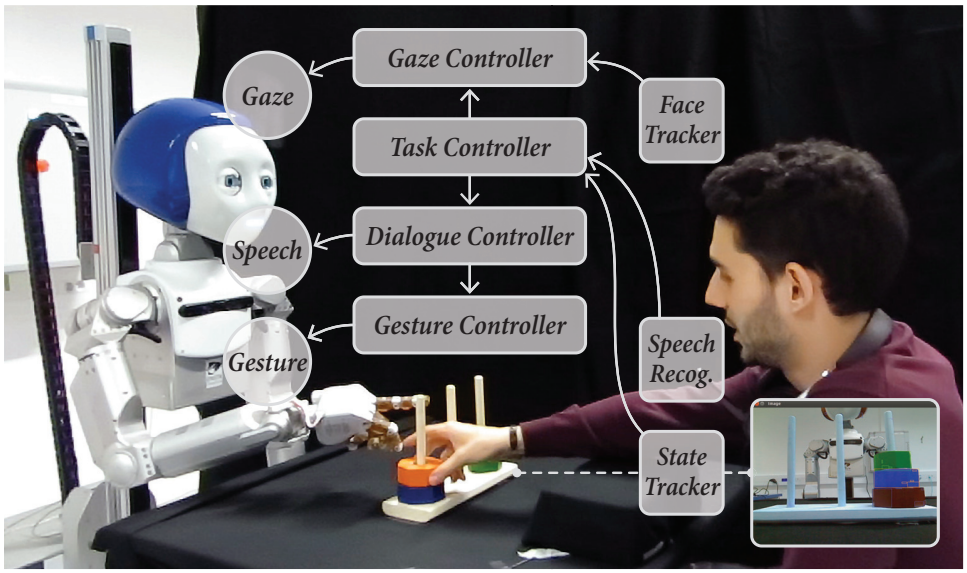

Figure 6.3: The implementation of the socially assistive robot. Participant face location, speech, and task state are tracked and passed to the dialogue controller, task controller, and gaze controller. These controllers determine the gaze, speech, and gestures of the robot. Rounded squares, circles, and rounded rectangles denote sensing, output, and control modules, respectively.

Tracking the participant and task state

A depth camera mounted in Meka's chest was used for face tracking. A small amount of noise is added to the tracking output so that the robot's gaze does not remain motionless when gazing toward the user's face. For puzzle tracking, a separate webcam is placed on the table and focused on the puzzle. Color blob tracking is employed to determine the current locations of each of the different colored puzzle pieces. Google speech recognition is utilized for capturing the user's requests for help and requests to terminate the task.

Task controller

This component manages the flow of the overall scenario. It keeps track of the current task phase (in-task or between-task) and continuously solves the puzzle from the current state so the robot can provide hints and help when requested. The robot can provide general strategies or suggest moves to make in completing the puzzle. If the user makes five bad moves in a row or does not make a move for ten seconds, the robot automatically provides help. The robot also randomly provides positive feedback when it detects that the user has made a good move. Providing these hints and feedback has been shown in previous work to positively affect people's motivation to engage in a task (Vallerand and Reid, 1984).

Generating robot behavior

Depending on the current phase of the scenario (introduction, in-task, between-task, closing), the dialogue controller generates the robot's speech appropriately. If a request for help is detected, the dialogue controller generates the response speech and the gesture controller generates a pointing gesture to the appropriate puzzle piece. The gaze controller generates gaze shifts according to the personality being expressed and the current phase of the interaction. The values in Table 6.2 are used to create distributions that the robot draws from when planning and generating gaze shifts toward the puzzle or toward the user. When gazing toward the puzzle, the robot looks toward blocks that are in motion, creating a stronger sense of responsiveness and lifelikeness.

Model Validation

After implementing the system, an online study was conducted to determine if the extroverted and introverted gaze behaviors generated by the gaze controller would actually be perceived as such. Four one-minute videos were filmed of the robot interacting with a human user in the Tower of Hanoi scenario. In two of the videos, the robot utilized the extrovert model of gaze behaviors, with the other two videos depicting the robot utilizing the introvert model. This study used a within-participant design; all four of the videos were shown to each participant in a random order. Following each video, participants rated the perceived extroversion of the robot on six five-point rating scales which have also been used in previous research to rate the perceived personality of robots (Lee et al., 2006b).

Participants in this study were recruited using Amazon's Mechanical Turk. Thirty participants were recruited (15 Female, 15 Male) with a mean age of 34.9 (SD = 8.5). Standard IP-filtering techniques were employed to limit participation to the United States and prevent multiple participation. In order to eliminate participants that were not focusing on the videos, participants were asked to indicate the color of the robot's head (blue) at the end of the study, discarding data from participants who failed to provide a correct answer. The study also tracked the amount of time that the browser window containing the video stimulus remained in focus on the participant's computer. Eight participants failed the color question and/or unfocused the stimulus window for a majority of the study time and were therefore eliminated from analysis, leaving 22 participants.

The effect of the robot's gaze behavior on participant ratings of the robot personality was analyzed using repeated-measures analysis of variance (ANOVA). Results indicated that participants did perceive a difference in the personality of the robot in the way that was intended. The extroversion rating of the robot was significantly higher in the extrovert gaze

behavior condition (M=3.79, SD=0.54) than in the introvert gaze behavior condition (M=3.53, SD=0.59), $F(1, 84) = 5.09, p = .027$. This result is supported by previous work which demonstrated that participants can accurately recognize a robot's intended personality based on its verbal and nonverbal behaviors (Lee et al., 2006b).

6.3 Experimental Evaluation

The validation study showed that the gaze manipulations indeed resulted in differential perceptions of the robot's personality. Presented next is a more comprehensive study of the effect of using these personality-expressing gaze behaviors in motivational interactions with human users. The goal of this study was to test the effect on compliance of matching or mismatching the robot's personality with that of the user.

Hypotheses

Three hypotheses were developed that predict the effect of matching the personality of the robot—expressed through gaze—to the personality of users, as well as the potential effect of users' intrinsic motivation.

Hypothesis 1—Matching the robot's personality to the user's personality will improve the user's subjective ratings of the robot's performance.

This hypothesis follows from similarity-attraction theory, which predicts that a person will be more attracted to others who match their personality than to those whose personalities do not match (Lee and Nass, 2003). Thus, this hypothesis predicts a strong interaction effect between user personality and robot personality.

Hypothesis 2—Matching the robot's personality to the user's personality will improve compliance with the robot's requests to engage in the task for a longer period of time.

This hypothesis also predicts a strong interaction effect between user personality and robot personality and follows from similarity-attraction theory. Previous work in socially assistive robots found a similar interaction effect on compliance between user personality (extrovert or introvert) and robot therapy style (challenging or nurturing) (Tapus and Matarić, 2008). The data collection presented above also lends some preliminary support for this hypothesis, as the two personality-matching dyads participated in the puzzle task longer than both personality-mismatching dyads.

Hypothesis 3—The user's intrinsic motivation for the task will interact with the personality-matching effect on compliance. Users with low intrinsic motivation will be more affected by personality-matching than users with high intrinsic motivation.

If a user has high intrinsic motivation for the puzzle-solving task, they would inherently find the task interesting or enjoyable and thus would not respond to external motivation attempts from a socially assistive robot (Ryan and Deci, 2000).

Participants

Forty participants were recruited for this experiment (16 Female, 24 Male) from a university campus. Participant ages ranged from 20 to 58 (M = 30.6, SD = 9.4). For feasibility purposes, all participants were healthy adults without need of physical or cognitive therapy, a strategy that has been employed in previous research on socially assistive robots (Tapus and Matarić, 2008). Participants' countries of origin included France, the United States, China, Romania, and Tunisia, and came from both technical and non-technical backgrounds. The experiment was implemented in both English and French (including the script of both the experimenter and the robot), and participants were allowed to choose the language with which they were most comfortable (10 chose English, 30 chose French).

Study Design & Procedure

The study followed a 2×2 between-participants study design, with participant personality (extrovert or introvert) and robot personality (extrovert or introvert) comprising the two factors. The study contained four total conditions representing each of the four personality combinations, with ten participants recruited for each of these conditions. The robot's gaze behavior was the only difference between robot personality conditions; experimenter instructions and the content of robot speech were held constant for all conditions.

All participants were asked to complete and submit the Big Five personality inventory prior to participation to determine their position on the extroversion-introversion spectrum (John and Srivastava, 1999). These items, as in the human-human task, took the form of five-point rating scales. A median split was used to separate participants into two groups. All participants with an extroversion score less than or equal to 3.0 were labeled as "introverted," with participants scoring higher than 3.0 labeled as "extroverted." Participants were also asked to complete and submit a questionnaire to assess their global motivation toward activities in their life. This questionnaire contains 28 items assessed on a seven-point scale, with constructs for both intrinsic motivation and extrinsic motivation (Guay et al., 2003).

Participants were randomly assigned to interact with either the extroverted or introverted robot. After receiving informed consent, the experimenter introduced the participant to the Meka robot and explained the task. Participants were told that they would be com-

pleting the Tower of Hanoi puzzle under the supervision of the robot, and that the robot would provide all the necessary instructions for the rules and for progressing through the various stages of the puzzle. Participants were seated in front of the robot, facing it at eye-level, with a table between them. The physical Tower of Hanoi puzzle was placed on the table between the robot and participant (as illustrated in Figure 6.2). The participant was clearly instructed that it was their decision as to when they wanted to terminate the interaction and that they could indicate this to the robot at any time they wished. A headset microphone was used for capturing the speech of the participants, while the robot's speech was projected through its own speakers.

The experimenter then started the system implementation and left the participant to interact with the autonomous robot. After initial introductions, the robot carefully explained the goal and rules of the puzzle and asked the participant to complete it. After the participant's first successful completion, the robot explained that it would be asking the participant to complete the same puzzle several times and reminded the participant that it was up to them to decide when they would like to stop. During the execution of each puzzle task—the *in-task phase*—the robot monitored the task and provided help if the participant got stuck or explicitly asked for help. Following the successful completion of each puzzle—the *between-task phase*—the robot first provided positive feedback and then asked the participant if they would like to continue. If the participant agreed, the robot indicated a new puzzle goal and asked the participant to begin another iteration of the task.

Three levels of difficulty were implemented for the puzzle. The easiest difficulty required the solution of the three-disk version of the puzzle; the medium difficulty used four disks; and the hardest difficulty involved five disks. The least number of disk movements that can be made for each level of the puzzle are 7, 15, and 31, respectively. When the participant completed the puzzle eight times at the same difficulty level (starting with the easiest), the robot asked them to increase the difficulty by adding another disk. If a participant reached the hardest difficulty level, they stayed at this difficulty until they decided to terminate the interaction.

When the participant indicated that they wished to terminate the interaction, the robot thanked and instructed them to fill out a follow-up questionnaire at a computer nearby. Once participants finished this final questionnaire, the experimenter returned and thanked them once again for their participation.

Measures

The study included three objective measures of participant compliance: total time spent working on the task, total number of puzzles solved, and total number of disks moved across all instances of the puzzle. The follow-up questionnaire contained several seven-point rating scales for assessing the performance of the robot. Five items from this questionnaire were combined into a single construct of *perceived robot performance*, including questions about the robot's skills as an instructor and motivator, as well as questions relating to the usefulness of the robot's information and advice. This construct was found to have good internal reliability (Cronbach's $\alpha = .75$).

The follow-up questionnaire also included open-ended questions about why the participant chose to participate for as long as they did and why they chose to eventually terminate the interaction. In order to obtain a more task-specific measure of intrinsic motivation, these open-ended responses were coded for explicit mention of the participant's inherent desire to solve the puzzles, without mention of any external motivation coming from the robot.

Results

Analysis of the data was conducted using a between-subjects analysis of variance (ANOVA). Participant personality, robot personality, and the interaction of both were modeled as fixed effects. Participant gender, language (English or French), and previous experience with robots (yes or no), were found to be non-significant covariates on all measures and are not discussed further. Participant background (technical or non-technical) was found to have a significant effect on some measures and it has been retained as a covariate in the statistical model. A Bonferroni correction was employed to control for the experiment-wise error in multiple comparisons.

Compliance

Regarding the measure of total participation time, there was no main effect of either participant personality, $F(1, 35) = 0.75, p = .39$, or robot personality, $F(1, 35) = 0.16, p = .69$. A significant interaction effect was observed, $F(1, 35) = 14.80, p < .001$, with a significant effect of extroverted participants participating longer with the extroverted robot, $F(1, 35) = 5.97, p = .039$, and for introverted participants participating longer with the introverted robot, $F(1, 35) = 8.97, p = .010$. Participant background also had a significant main effect on participation time, with non-technical participants participating for longer than those from technical backgrounds, $F(1, 35) = 7.31, p = .011$.

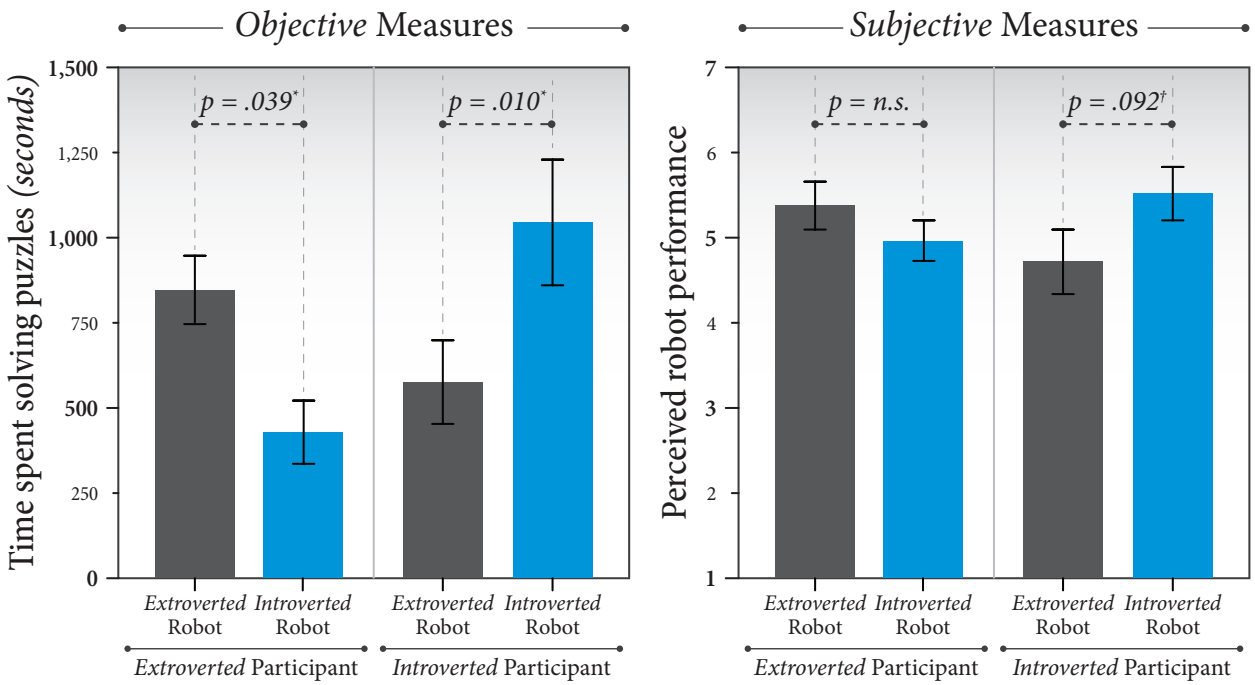

Figure 6.4: **Left:** Objective results of compliance for total participation time. A personality matching effect predicted by similarity-attraction theory was found. **Right:** Subjective results of perceived robot performance. Introverted participants reported a marginal preference for the introverted robot. (*) and (†) denote $p < .05$ and $p < .10$, respectively.

On the total puzzles solved measure, there was no main effect of either participant personality, $F(1, 35) = 0.36, p = .55$, or robot personality, $F(1, 35) = 0.57, p = .45$. A significant interaction effect was found, $F(1, 35) = 8.07, p = .007$, with a significant effect of extroverted participants solving more puzzles with the extroverted robot, $F(1, 35) = 6.51, p = .031$, but no significant effect among introverted participants $F(1, 35) = 2.16, p = .30$. Participant background did not have a significant effect on this measure, $F(1, 35) = 0.51, p = .48$.

On the measure of total disks moved across the entire interaction, there was no main effect of either participant personality, $F(1, 35) = 0.14, p = .71$, or robot personality, $F(1, 35) = 1.72, p = .20$. The analysis found a significant interaction effect, $F(1, 35) = 5.42, p = .026$, with a significant effect of extroverted participants moving more disks with the extroverted robot, $F(1, 35) = 6.65, p = .028$, but no significant effect among introverted participants $F(1, 35) = 0.52, p = .94$. Participant background was not found to have a significant effect on this measure, $F(1, 35) = 1.03, p = .32$.

Perceived Robot Performance

On the subjective rating of the robot's performance, there was no main effect of either participant personality, $F(1, 35) = 0.25, p = .62$, or robot personality, $F(1, 35) = 0.58, p = .45$. A significant interaction effect was found, $F(1, 35) = 4.70, p = .037$, with no significant effect among extroverted participants, $F(1, 35) = 0.99, p = .66$, and a marginal preference among introverted participants for the introverted robot, $F(1, 35) = 4.27, p = .092$. Participants with a non-technical background also expressed marginally higher ratings of the robot's performance, $F(1, 35) = 3.10, p = .087$. Results for compliance and perceived robot performance are visually presented in Figure 6.4.

Motivation

A regression analysis was conducted on the effect of the participant's reported global intrinsic motivation (collected in the pre-test survey) on all measures of compliance. No significant results were found for any of the measures. However, significant effects were observed for the task-specific measure of intrinsic motivation, in which participant responses were coded to the open-ended question: "Why did you participate for as long as you did?" Participants who indicated an inherent interest in solving the puzzle, rather than participating because of the presence of the robot and its feedback, were labeled as having high intrinsic motivation. Twenty-seven participants (14 extroverts, 13 introverts) were found to have high intrinsic motivation. A three-way interaction effect was observed between this measure of intrinsic motivation, participant personality, and robot personality on the compliance measure of total participation time, $F(1, 34) = 5.62, p = .006$. Among participants with high intrinsic motivation, there was no significant effect of personality among extroverted participants, $F(1, 34) = 1.43, p = .24$, or introverted participants, $F(1, 34) = 2.08, p = .16$. However, among participants without high intrinsic motivation, extroverted participants participated for significantly longer with the extroverted robot, $F(1, 34) = 15.84, p < .001$ and introverted participants participated for significantly longer with the introverted robot, $F(1, 34) = 19.44, p < .001$. The results involving intrinsic motivation are visualized in Figure 6.5.

Discussion

The goal of the experimental evaluation was to test the effectiveness of matching a socially assistive robot's personality—as expressed via the adaptive designed gaze behaviors—to that of the participant in a repetitive task requiring persistent motivation from the robot. The robot's personality was manipulated purely through its gaze behavior, gazing much

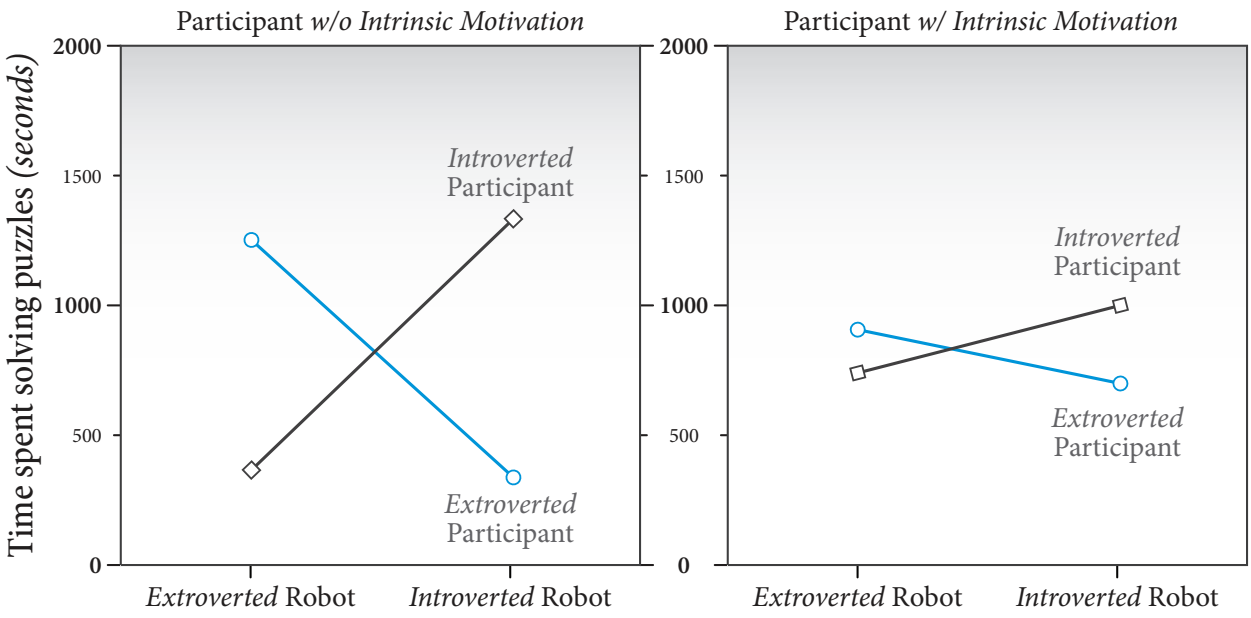

Figure 6.5: The interaction effect on compliance predicted by similarity-attraction theory is only present for participants that were not found to have high intrinsic motivation for the task.

more toward the participant when expressing an extroverted personality and much more toward the task space when expressing an introverted personality. These behaviors were validated in an online study to express their intended personality.

The first hypothesis predicted that, in line with similarity-attraction theory, participants would give higher subjective ratings to the performance of a robot that matches their personality. This hypothesis was partially supported in the experiment, in that introverted participants reported a marginal preference for the introverted robot behaviors. Extroverted participants reported no difference in ratings. Introverts may have been more consciously sensitive to the behaviors of the robot, as previous work has shown introverts have a superior detection rate and perceptual sensitivity than extroverts (Davies et al., 2013). In a study involving the rating of other people, previous work has also found that introverts preferred other introverts on the measures of "reliable friend" and "honest and ethical," while extroverts were ambivalent in these measures (Hendrick and Brown, 1971).

The experiment provided support for the second hypothesis, which predicted that participants would comply more with robots that matched their personality. On the measure of total participation time, both extroverts and introverts exhibited significantly greater compliance with the personality-matching robot. However, in the measures of total puzzles solved and total disks moved, only extroverts exhibited significantly greater compliance with the personality-matching robot. Previous work in HCI has shown a similar

result in that matching a synthesized voice's personality to a user's personality improved feelings of social presence, but only for extroverts (Lee and Nass, 2003).

The third hypothesis predicted that the intrinsic motivation of participants and the personality-matching effect would interact. Some support was found for this hypothesis, but not from the pre-test survey asking participants to rate their intrinsic motivation towards tasks in general. Instead, after coding participants' open-ended responses to a post-study interview question asking why they participated for as long as they did, it was found that some participants were more intrinsically motivated to solve the puzzles, whereas others were more extrinsically motivated by the robot. After splitting the participants into these two groups, a significant interaction was found between intrinsic motivation and personality matching. For both extroverts and introverts, the personality-matching robot was most effective in motivating those who were not highly intrinsically motivated to solve the puzzles. It is important to note, however, that splitting the population into groups with intrinsic and extrinsic motivation resulted in relatively small sample sizes and that follow-up work with a larger participant population must be carried out to more conclusively establish this relationship.

This work further demonstrates the importance of designing social technologies that can adapt to user characteristics. It has shown that a robot that matches the personality expressed by its gaze behavior to the personality of its user can improve compliance and subjective perceptions of the robot's performance. These outcomes are particularly critical for socially assistive robots that must motivate their users to engage in physical or cognitive exercises, especially when these exercises are repetitive or boring. Social theories such as similarity-attraction must be leveraged and further studied in human-robot interaction, as they can have powerful implications for the effectiveness of these interactions.

6.4 Chapter Summary

This chapter focused on the social variables of *personality* and *intrinsic motivation* for the design of adaptive gaze mechanisms for agents. Specifically, it presented the design and evaluation of gaze behaviors for socially assistive robots that allow the robot to match the personality of the user, thereby more effectively motivating users to repeatedly engage in a therapeutic task. Socially assistive robots are envisioned to provide social and cognitive assistance where they will seek to motivate and engage people in therapeutic activities. Due to their physicality, robots serve as a powerful technology for motivating people.

This chapter presented work on matching a robot's personality—expressed via its gaze behavior—to that of its users. It focused on the extroversion dimension of the Big

Five personality model (John and Srivastava, 1999) as it is the most accurately observable and expressible dimension of personality expressed by nonverbal behaviors over short timescales (Lippa and Dietz, 2000). An online study with 22 participants confirmed that the robot's gaze behavior can successfully express either an extroverted or introverted personality. A laboratory study with 40 participants demonstrated the positive effect of personality matching on a user's motivation to engage in a repetitive task. It also demonstrated the importance of taking the user's intrinsic motivation into account when attempting to produce external motivation and increase compliance.

7 GENERAL DISCUSSION

The primary goal of this dissertation is to present algorithms and techniques that model situated gaze mechanisms for embodied agents, focusing particularly on contingencies with various user behaviors, user characteristics, and task states. Gaze shifting was the most fundamental mechanism; the three other mechanisms—aversions, coordination, and adaptivity—all utilized the gaze shift model as a core component in generating gaze motions.

All of these mechanisms have been demonstrated to positively affect an array of high-level social and cognitive processes in human users. Furthermore, this dissertation presented several system implementations that generate gaze behavior based on these models, evaluated the implementations in specific task contexts, and demonstrated the challenges and tradeoffs in designing behavioral mechanisms across virtual and physical agent presentations. A primary contribution of this dissertation is new knowledge on how low-level gaze variables might be controlled and manipulated to achieve high-level social and cognitive effects, connecting existing knowledge in the social sciences to a more computational foundation.

Each of the mechanisms presented in this dissertation was comprised of several different low-level variables. Some of the most important variables across the three higher-level mechanisms—aversions, coordination, and adaptivity—were those that related to *timing*. For example, gaze aversions were found to be most effective in conveying positive subjective impressions and enabling an agent to better handle turn-taking only when the aversions were deployed with humanlike timings (in relation to speech utterances). Similarly, agents that timed their gaze motions reactively to the user's gaze motions were better at collaborating than agents that gazed unreactively. And finally, the timing and distribution of gazes toward a task space and toward a user make the difference in conveying introversion vs extroversion, which is critical in motivating users in a rehabilitation scenario.

Although timing was the most studied variable for evaluating the mechanisms in this dissertation, many other variables are present in the models and could warrant their own evaluations. For example, the importance of the spatial dimension was not evaluated for any of the mechanisms, with most experimental conditions controlling for that factor, e.g., directing gaze aversions according to the distribution of aversion directions found in the human-human data collection for all conditions. Thus, open questions still remain as to the importance of spatial variables and other unstudied variables for achieving maximum effectiveness for each gaze mechanism. Was it important to follow the distributions from human-human data in terms of aversion direction? What if the agent had directed all of

its gaze aversions to the side, or only downward? What would this decidedly non-human pattern of gaze aversion direction mean for the interaction? Similar questions could be raised for each mechanism.

The remainder of this discussion addresses related issues and questions. First will come a discussion of the methodological validity of the research presented in this dissertation, followed by discussion of the issue of generalizability and applicability for extending these mechanisms to new users, scenarios, environments, and so on. Then this chapter will discuss the specific and inherent technical challenges involved in this work and finally propose open questions and future research directions on the topic of social gaze modeling for interactive virtual agents and robots.

7.1 Methodological Validity

In order to design each gaze mechanism, it was important to understand the social system being replicated as it occurs in humans. The *understanding* phase thus focused on human-human data collection, providing a computational understanding of the mechanism in question. This understanding enabled the design and *building* of models and techniques, integrated into a sociotechnical system for agents. Implementation of the sociotechnical system was followed by *evaluating* whether or not it was a good one, measuring interaction outcomes in a similar context to where the initial social system was studied.

The chosen methodology reflects an explicit choice to develop human-centered models of gaze constructed from theory and direct observations of human behavior. Social interaction, as with any other complex system, is made up of many interrelated components and mechanisms that interact with each other. Therefore, I argue that designing artificial agents to work within a social interaction system needs to be grounded in a deeper understanding of these components and mechanisms and the relationships among them. However, this choice raises the following question: Is human data-driven modeling the best approach to designing social gaze behaviors, and are the mechanisms presented here the best representations for the modeled behaviors?

I readily acknowledge that human-based modeling is not the only way of capturing the variables and mechanisms of social gaze behavior. Other approaches might primarily rely on a designer's intuition and guidelines developed through an iterative process, e.g., iterative or formative design. Furthermore, different methodologies might be more appropriate for different agent designs and fidelities. For instance, while a human-based approach might create gaze mechanisms that work for embodied agents that are moderately to highly humanlike in appearance, other approaches might create behaviors that are

more appropriate and effective for agents with abstract designs. Very little previous work has addressed this specific issue, except for Pejsa et al. (2013) who have demonstrated a set of both existing and novel animation techniques that can improve the quality of gaze motions for highly abstract and stylized virtual agents. These techniques could not have been determined strictly from studying human interactions.

This dissertation does not answer the question of whether or not the chosen methodology is the best possible one, which would require utilizing different methodologies toward achieving identical mechanisms and comparing their effectiveness. However, I would again argue that the complex nature of social interaction requires the design of artificial gaze mechanisms to be grounded in a deeper understanding of the variables identified in existing social science theories. However, even within the chosen methodology, a number of scoping decisions had to be made in order to make the research feasible. The potential limitations for methodological validity caused by these decisions are discussed next.

Choosing Variables to Explore

Social behavior is an infinitely complex space for design variables and the relationships among them. In modeling gaze mechanisms, I made design decisions to focus on variables that I found to be the most salient and important while omitting others. These decisions were generally made based on my intuitions and experience. While I sought ways to formalize some of these decisions by grounding them in social science theory or empirical results from open-ended pretesting, I could not formalize and validate all of them given the complexity of the interaction space and the limited time and resources of this dissertation research. At various stages in the methodology, seeking external validation on each decision could improve the overall validity of the process. However, because this process might involve hundreds or thousands of design decisions when working in a complex interaction space, intuition and experience will always inform to some degree what decisions and analyses should be validated.

As an example of modeling variables that were not considered, a limiting assumption of the gaze aversion model is that the gaze aversion behaviors generated are stable over time, while in reality these behaviors likely change over the course of a conversation due to increasing familiarity with the interlocutor, changing emotions and level of comfort, and so on. In fact, previous research has shown that the total amount that conversational partners look at each other often decreases over the course of a conversation (Abele, 1986). Similarly, gaze aversions during speech become more common in later conversations between counselors and clients (Schulman and Bickmore, 2012). Gaze aversions to regulate

intimacy while listening should also change in frequency and length based on the intimacy of the topic at hand and the familiarity of the conversational participants.

As modeled in this dissertation, the gaze aversion mechanism is also only responsive to the speech of the human user but not to their gaze. By tracking the user's gaze, as was later done for the gaze coordination mechanisms, the agent might be able to more effectively regulate mutual gaze throughout the conversation and recognize gaze aversions of its user. Interactively aligning gaze aversions is particularly important to ensure that the agent's gaze aversions are recognized by the user, for instance, by producing a turn-taking gaze aversion only when the user is looking toward the robot.

The design of the gaze aversion mechanisms did not consider the above variables as a part of the design as a means to simplify the design space. However, whether including these variables in the design might have changed the measured social and cognitive outcomes remains unknown. A natural next step in future work would be to develop models of gaze that, e.g., consider interaction history and dynamically adjust gaze aversion strategies over time.

On the evaluation side, this dissertation cannot provide strong claims about the relationship between specific model parameters and individual or collective interaction outcomes. Most of the current evaluations compare the entire model against not using the model at all, or using a reduced or simplified version of the model. However, all of the models presented in this dissertation have dozens of parameters that could each be individually explored to test the sensitivities of their effectiveness. For example, is it important that an agent uses a cognitive gaze aversion one second before speaking, rather than half a second? Multivariate evaluation has recently been proposed in the HRI literature as a means to start answering questions such as these (Huang and Mutlu, 2014). Multivariate evaluation builds multiple regression models to capture the relationship between a set of design variables and an interaction outcome as a linear system in order to determine which design variables predict the interaction outcome and the extent to which each variable affects it. Future work in this space should consider combining conventional evaluation methods with multivariate evaluation to create a more holistic understanding of the entire design space.

Granularity

Within the complex design space of social gaze interaction, I also made a number of decisions pertaining to the level of detail in modeling gaze mechanisms. For example, although the gaze aversion mechanisms presented in this dissertation were closely tied

to the conversational states of speaking and listening, gaze aversions are also affected by the content and structure of speech. Previous research by Cassell et al. (1999b) identified relationships between gaze behavior and the information structure of utterances, specifically the theme and rheme of sentences. Future work should integrate knowledge such as this with the gaze aversion controller presented here. Similarly, the gaze adaptivity work differentiated between on-task and between-task times, but future work will need to include finer-grained task analyses.

Future work on gaze coordination should also further investigate the temporal aspects of the gaze behaviors observed in reference-action sequences. The current work divides a reference-action sequence into an ordered sequence of four to five phases, but the gaze fixations within these phases are aggregated, and the low-level ordering of fixations is lost. While scanpath analysis is commonly used for analyzing temporal characteristics of gaze, scanpaths that result from this analysis only represent the gaze behaviors of individuals. The analysis presented in Chapter 5 attempted to extract generalizable patterns of gaze behavior by aggregating data across multiple dyads and abstracting away the variability in gaze that results from individual differences and changing contextual factors. Ideally, the data collected for any gaze or other behavioral mechanisms should be modeled at multiple levels of information structure, with intermediate testing for how well each level predicts positive outcomes in order to automatically choose the right level of granularity for implementation.

Separating Gaze from Other Cues

An important methodological limitation to consider is that across all of this work, gaze was singled out from the full set of nonverbal cues that humans make use of in interaction. In everyday human communication, other behavioral mechanisms such as facial expressions, arm and head gestures, and posture are utilized to form a full range of visible behavior alongside spoken utterances. For instance, previous research suggests a strong interaction between gaze and interpersonal distance (Argyle and Dean, 1965), and future gaze models should consider the proxemic aspect of the social situation in generating gaze behaviors. Gestures also play an important role in conversations, supporting speech content and communicating information that cannot be efficiently conveyed through speech alone (Chawla and Krauss, 1994; Cassell et al., 2007; Becvar et al., 2008).

A number of separate research efforts have investigated different social cues for agents and robots, including gaze, speech, gesture, proxemics, and so on (Huang and Mutlu, 2014; Mumm and Mutlu, 2011b). These research efforts have made great strides in understanding

how to design behavioral cues for agents and what effect they might have in various contexts. For instance, body orientation has been shown to play an important role in communicating an agent's direction of attention, in addition to the head and eye movements that make up overall gaze behavior (Pejsa et al., 2015).

Furthermore, when agents have a humanlike appearance, subtle idle behaviors such as breathing, blinking, and fidgeting might be required to create an impression of lifelikeness. A mismatch in lifelikeness between an agent's appearance and behavior will likely weaken it effectiveness. For example, an agent that appears to have the ability to produce facial expressions will raise expectations of doing so; it will need to actually produce facial expressions in order to meet these expectations. In most of the implementations presented in this dissertation, lifelike behaviors such as blinking, subtle fidgeting, periodic smiling, and eyebrow raising were designed into the agent's behaviors, but these "extra" motions were held constant across all evaluation conditions in order to focus on gaze and control for experimental variability. Deeper consideration of whether user expectations are met in terms of the agent's overall behaviors would improve the validity of social and cognitive outcome measures in human-agent interactions.

How can we design and implement the full range of situated social cues for embodied agents? Future research will need to develop higher-level, more holistic models that can tie social cues together, along with studies that can measure how the cues interact with each other. In general, the interactions afforded by embodied technologies are powerful but incredibly complex. In order to create these interactions, designers will likely need new methods and tools that can aid them in confronting such complexity.

7.2 Generalizability

The experimental evaluations of the gaze mechanisms presented in this dissertation demonstrated a number of significant social and cognitive outcomes arising from manipulations in the underlying models. However, all of these results were obtained with specific populations, in specific social and task scenarios, and using specific agent platforms. Therefore, questions remain regarding the generalizability of these results. Do the results presented here extend into other user populations, tasks and interaction scenarios, agent platforms, or into real-world contexts? How could more generalizability be achieved? This section addresses these questions.

Scaling behavioral models, such as the situated gaze mechanisms presented in this dissertation, to different aspects of collaborative interaction, different roles, more complex tasks, more users, and more complex environments, is an open challenge that future work

will need to explore. Accomplishing such extensions will require additional data collection to design more robust behavior models, more sophisticated task models to support an array of agent roles, and more sophisticated strategies to accurately infer user states and intents in open-ended scenarios. Future work will need to combine and coordinate several such models, scalable in such a way that new models can be easily added or adapted to accommodate unforeseen situations. Such scalability will allow embodied agents to be able to produce and respond to gaze effectively in a wide array of interaction roles and contexts.

User Characteristics

Studies that compare the results presented in this dissertation across a wider variety of user groups would significantly improve their generalizability. For gaze adaptivity, a single dimension of individual differences was considered—extroversion and introversion—but there are of course many others. People might utilize gaze differently across gender, culture, language, other personality dimensions, age groups, and so on. Therefore, I cannot make strong claims that the mechanisms presented here would be equally effective for user groups and contexts different from what I studied. Future work should look at how designed behaviors could be extended to agents that work in different cultural contexts, use different languages, and interact with people with different demographic and personality attributes.

Research on gaze in human communication suggests that *gender* has a significant effect on both the production and perception of gaze cues (Argyle and Ingham, 1972; Abele, 1986; Bente et al., 1998; Bayliss et al., 2005). Gender was considered at various points in the modeling and analysis of the gaze mechanisms presented in this dissertation. For validating the model of gaze shifts, both a female and male virtual agent were developed and matched to participant gender. Participants rated gaze shifts performed by the female agent as significantly more natural than those performed by the male agent, but communicative accuracy of the gaze shifts performed by the male agent was significantly higher than that of the shifts performed by the female agent. However, these measured outcomes may be due to any one of a number of differences in the characters' designs not having to do with gender.

For modeling gaze aversions, an equal mix of male-male, female-female, and female-male dyads were recruited for human-human data collection in order to account for possible gender differences. For the gaze coordination and gaze adaptivity mechanisms, the gender of the participant was modeled as a covariate in the experimental data analyses, and in

both cases were found to be nonsignificant predictors on any of the outcome variables.

Gaze behavior is also sensitive to *culture* (Argyle and Cook, 1976). Designed gaze mechanisms and the social and cognitive outcomes that they lead to are limited to the cultural context and language of each particular study. Native English-speaking American participants were hired most often in this work to evaluate agent gaze mechanisms. However, the gaze adaptivity model was developed and evaluated in France, with a much wider set of cultural and language backgrounds. The study was conducted in both English and French, and participants were allowed to choose the language they were most comfortable with. This variable did not have an effect on the outcomes, which has promising—if still limited—implications for the cultural generalizability of the adaptive gaze model. Further experiments are required to understand whether the results from the experiment can generalize to other cultural contexts.

The physical and mental *abilities* of intended users are likewise very important to consider. The gaze adaptivity mechanism was primarily developed for socially assistive robots that might take on roles in elder care or stroke therapy. However, a significant limitation of the gaze adaptivity work is that the study population was comprised of healthy adults all under the age of 60. This choice was made for the purposes of feasibility in testing a novel idea for socially assistive robots, similar to what has been done in previous work in this domain (Tapus and Matarić, 2008). I expect the findings to hold for the targeted populations that need socially assistive robots, such as the elderly or post-stroke patients, but this expectation should be tested in future work.

Task and Interaction Context

The tasks devised for data collection and experimental evaluation also place some limitations on the generalizability of results. For instance, although data on gaze aversions was collected during human-human conversations about movie preferences, whether the results from those studies would generalize to different tasks or conversation topics is unknown. Conversation topic has been confirmed in the human communication literature to affect how much people look at each other in general (Abele, 1986), but it was not considered in the current work.

One task instance of collaboration was also explored for the gaze coordination model: assembling toy sandwiches. This task was designed such that it involves generic aspects of collaboration that should map well to other tasks, such as verbal referencing, physical actions, monitoring behaviors, and so on. The gaze coordination model handles an important subset of collaborative interactions, in which embodied agents will serve in instructional

roles and need to produce verbal references that users are intended to act upon. Many complex real-world tasks comprise the same sorts of reference-action sequences that were studied here, thus, I expect the gaze coordination model to apply to a wide range of activities that interactive embodied characters might engage in with human users. However, future research should support this by implementing the gaze coordination model in other task contexts to empirically assess its generalizability.

Within task context is the issue of roles. Another limitation of the gaze coordination model presented in this dissertation is that it only handles the instructor role. The worker role, and roles in other types of collaborations, should be considered in future work. For example, the agent might need to accomplish actions based on user instructions. Recent work has started to address this need for robots, exploring how a collaborative robot in the worker role might react to user gaze in order to increase efficiency and naturalness in the interaction (Huang and Mutlu, 2016).

Future work should also explore qualitatively different interaction contexts, such as competitive interactions in which the user and agent do not share the same goals. Additionally, handling multiparty interaction scenarios in which the agent must distribute its gaze to more than one user throughout a task would require significant extensions to the mechanisms presented here.

Agent Design

An inherent limitation of research on embodied agents is imposed by the chosen visual and physical designs of the agent platforms themselves. While the gaze mechanisms in this dissertation have been applied and evaluated on several different virtual agents and robots, whether the positive results found would generalize to interactions with other agents is unknown. Previous research has shown that people's perceptions of an agent's characteristics can affect their responses to and subjective impressions of the agent. For example, Powers and Kiesler (2006) showed that a robot's physical characteristics, such as whether it had a male or female voice, the fundamental frequency of its voice, and the length of its chin predicted participants' rating of how knowledgeable and sociable they found the robot and whether they would follow its health advice.

Future work needs to test the generalizability of the mechanisms presented here to interactions with other agents in order to gain a better understanding through systematic studies of how different agent characteristics might shape people's perceptions of and responses to the agent. Furthermore, new methods and models may need to be developed in order to adapt gaze and other behavioral mechanisms to vastly different agent

morphologies, e.g., agents with highly stylized or abstract designs (Pejsa et al., 2013).

Short Interactions in the Lab

All evaluations presented in this dissertation were conducted in a controlled laboratory setting wherein participants interacted with one or more agents for at most an hour, often for only 15 minutes or so. Thus, questions of generalizability arise in terms of utilizing these gaze mechanisms in real-world contexts and for long-term interactions. Whether the social and cognitive outcomes demonstrated in this work could be obtained in less controlled environments is unknown. For instance, would an agent's use of referential gaze shifting (Chapter 3) lead to greater information recall in a real-world classroom over longer periods of interaction?

To answer questions like this, future work would need to situate designed gaze behaviors in real-world scenarios throughout the methodology. Real-world contexts should be analyzed as they currently exist, more robust and adaptive models should be developed to reflect understanding of the context in a flexible manner, and evaluations should be conducted by deploying agent systems "in the wild" to determine their effectiveness with real users outside of the lab setting. For example, testing whether an agent could use gaze cues to adaptively express personality in something like a public information kiosk with individuals who are not paid to interact with the agent would provide important insights into the generalizability of that mechanism.

Modeling Simplifications

For feasibility purposes, modeling for each of the gaze mechanisms required making a number of simplifications. For gaze adaptivity, the participant population was median-split for both the modeling and the evaluation studies to establish "introverted" and "extroverted" groups in order to investigate the effect of matching/mismatching personality categories with balanced sampling. To address the potential limitations of this simplification, future work should model extroversion as a continuum, and the embodied agent should dynamically adjust its behaviors to match the user's location on this continuous spectrum.

Other personality dimensions should also be modeled. Previous work has shown that neuroticism and openness have consistent correlations with a person's gaze behavior, as measured by fixation frequencies and durations (Rauthmann et al., 2012). However, previous research has also demonstrated that personality and nonverbal behavior are not always linked in simple ways (Gifford, 2006). Personality can be expressed differently in different contexts, group compositions, cultures, and combinations, and this richness

should be taken into account in future gaze models to align agent gaze with intended personalities. Additionally, the gaze adaptivity mechanisms should be extended in future work to detect user characteristics such as personality *in situ*, rather than requiring the user to explicitly provide this information, e.g., by filling out a questionnaire before using the system as was done in this work.

Simplifications were similarly made in the gaze coordination model, which focuses solely on agents making verbal references to task objects in the form of reference-action sequences. These sequences are a fundamental building block of collaborative interactions, but other elements will need to be addressed to extend beyond such discrete and well-defined sequences, such as a more general model of turn-taking, coordinating gaze during non-referential speech, providing verbal and nonverbal feedback to users, and so on. For example, the gaze coordination model does not tell an agent where to gaze when making small talk, or when giving a monologue, or when interacting with more than one user.

A general question also remains in terms of how to combine the gaze mechanisms presented here, along with behavioral mechanisms developed by other researchers, into a single cohesive model for embodied agents to use. A meta-model could be developed to structure various studied mechanisms in a cognitively and computationally plausible manner. This approach requires new evidence on how the different gaze mechanisms are related and work together as well as a cognitive architecture that can plan and accommodate the organization of the different mechanisms.

This cognitive architecture could plan out gaze motions alongside other verbal and nonverbal behaviors based on higher level goals. For example, an embedded gaze aversion model might indicate that a gaze aversion (Chapter 4) needs to happen soon, while the coordination model (Chapter 5) is simultaneously indicating that the agent should be gazing toward a certain task object. The cognitive architecture might then include a planning algorithm that combines those requests into a single gaze shift, aligning the agent's head (if applicable) in accordance with the agent's goals to be more affiliative or referential (Chapter 3), and maintaining gaze toward that target for the amount of time that would reinforce its intended personality expression (Chapter 6).

Applicability

Relevant to all of the above generalizability issues is the related question of applicability. For each of the mechanisms presented in this work, there is a need to discuss the extent to which the model is applicable to scenarios and contexts outside of what was specifically evaluated. For example, the gaze coordination model (Chapter 5) was derived and evaluated within a

single scenario—training a user on how to construct specific sandwiches out of toy materials. However, this model task was specifically developed to represent a broader class. Similarly, the gaze aversion (Chapter 4) and gaze adaptivity (Chapter 6) models were designed and evaluated within structured dyadic conversations and physical puzzle-solving scenarios respectively. The gaze shift model (Chapter 3) is a lower-level mechanism that is broadly applicable (in fact, it was utilized to drive gaze shifts in all implementations and scenarios in the remainder of the dissertation).

For the mechanisms of gaze aversion, gaze coordination, and gaze adaptivity, there are thus three categories of scenarios to consider: (1) those that the model would apply directly to, (2) those that would require some degree of modification to the model, and (3) those that are fully outside the model's scope.

Direct Application

The gaze coordination model has direct application to instructional object-based tasks (physical or VR) in which an agent is training a user on how to act on one or more objects (building sandwiches, preparing recipes, assembling furniture, fixing a bicycle, arranging table settings, etc). These tasks all contain the essential elements addressed by the model: making mutual gaze, shifting and responding to joint attention cues, tracking action intent, etc. Underlying implementations would differ in terms of dialogue handling and object tracking, but the model would still apply. An open question for future work concerns the sensitivity of the precise timing parameters collected from the sandwich-building task.

Similarly, the gaze aversion model can be directly applied to structured conversations with primarily question-answer dialogue turns. The gaze adaptivity model is directly applicable to scenarios in which an agent teaches and motivates known introverted or extroverted participants in a task that can be segmented into in-task (the user is engaged in the task) and between-task (the agent provides instruction and motivation) phases.

Requires Modification

The gaze coordination model would require some modification for fully collaborative physical tasks in which the agent is not just instructing, but receiving instructions from the user and taking actions of its own. The model would need to be extended to account for agent gaze behaviors in more fluid and open-ended roles. Other tasks in this category include those that do not involve taking actions on objects, such as an agent tour guide. Parts of the model are still applicable (mutual gaze, gazing to referents, following joint

attention, etc) but others would need to be adapted, e.g., there would no longer be an "action" phase.

The gaze aversion model would require some augmentation in order to be employed in more open-ended conversations—these scenarios would also require significantly more advanced dialogue modeling—as well as multi-party conversations. The model would need to be integrated with other models, e.g., gaze coordination, in more situated contexts with task-relevant objects and other situational attractors for the agent's gaze. The gaze adaptivity model would need to be modified in order to handle other personality dimensions, more open-ended interactions, tasks with completely different structure, tasks with more than one object to gaze at, and multi-party interaction.

Outside the Scope

Outside the scope of what the gaze coordination model could handle are casual conversations not situated in an environment of relevant objects, or traditional tutoring scenarios where the agent is conveying abstract information and checking for understanding. There are a number of existing models that have been developed in the virtual agent and human-robot interaction literature that could be applied in these situations, including the gaze aversion model presented in Chapter 4. The gaze adaptivity model is not applicable in scenarios where user personality is completely unknown or where motivation is not a primary goal.

7.3 Technical Challenges & Limitations

In designing situated gaze mechanisms utilizing the presented methodology, I faced a number of technical challenges that pose significant bottlenecks in advancing the state-of-the-art for human-agent interaction. Limitations arise in capturing human behaviors, accurately modeling those behaviors, and developing practical system implementations that can not only express the specific behavior but also handle all sensing and logic required by the intended task context. In general, applying new techniques from other areas such as natural language processing and synthesis, discourse and dialog modeling, computer vision, and machine learning to these problems might significantly help future work on designing situated social behaviors to overcome some of the technical challenges.

Behavioral Coding

Development of the gaze aversion, gaze coordination, and gaze adaptivity mechanisms required a large amount of manual behavioral coding on collected human-human data. For gaze aversions, videos were coded for aversion timing, direction, and purpose. For gaze coordination, gaze fixations were coded for the semantic target and all speech was transcribed. For gaze adaptivity, videos were coded for task-directed and person-directed gazes. All of these projects required the use of human coders, which puts significant limitations on the amount of data that can be coded, the coding categories, and how biases and error can be introduced during the modeling process. Future work should look to automating this process by utilizing computer vision techniques along with estimations for filling in missing data and correcting errors through semi-supervised machine learning techniques.

Practical Implementation Limitations

Each of the gaze mechanisms was situated in a particular task implementation that was designed with sufficient fidelity for the purposes of testing the associated gaze mechanism, but nothing more. For example, a limitation of the gaze coordination mechanism as presented in this work is that the simple sandwich-building task it was embedded in has a very limited range of instructional utterances and worker actions that are possible. Users can only inquire about a very limited set of objects (the ingredients present on the table) and the agent is only capable of tracking a very limited set of physical actions (moving an ingredient to the bread). This implementation was chosen not only due to practical limitations in time and resources, but also because I was simply not researching ways to develop the best possible collaborative sandwich-making system. Instead, I needed to build a system with sufficient interactional fidelity such that I could test specific hypotheses related to the gaze coordination mechanism. I believe I succeeded in that goal, but more external validating across other tasks (see the related point on generalizability above) would be required to sufficiently prove that claim.

Modeling Limitations

A significant limitation across all of these mechanisms pertains to the underlying model representations. All models generally took the form of hybrid heuristic/stochastic models, often including probabilistic state machines to computationally operate each gaze mechanism. While these representations were sufficient to model the amount of human

behavioral data collected for the mechanisms considered in this work, future work should look into finding more scalable ways to represent large amounts of sequential data using techniques such as Hierarchical Hidden Markov Models or Conditional Random Fields. The modeling employed in this dissertation also potentially misses significant edge cases of behavior due to the statistical approach of collapsing data into averaged distributions.

Building more sophisticated computational representations of situated gaze mechanisms will require processing and learning from larger amounts of data from wider contexts. Machine learning techniques will facilitate the process of finding behavioral patterns across many interactions, represent these patterns in temporal probabilistic frameworks, and enable flexible real-time generation of interactive behaviors. These techniques can also provide estimations for missing data and errors that occur in sensing and inference. Using these technologies will further facilitate the studying of complex interaction processes, for instance, taking into account multiple participants and shifting gaze targets in understanding multi-party behaviors in dynamic scenes.

7.4 Open Questions & Future Work

In addition to future work based on the above methodological and technical limitations, a number of interesting open questions can be posed for designing more effective situated gaze and other behavioral mechanisms for socially interactive systems. This section presents open questions and ideas on how future work can move beyond human-based modeling, integrate embodied agents into everyday life, analyze behavioral mechanisms at micro-scale, utilize agent gaze mechanisms in simulated social interaction as research tools for psychologists, and develop more general and automatic strategies for considering agent affordances to maximize their interaction effectiveness.

Moving Beyond Human-Based Models

A growing body of literature has started to suggest that agent and robot gaze is not necessarily afforded the same special cognitive status as human gaze (Meltzoff et al., 2010; Admoni et al., 2011; Okumura et al., 2013), providing evidence that non-human-based models might be worth exploring for designing artificial gaze mechanisms. As mentioned in Chapter 2, people generally exhibit a reflexive cueing effect in which they unavoidably shift their attention in the direction of another person's averted eye gaze. This effect suggests that gaze is processed in a separate neural pathway from other directional symbols, such as arrows. A test of the reflexive cueing effect utilizing both highly anthropomorphic and

highly stylized robots has shown that robots failed to elicit reflexive cueing in people, suggesting that robots may be cognitively processed more like arrows than like human faces (Admoni et al., 2011).

Infants have also recently been shown to disregard robot gaze while treating human gaze as meaningful. When infants are shown videos of robots and humans looking at objects, they can follow the robot gaze as well as the human gaze. However, the infants look longer at, and show a preference for, objects gazed at by the human over objects gazed at by the robot (Okumura et al., 2013). Only after infants observe a robot engage in a socially communicative exchange with an adult do they follow a robot's directional gaze (Meltzoff et al., 2010). This suggest that, even to infants, robot gaze is not automatically processed as being meaningful in the same way as for human gaze. Future work should further explore this issue, fully mapping out the differences in how people perceive agent gaze shifts vs human gaze shifts, and what those differences might mean for designing potentially non-humanlike gaze motions that work more effectively for virtual agents and robots.

Integrating Agents into Everyday Life

How can we integrate embodied agents into everyday life as educational tutors, peers for play, work collaborators, therapy aids, etc? The research in this dissertation involved manually exploring large amounts of data to identify variables or patterns of interest, followed by implementing these variables and patterns into hybrid stochastic/heuristic models. While this methodology has produced a number of models that sufficiently capture desired mechanisms and effectively facilitate targeted human-agent interacts, the models are often not robust or flexible enough to immediately extend into real-world deployments. Thus, future research in this space must move toward machine-learned models (e.g., probabilistic graphical models) derived directly from the data on human behaviors, while still sufficiently exploring the source data to ensure understanding of the patterns and behaviors being generated. To complement these efforts, future work should include longitudinal, *in situ* studies of people's interactions with socially interactive systems in realistic settings.

Analyzing Gaze Mechanisms at Finer Scale

Very little previous work has explored humans' micro-scale responses to embodied agent gaze. Future work could take a conversation analytic approach, plotting sequences of human-robot gaze (possibly along with speech) and comparing them to human-human

sequences of interaction. These comparisons could illuminate differences at a finer granularity than the high-level investigations in this dissertation wherein different models were compared against each other to determine their overall effect on high-level outcomes like rapport and task efficiency. Previous work has demonstrated that people's fine-grained responses to robot gaze differ from their responses to human gaze in terms of reflexive gaze cueing (Admoni et al., 2011). In the moments just before naming an object, people spend more time ensuring joint attention by looking at their partner's face than at the object if their partner is a robot, but more time looking at the object than at their partner's face if their partner is another human (Yu et al., 2012). Findings such as these warrant further exploration in human-agent interaction. Are these fine-scale differences due to a novelty effect, or do they go away with further exposure to the agent? What other kinds of fine-scale patterns can be observed in the gaze behavior of humans that interact with agents utilizing various artificial gaze mechanisms?

Agent Gaze Mechanisms as Psychological Research Tools

How can we use socially interactive systems as basic research tools to study aspects of human communication and cognition? Utilizing artificial embodied agents to produce these cues in a tightly controlled manner is a methodology referred to as *simulated social interaction*, and is an emerging experimental paradigm in psychology and cognitive science (Byom and Mutlu, 2013). This technique involves generating social behavior in artificial agents, such as virtual characters or robots embedded in realistic environments. Simulated social interaction allows for a great deal of experimental control and ecological validity. Participants interact with simulated agents whose behaviors are precisely controlled to reflect experimental manipulations and that respond to changes in the participants' behaviors, affording more realistic interactions than static stimuli can accomplish. Simulations of social stimuli follow computational representations of human behavior, which provide the experimenter with control parameters for the behavior or mechanism under study and the ability to create experimental manipulations that are impossible or infeasible for human confederates to perform.

More General Consideration of Agent Affordances

Chapter 4 discussed the opportunities and inherent challenges in mapping the gaze aversion mechanism from a virtual agent platform to a physical robot platform. In general, mapping gaze mechanisms from virtual agents, which have nearly unlimited capabilities, to physical robots, which are constrained by hardware, is not trivial (Ruhland et al., 2015).

Understanding the specific effects of each capability and affordance will be necessary to allow researchers to avoid over-generalizing their findings from virtual agents to physical robot interactions.

The variability in appearance and capability of agent eyes is wide. Because studies on human-agent interaction are conducted with many different robots and agents, the results of one study may not directly transfer to different embodiments. Researchers will need to develop new methods for transferring the communicative intent of gaze movements, moving beyond the literal details. To achieve this transfer, researchers will need to encode example motions based on higher-level properties, such as implied direction of gaze, rather than the specific measurements of the orientation of the eyes and the head. They will also need to generalize existing control models to consider multimodal input and output variables at a higher-level than the specified variables. For example, rather than computing eye direction specifically, a model might generate an overall gaze direction and a retargeting process might specify how the gaze shift toward this direction should be executed given the affordances available to and the constraints of the specific agent representation. Finally, they will have to add characteristics of the agent representation as parameters to the models. For example, knowing that the character has big eyes might automatically result in dampened saccades to avoid jerky movements (Pejsa et al., 2013).

Robots with physical embodiments but animated eyes, such as Baxter[1], the IROMEC robot[2], and Chester (Vázquez et al., 2014), present an interesting test case for the effect of embodiment on gaze perception. These robots may help separate the effects of physical eyes, as opposed to physical bodies, when examining gaze in human-robot interactions. Essentially, these robots have two-dimensional animated eyes embedded in a three-dimensional physical head. Interesting research questions here include: How should head-eye coordination unfold for gaze shifts on these systems? How much does the Mona Lisa effect (Al Moubayed et al., 2012) influence perceptions of gaze motions? Recent work has begun testing the ability of users to perceive small and large gaze shifts of people represented on telepresence displays when their gaze motions are carried out only in the video image, only on the physical motion of the telepresence system, or using a combination of both (Kawaguchik et al., 2015).

Completely non-human embodiments also create challenges in designing social behavior mechanisms. The most important use of gaze is to express attention, but how can attention be expressed on very non-humanlike agents that do not have heads or eyes? How can a system express attention without any kind of embodiment whatsoever? These

[1]http://www.rethinkrobotics.com/baxter/
[2]http://www.iromec.org/

questions again touch on the general problem of limited affordances. Designers will need to focus on the desired outcomes, understand how people achieve those outcomes, then iterate on designs that target those outcomes with the affordances and capabilities available to the system.

8 CONCLUSION

Four mechanisms of social gaze for embodied agents were presented in this dissertation (Figure 8.1). The first step in designing social gaze into artificial systems was to determine the precise mechanics of *how* to carry out gaze motions. Chapter 3 presented the development and evaluation of the most basic gaze mechanism, *gaze shifts*, coordinating both head and eye movements to direct an agent's attention from one focal point to another (Figure 8.1A). A model of gaze shifts was presented that was validated to achieve human-like gaze for virtual agents. The validation study demonstrated that this physiologically inspired model provides parametric control and achieves both communicative accuracy and subjective naturalness.

A follow-up evaluation of the gaze shift model demonstrated that it can achieve interesting subjective and objective effects through manipulation of low-level gaze parameters. This evaluation shows how designers of embodied agents can use a subtle gaze parameter like *head alignment* to reach different desired outcomes. If the agent designer wants human interlocutors to pay more attention to specific objects in the environment, possibly to learn more about them, the agent could be programmed to use high head alignment when gazing to those objects. Similarly, if the agent designer wants the agent to build a stronger relationship with the human interlocutor, increasing feelings of, e.g., rapport and trust, the agent should be programmed to use high head alignment when gazing towards the human. The model of gaze shifts offers a simple and effective means to control the low-level gaze parameters found in physiological research. By creating a mechanism for synthesizing gaze shifts in a natural, yet parameterized fashion, this model serves as a building block for creating effective social and communicative gaze behaviors. Embodied agent designers can use and build off of this core model to create rich, compelling gaze behaviors for their agents that accomplish the high-level effects they wish to achieve, as was done with the remaining gaze mechanisms in this dissertation.

The gaze shift model enables agents to execute gaze movements naturally and effectively, but the next consideration is *when* to carry out gaze shifts. This is particularly important in dyadic conversations with agents, in which the agent will by default make continuous eye contact with its human interlocutor. Chapter 4 presented the development of a conversationally situated model of gaze, focusing on the mechanism of *gaze aversion*—the intentional redirection of gaze away from the face of an interlocutor (Figure 8.1B). Gaze aversion is an important nonverbal cue that serves cognitive, intimacy-modulating, and floor management functions in conversations. Chapter 4 presented a data collection study to identify precise spatial and temporal parameters of gaze aversion from a video corpus

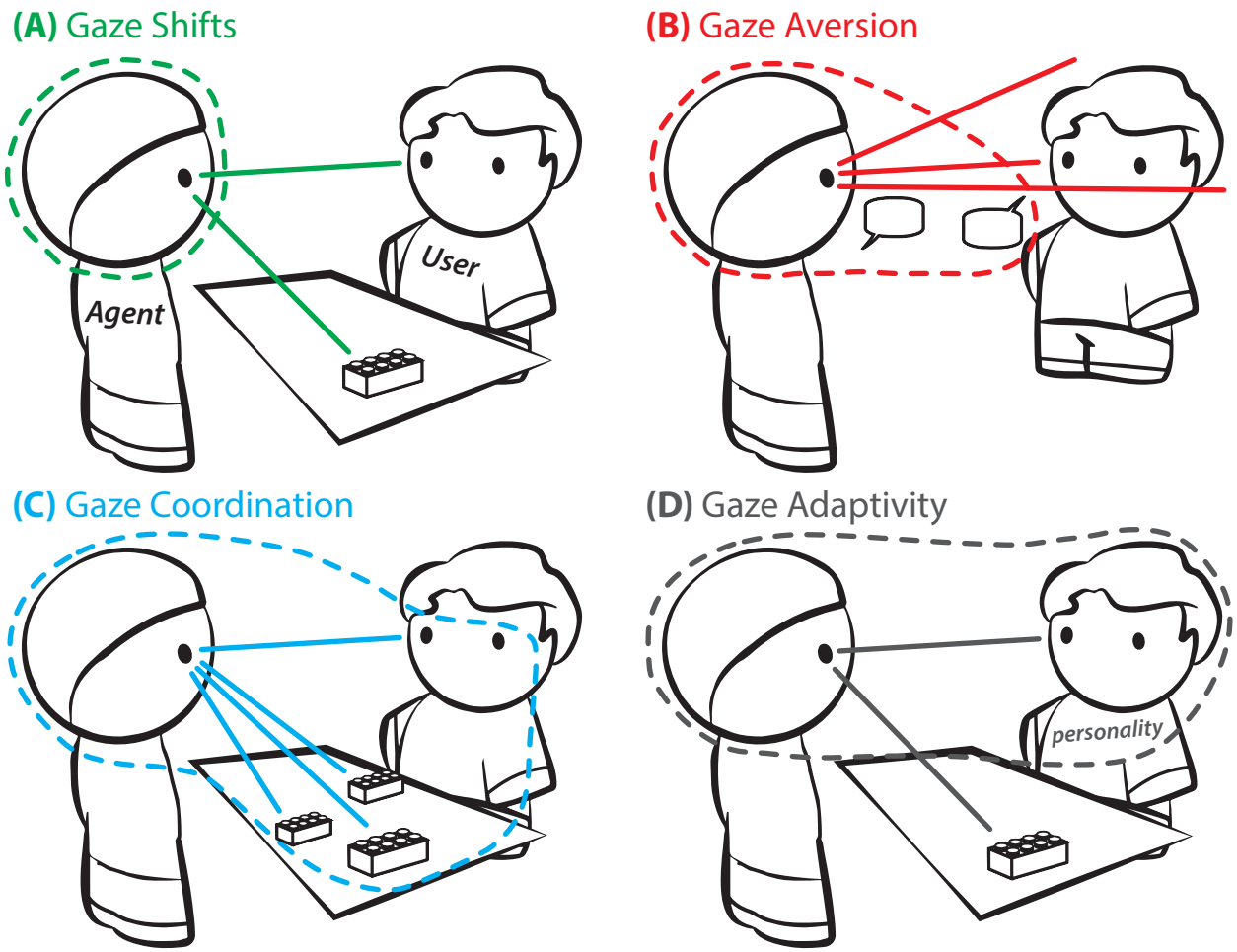

Figure 8.1: To review, this dissertation presented new understanding, models, and evaluation of four gaze mechanisms for virtual agents and social robots: (A) gaze shifts, (B) gaze aversion, (C) gaze coordination, and (D) gaze adaptivity.

of human-human interactions. Work was then presented to implement these behaviors on both virtual agent and humanlike robot systems. Evaluations of the designed gaze aversion behaviors generated by these systems demonstrated that they are perceived as intentional when expressed by an embodied artificial character, and that agents can use gaze aversions to appear more thoughtful and effectively manage the conversational floor. These results have important implications for designers of human-agent interactions. Gaze aversions should be considered as an important and useful cue for developing effective conversational interactions between humans and agents.

In more complex interactions involving physical collaboration over a shared task space, the agent must be able to effectively distribute its gaze across relevant task objects in addition to its human collaborator. By coordinating the agent's gaze with gaze motions tracked

from the human collaborator, the agent can be become more tightly situated in the task and improve collaborative outcomes. This goal is captured in a mechanism referred to as *gaze coordination*, modeled and evaluated in Chapter 5 (Figure 8.1C). Work was first presented to develop a deeper understanding of coordinated referential gaze in collaborating dyads. It turns out that although people coordinate their gaze with each other in a complex and dynamic process, general patterns can be observed and extracted.

The behavioral context for analyzing gaze coordination was the *reference-action sequence*, a pattern of interaction in which one member of the dyad makes a verbal reference to an object in the shared workspace that the other member is expected to act upon in some way. A dyadic sandwich-making task was chosen to study collaborative interactions that contain a large number of such sequences. A series of analyses of data collected in this task—utilizing epistemic network analysis—revealed how gaze coordination unfolded throughout an interaction sequence, how the gaze behaviors of individuals aligned at different phases of the interaction, and what gaze patterns indicated breakdowns and repairs in the interaction. Arguments were presented that the characterization of these patterns will generalize beyond this specific task to any interactions that involve reference-action sequences, as these sequences are commonly observed across many kinds of interactions. In addition to contributing to the growing body of knowledge on the coordination of gaze behaviors in joint activities, these analyses offer a number of design implications for technologies that engage in dyadic interactions with people.

It was next demonstrated how embodied agents stand to benefit from tracking user gaze and the knowledge of these coordinated gaze patterns. Grounded in data collected from pairs of collaborating people, A number of subtle features of human gaze coordination were identified, including timings, spatial mappings, and repair strategies. These features were built into a model of gaze coordination that enabled interactive virtual characters to interpret the gaze of their users and generate their own gaze to effectively communicate coordinated behaviors. This model also enabled virtual characters to achieve more efficient verbal referencing by signaling attention to the user and to items in the environment appropriately over time, and infer the user's current state and goals—such as confusion leading to an impending request for repair—from the user's gaze.

A user study demonstrated that gaze coordination mechanisms improve efficiency, subjective quality, and the overall level of behavioral coordination between people and interactive agents. In order to make the model more practical for users without access to eye tracking systems, techniques were introduced that allow head tracking to serve as a proxy for more precise eye tracking. A second user study demonstrated that the relaxed model retains much of the effectiveness of gaze coordination that was achieved by full eye

tracking. Furthermore, the third user study demonstrated that using this model with the head tracking available in a head-mounted display provided sufficient fidelity to improve user experience and interaction with a character in a virtual reality setting. In general, gaze coordination is a powerful strategy that results in more immersive and fluent interactive experiences with embodied agents.

Finally, it is important to consider that a one-size-fits-all approach limits an agent's ability to account for cultural and individual differences across human users. Chapter 6 presented a mechanism of *gaze adaptivity* that demonstrates how the timing of an agent's gaze shifts can be manipulated in order to express extroversion or introversion, and how this personality expressed via gaze can be matched to a user's personality in order to improve motivation in a rehabilitation setting (Figure 8.1D). That chapter specifically presented the design and evaluation of gaze mechanisms for socially assistive robots that allow the robot to match the extroversion dimension of personality to the user, thereby more effectively motivating users to repeatedly engage in a therapeutic task. These robot behaviors were designed by analyzing the gaze behavior of participants in a human-human data collection study, and were validated in an online study to express either an extroverted or introverted personality. The evaluation also demonstrated the importance of taking the user's intrinsic motivation into account when attempting to produce external motivation and increase compliance.

This work on gaze adaptivity demonstrated the importance of designing social technologies that can flexibly adapt to user behaviors and characteristics. An embodied agent that matches the personality expressed by its gaze behavior to the personality of its user can lead to positive outcomes of compliance and subjective perceptions of the agent's performance. These outcomes are particularly critical for socially assistive robots that must motivate their users to engage in physical or cognitive exercises, especially when these exercises are repetitive or boring.

The key element tying all of these mechanisms together is that they are *situated*. A critical component of the central thesis is the claim that in order to be truly effective, an agent's gaze behaviors must be tightly linked and responsive to the environment and context in which they are deployed. Generating effective gaze behavior requires more than simply pointing the agent's eyes in the "right" direction. Much of the subtlety and communicativeness of gaze comes from the details of timing, the usage of both the eyes and head, and how it is deployed contingently on a number of multimodal features in the interaction. The approaches to generating gaze movements for agents presented in this dissertation aim to capture and represent these complex contingencies in a manner that provides sufficient fidelity to serve as effective gaze mechanisms and appear natural. In

order to be the most useful to interaction designers, these mechanisms were designed to be highly controllable, adaptable to the agent's goals, and flexible enough to work for both virtual and physical embodiments.

A particular strength of the research methodology employed in this work is that it is grounded in existing research and new data collection of human-created gaze behaviors. While further work remains in order to improve the validity of various modeling and implementation decisions as well as the generalizability of the evaluation results, this dissertation provides a major step towards designing effective social gaze behaviors for embodied agents using a theoretically and empirically grounded methodology and understanding their social and cognitive impacts on people who interact with such agents.

8.1 Contributions

To review, this dissertation makes a number of design, systems, and empirical contributions to research on human-robot interaction (HRI), intelligent virtual agents (IVA), human-computer interaction (HCI), multimodal interaction, and human communication.

Design Contributions

The design contributions of this dissertation advance our understanding of human gaze mechanisms from a computational point of view. This dissertation contributes new knowledge and computational models of human gaze mechanisms, along with a set of models and controllable parameters—such as gaze target, gaze triggers, frequency, and duration—that interaction designers can use to create gaze behaviors for agents that can be manipulated to obtain targeted social and cognitive outcomes.

- New knowledge of how low-level gaze variables and gaze mechanisms might achieve high-level social and cognitive effects, connecting current knowledge in the social sciences with a more computational foundation (Chapters 3-6).
- A computational model of gaze shifts—a fundamental building block of overall gaze behavior, intentionally redirecting gaze to specific targets—that coordinates eye and head movements in a humanlike way (Chapter 3). This model serves as a core component of all the subsequent gaze models.
- A computational model of gaze aversion—intentional shifting of gaze away from a partner's face—that specifies when and in which direction agents should avert their

gaze when conversing with people in order to appear more thoughtful, regulate turn-taking, and maintain a comfortable level of intimacy (Chapter 4).

- A demonstration of how a new analysis technique, Epistemic Network Analysis, can be used to obtain a detailed and nuanced understanding of coordinated referential gaze patterns arising in physical dyadic collaborations. In particular, this analysis revealed (1) how a collaborating dyad's gaze behaviors unfold over the course of an interaction, (2) how the alignment of gaze behaviors shift throughout the interaction, and (3) how coordinated gaze behaviors differ in interaction sequences that include breakdowns and/or repairs (Chapter 5).
- A computational model of gaze coordination that specifies how an agent should deploy its gaze over a shared collaborative workspace, tightly linked to the human user's gaze, speech, and actions (Chapter 5).
- A computational model of personality-expressive gaze that specifies the frequencies and lengths of gazes toward a shared task space and toward the human partner in order to convey introversion or extroversion (Chapter 6).

Systems Contributions

Each of the computational gaze models listed above was implemented on a virtual agent and/or humanlike robot platform. Each system also required additional competencies to be implemented in order for the agent to be able to autonomously perform the task associated with the scenario it was embedded in.

- A comprehensive virtual agent framework, implemented in the Unity game engine, that was utilized for all virtual agent scenarios presented in this dissertation. This framework includes animated character models as well as behavior modules for gaze, speech, gestures, task logic, and so on (Chapters 3-5).
- An implementation of the gaze shift model for virtual agents in an educational scenario in which the agent gave one-sided lectures to a human listener while periodically gazing toward visual content supportive to the lecture (Chapter 3).
- An implementation of the gaze aversion model that autonomously plans and executes gaze aversions for virtual agents in a scenario where the agent engages in a structured conversation with a human user (Chapter 4).

- An implementation that retargets the gaze aversion model to a robot platform with more limited affordances, requiring a number of new techniques—idle motion, face tracking, and predictive filtering—in order to make it successful in a similar structured conversation scenario (Chapter 4).
- An implementation of the gaze coordination model in a virtual agent system that can autonomously collaborate with human users in a sandwich-making training scenario, including gaze tracking, speech recognition, and action tracking (Chapter 5).
- An extension of the gaze coordination implementation to utilize head pose tracking as a lower fidelity and lower cost proxy for full gaze tracking (Chapter 5).
- An implementation of the gaze coordination model and sandwich-making scenario in head-mounted virtual reality with a virtual agent (Chapter 5).
- An implementation of the personality-expressive gaze model on a robot platform that can autonomously instruct, motivate, and monitor people in a Towers of Hanoi puzzle-solving task (Chapter 6).

Empirical Contributions

In addition to modeling and implementation, each situated gaze mechanism presented in this dissertation was evaluated in one or more user studies, contextualized in specific scenarios. These studies provide a better understanding of the social, cognitive, and behavioral outcomes achievable via carefully deployed gaze mechanisms in human-agent interaction. All studies utilized carefully designed experimental paradigms for studying how subtle manipulations in situated gaze mechanisms can target specific outcomes under various conditions.

- Evidence across all studies that seemingly subtle manipulations in an agent's gaze behavior can lead to powerful high-level interaction outcomes (Chapters 3-6).
- Evidence that the presence of an agent can improve recall in an educational scenario, compared with having the same lecture content expressed through audio alone (Chapter 3).
- Evidence that head alignment, one parameter of the gaze shift model, can be manipulated to achieve either better rapport or better recall in a lecture-style educational scenario (Chapter 3).

- Evidence that virtual agents using gaze aversions generated by the presented computational model were perceived as thinking, elicited more disclosure from human interlocutors, and effectively managed turn-taking (Chapter 4).
- Evidence that gaze aversions expressed by social robots utilizing the presented gaze aversion model are perceived as intentional, and that robots can use gaze aversions to appear more thoughtful and effectively manage the conversational floor (Chapter 4).
- Evidence that the gaze coordination model can improve collaborative outcomes for a virtual agent task-training system in terms of task time, number of errors made, recall, and the subjective preferences of users (Chapter 5).
- Evidence that the gaze coordination model is comparably effective even when gaze tracking is replaced with lower fidelity head pose tracking (Chapter 5).
- Evidence from an online study that the personality-expressive gaze model is effective in conveying introversion or extroversion noticeably for robots (Chapter 6).
- Evidence that matching a robot's personality—introversion or extroversion expressed via gaze alone—to that of a user can positively motivate them to participate longer in a repetitive task, particularly when their intrinsic motivation going into the task is low (Chapter 6).

8.2 Closing Remarks

While describing classic Disney animation principles, Thomas and Johnston (1981) observed that for animating characters, "the eyes are the most important part of an expression, and must be drawn with extreme care. Any jitter or false move ... destroys both communication and believability." Interactivity only serves to make this problem more difficult. Designing gaze behaviors for virtual agents and robots that can achieve specifically targeted high-level outcomes has long been a difficult problem. This dissertation has demonstrated several mechanisms by which embodied agents can deliver social and cognitive benefits through their gaze behavior, providing support for the central thesis that humanlike gaze mechanisms can enable both virtual agents and social robots to more effectively communicate with human users in situated interaction contexts.

The four mechanisms that were presented show that embodied agents can use their gaze to improve users' comprehension of information, increase feelings of rapport, regulate the flow of conversation, modulate intimacy, improve collaboration outcomes, and increase

motivation. I have also argued that these benefits can be achieved by following a process of gaining a theoretically and empirically grounded understanding of human gaze behaviors as they are used in typical dyadic interaction, carefully designing computational models for embodied agents that reflect and facilitate these situated behaviors, and testing how these mechanisms could be manipulated to achieve particular social and cognitive outcomes.

With this dissertation work, I have come to appreciate gaze as an incredibly complex and powerful social cue. Through subtle changes in gaze, people can achieve a wide range of social and communicative goals, affecting their partner of interaction in a variety of different ways. Gaze cues, as with all embodied communication cues, hold a strong fundamental connection with key social, cognitive, and task outcomes. This connection reveals an opportunity for designing effective interactions with embodied agents. By presenting models that synthesize gaze mechanisms in a natural, yet parameterized fashion, as well as contingently on the human user's behaviors and characteristics, this dissertation provides a framework for creating high-level social and communicative behaviors. By manipulating and combining these mechanisms in different ways, I believe that embodied agents will have access to a rich new source of possible gaze behaviors, resulting in human-agent interactions that are more effective and rewarding.

A APPENDIX: STUDY QUESTIONNAIRES

Study questionnaires for subjective measures from the evaluations of each of the four gaze mechanisms (Chapters 3-6) are presented on the following pages.

Gaze Shifts (Chapter 3)

Part I. Your impressions of the lecturer

*The following part consists of questions regarding **your observation of the lecturer**. Read each question or statement and indicate your answer in a scale from 1 to 7.*

The lecturer's appearance was very masculine/very feminine.

Very masculine	1	2	3	4	5	6	7	Very feminine

The lecturer's voice was very masculine/very feminine.

Very masculine	1	2	3	4	5	6	7	Very feminine

Please rate the lecturer on the following scales:

Unfriendly	1	2	3	4	5	6	7	Friendly
Attractive	1	2	3	4	5	6	7	Unattractive
Likeable	1	2	3	4	5	6	7	Not likeable
Unhelpful	1	2	3	4	5	6	7	Helpful
Natural	1	2	3	4	5	6	7	Unnatural
Incompetent	1	2	3	4	5	6	7	Competent
Knowledgeable	1	2	3	4	5	6	7	Ignorant
Unintelligent	1	2	3	4	5	6	7	Intelligent
Trustworthy	1	2	3	4	5	6	7	Untrustworthy
Dishonest	1	2	3	4	5	6	7	Honest
Novice	1	2	3	4	5	6	7	Expert

Please indicate your agreement/disagreement with the following statements:

The lecturer liked me.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I liked the lecturer.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I believe that the lecturer is similar to me.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I believe the lecturer thinks in the same way I do.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I believe the lecturer connected well with me.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I believe the lecturer effectively used the map to convey information to me.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I believe the lecturer was good at communicating.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

I would willingly disclose a great deal of positive and negative details about myself to the lecturer.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

The lecturer presented the material in a way that kept me interested.

Strongly agree	1	2	3	4	5	6	7	Strongly disagree

How much did the lecturer direct his/her attention towards **you**?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

How much did the lecturer direct his/her attention towards **the map**?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Part II. Your impressions of the content of the lecture

*The following part consists of questions regarding **the content of the lecture**. Rate the content of the lecture for each scale from 1 to 7.*

Unreliable	1	2	3	4	5	6	7	Reliable

Boring	1	2	3	4	5	6	7	Enjoyable
Difficult	1	2	3	4	5	6	7	Easy

Part III. About You

The following part consists of questions regarding you, and how you felt during the lecture. Please indicate ***how you felt during the lecture*** *for each scale from 1 to 7.*

Engaged	1	2	3	4	5	6	7	Not engaged
Dissatisfied	1	2	3	4	5	6	7	Satisfied
Attentive	1	2	3	4	5	6	7	Inattentive
Disinterested	1	2	3	4	5	6	7	Interested
Focused	1	2	3	4	5	6	7	Unfocused
Confident	1	2	3	4	5	6	7	Unconfident

Gaze Aversion (Chapter 4)

Task 1 (Thinking)

Please rate the virtual agent in terms of its responsiveness to your questions:

How acceptable was the responsiveness of the virtual agent?

Very Unacceptable	1	2	3	4	5	6	7	Very Acceptable

Task 2 (Thoughtfulness)

Please indicate your agreement/disagreement with the following statements regarding the virtual agent's response to your question:

The virtual agent gave a thoughtful response.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent gave a truthful response.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent disclosed a lot in the response.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent behaved naturally while giving the response.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent gave a creative response.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Task 3 (Disclosure)

Please indicate your agreement/disagreement with the following statements regarding the virtual agent:

The virtual agent liked me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I liked the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe that the virtual agent is similar to me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent thinks in the same way I do.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent connected well with me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent was good at communicating.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I would willingly disclose a great deal of positive and negative details about myself to the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I felt comfortable conversing with the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent behaved in a natural way.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Task 4 (Turn-taking)

Please indicate your agreement/disagreement with the following statements regarding the virtual agent:

The virtual agent liked me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I liked the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe that the virtual agent is similar to me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent thinks in the same way I do.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent connected well with me.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I believe the virtual agent was good at communicating.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I would willingly disclose a great deal of positive and negative details about myself to the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

I felt comfortable conversing with the virtual agent.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

The virtual agent behaved in a natural way.

Strongly disagree	1	2	3	4	5	6	7	Strongly agree

Gaze Coordination (Chapter 5)

Please rate the particular version of Jason that you just interacted with. Rate the extent to which you agree or disagree with the following statements, where 1 means you strongly disagree and 7 means you strongly agree:

I believe that this version of Jason was ... *

	1. Strongly disagree	2.	3.	4.	5.	6.	7. Strongly agree
... competent.	○	○	○	○	○	○	○
... human-like in appearance.	○	○	○	○	○	○	○
... human-like in behavior.	○	○	○	○	○	○	○
... conscious of me.	○	○	○	○	○	○	○
... listening to me when I talked.	○	○	○	○	○	○	○
... dedicated to the tasks.	○	○	○	○	○	○	○
... knowledgeable.	○	○	○	○	○	○	○
... lively.	○	○	○	○	○	○	○
... engaged in the task.	○	○	○	○	○	○	○
... expressive.	○	○	○	○	○	○	○
... watchful.	○	○	○	○	○	○	○
... good at following social norms.	○	○	○	○	○	○	○
... attentive.	○	○	○	○	○	○	○
... unqualified for this task.	○	○	○	○	○	○	○
... intelligent.	○	○	○	○	○	○	○
... sensitive to my needs.	○	○	○	○	○	○	○

	○	○	○	○	○	○	○
... an active part of the task.	○	○	○	○	○	○	○
... responsible.	○	○	○	○	○	○	○
... helpful.	○	○	○	○	○	○	○
... observant.	○	○	○	○	○	○	○
... experienced.	○	○	○	○	○	○	○
... excited to help me.	○	○	○	○	○	○	○

In 1-2 sentences, please answer the following questions in your own words:

What did you like about Jason's behaviors as an instructor in this task? *

What did you not like about Jason's behaviors as an instructor in this task? *

Gaze Adaptivity (Chapter 6)

Indicate to what extent you agree or disagree with the following statements:

I believe that Meka . . . *

	1. Strongly disagree	2. Disagree	3. Somewhat disagree	4. Neither agree or disagree	5. Somewhat agree	6. Agree	7. Strongly agree
1. ... was a good instructor.	○	○	○	○	○	○	○
2. ... provided helpful advice on how to solve the puzzles.	○	○	○	○	○	○	○
3. ... paid close attention to me.	○	○	○	○	○	○	○
4. ... was good at motivating me to keep working on the puzzles.	○	○	○	○	○	○	○
5. ... frequently made or attempted to make eye contact with me.	○	○	○	○	○	○	○
6. ... influenced me to work longer.	○	○	○	○	○	○	○
7. ... moved its eyes and head in a natural way.	○	○	○	○	○	○	○
8. ... provided enough information so I could solve the puzzles efficiently.	○	○	○	○	○	○	○
9. ... provided encouraging feedback during the puzzles.	○	○	○	○	○	○	○
10. ... helped me to improve my skill at solving the puzzles.	○	○	○	○	○	○	○

In 1-2 sentences, please answer the following questions in your own words:

What did you like about Meka's behavior? *

What did you not like about Meka's behavior? *

What was your motivation to work on the puzzles for as long as you did? *

REFERENCES

Abele, Andrea. 1986. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior* 10(2):83–101.

Admoni, Henny, Caroline Bank, Joshua Tan, Mariya Toneva, and Brian Scassellati. 2011. Robot gaze does not reflexively cue human attention. In *Proceedings of the 33rd annual conference of the cognitive science society, boston, ma, usa*, 1983–1988. Citeseer.

Admoni, Henny, Christopher Datsikas, and Brian Scassellati. 2014. Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the 36th annual conference of the cognitive science society (cogsci '14)*.

Admoni, Henny, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. 2013. Are you looking at me?: Perception of robot attention is mediated by gaze type and group size. In *Proceedings of the 8th acm/ieee international conference on human-robot interaction (hri '13)*, 389–396. IEEE Press.

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

Al Moubayed, Samer, Jens Edlund, and Jonas Beskow. 2012. Taming mona lisa: Communicating gaze faithfully in 2d and 3d facial projections. *ACM Transactions on Interactive Intelligent Systems* 1(2):11:1–11:25.

Al Moubayed, Samer, and Gabriel Skantze. 2012. Perception of gaze direction for situated interaction. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, 3. ACM.

Albrecht, Irene, Jörg Haber, and Hans-Peter Seidel. 2002. Automatic generation of nonverbal facial expressions from speech. In *Advances in modelling, animation and rendering*, 283–293. Springer.

Anderson, Clare, AW Wales, and James A Horne. 2010. Pvt lapses differ according to eyes open, closed, or looking away. *Sleep* 33(2):197–204.

Andrist, Sean, Iolanda Leite, and Jill Lehman. 2013a. Fun and fair: influencing turn-taking in a multi-party game with a virtual agent. In *Proceedings of the 12th international conference on interaction design and children*, 352–355. ACM.

Andrist, Sean, Bilge Mutlu, and Michael Gleicher. 2013b. Conversational gaze aversion for virtual agents. In *Intelligent virtual agents*, 249–262. Springer.

Andrist, Sean, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: Matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 3603–3612. ACM.

Andrist, Sean, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. 2012a. Designing effective gaze mechanisms for virtual agents. In *Proceedings of the 2012 acm annual conference on human factors in computing systems*, 705–714. ACM.

———. 2012b. A head-eye coordination model for animating gaze shifts of virtual characters. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, 4:1–4:6. ACM.

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*, 25–32. ACM.

Argyle, M. 1988. *Bodily communication*, vol. 581. Taylor & Francis.

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

Argyle, Michael, and Janet Dean. 1965. Eye-contact, distance and affiliation. *Sociometry* 289–304.

Argyle, Michael, and Roger Ingham. 1972. Gaze, mutual gaze, and proximity. *Semiotica* 6(1):32–49.

Bailenson, Jeremy N, Jim Blascovich, Andrew C Beall, and Jack M Loomis. 2001. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence* 10(6):583–598.

———. 2003. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin* 29(7):819–833.

Bailenson, J.N., and N. Yee. 2005. Digital chameleons. *Psychological Science* 16(10):814.

Bailenson, J.N., N. Yee, D. Merget, and R. Schroeder. 2006. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments* 15(4):359–372.

Bailly, Gérard, Stephan Raidt, and Frédéric Elisei. 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication* 52(6):598–612.

Bainbridge, WilmaA., JustinW. Hart, ElizabethS. Kim, and Brian Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics* 3(1):41–52.

Baldwin, Dare A. 1995. Understanding the link between joint attention and language. *Joint attention: Its origins and role in development* 131–158.

Bard, Ellen Gurman, Robin Hill, and Manabu Arai. 2009. Referring and gaze alignment: Accessibility is alive and well in situated dialogue 1246–1251.

Baron-Cohen, Simon, Ruth Campbell, Annette Karmiloff-Smith, Julia Grant, and Jane Walker. 1995. Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology* 13(4):379–398.

Bavelas, Janet Beavin, Linda Coates, and Trudy Johnson. 2002. Listener responses as a collaborative process: The role of gaze. *Journal of Communication* 52(3):566–580.

Bayliss, Andrew P, Giuseppe di Pellegrino, and Steven P Tipper. 2005. Sex differences in eye gaze and symbolic cueing of attention. *The Quarterly Journal of Experimental Psychology* 58(4):631–650.

Bayliss, A.P., M.A. Paul, P.R. Cannon, and S.P. Tipper. 2006. Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic bulletin & review* 13(6):1061–1066.

Beattie, Geoffrey W. 1981. A further investigation of the cognitive interference hypothesis of gaze patterns during conversation. *British Journal of Social Psychology* 20(4):243–248.

Becker, WOLFGANG, and ALBERT F Fuchs. 1988. Lid-eye coordination during vertical gaze changes in man and monkey. *Journal of neurophysiology* 60(4):1227–1252.

Becvar, Amaya, James Hollan, and Edwin Hutchins. 2008. Representational gestures as cognitive artifacts for developing theories in a scientific laboratory. In *Resources, co-evolution and artifacts*, 117–143. Springer.

Bee, Nikolaus, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, and Marc Cavazza. 2010. Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, 9:1–9:8. ACM.

Beebe, S.A. 1976. Effects of eye contact, posture and vocal inflection upon credibility and comprehension.

Bennewitz, Maren, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. 2005. Integrating vision and speech for conversations with multiple persons. In *Intelligent robots and systems, 2005.(iros 2005). 2005 ieee/rsj international conference on*, 2523–2528. IEEE.

Bente, Gary, William C Donaghy, and Dorit Suwelack. 1998. Sex differences in body movement and visual attention: An integrated analysis of movement and gaze in mixed-sex dyads. *Journal of Nonverbal Behavior* 22(1):31–58.

Beskow, J. 1997. Animation of talking agents. In *Proceedings of avsp*, vol. 97, 149–152. Citeseer.

Bevacqua, Elisabetta. 2013. *A survey of listener behavior and listener models for embodied conversational agents*. Taylor & Francis.

Bickmore, T., and J. Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the sigchi conference on human factors in computing systems*, 396–403. ACM.

Bickmore, Timothy W, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *Chi'05 extended abstracts on human factors in computing systems*, 1212–1215. ACM.

Bickmore, Timothy W, and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12(2):293–327.

Binder, Marc D., Nobutaka Hirokawa, and Uwe Windhorst, eds. 2009. *Gaze shift*, 1676–1676. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bohus, Dan, and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, 5. ACM.

Boucher, Jean-David, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. 2012. I reach faster when i see you look: Gaze effects in human-human and human-robot face-to-face cooperation. *Frontiers in neurorobotics* 6(3):1–11.

Brandes, Ulrik, and Thomas Erlebach. 2005. *Network analysis: methodological foundations*, vol. 3418. Springer Science & Business Media.

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

Breazeal, Cynthia. 2003. Toward sociable robots. *Robotics and autonomous systems* 42(3): 167–175.

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 *2005 ieee/rsj international conference on intelligent robots and systems*, 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.

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.

Brennan, Susan E, JE Hanna, GJ Zelinsky, and Kelly J Savietta. 2012. Eye gaze cues for coordination in collaborative tasks. In *Duet 2012 workshop: Dual eye tracking in csce. 2012 acm conference on computer supported cooperative work*, vol. 9.

Brooks, Andrew G, and Cynthia Breazeal. 2006. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction*, 297–304. ACM.

Brooks, Rodney A, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew M Williamson. 1999. The cog project: Building a humanoid robot. In *Computation for metaphors, analogy, and agents*, 52–87. Springer.

Brown-Schmidt, Sarah, Ellen Campana, and Michael K Tanenhaus. 2005. Real-time reference resolution by naïve participants during a task-based unscripted conversation. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* 153–171.

Bruno, Barbara, Fulvio Mastrogiovanni, and Antonio Sgorbissa. 2013. Functional requirements and design issues for a socially assistive robot for elderly people with mild cognitive impairments. In *Ro-man, 2013 ieee*, 768–773. IEEE.

Bull, Ray, and Elizabeth Gibson-Robinson. 1981. The influences of eye-gaze, style of dress, and locality on the amounts of money donated to a charity. *Human Relations* 34(10): 895–905.

Buller, David B, and Judee K Burgoon. 1998. Emotional expression in the deception process. *Handbook of communication and emotion* 381–402.

Burgoon, J.K., D.A. Coker, and R.A. Coker. 1986. Communicative effects of gaze behavior. *Human Communication Research* 12(4):495–524.

Burroughs, Nancy F. 2007. A reinvestigation of the relationship of teacher nonverbal immediacy and student compliance-resistance with learning. *Communication Education* 56(4):453–475.

Busso, Carlos, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. 2008. Learning expressive human-like head motion sequences from speech. In *Data-driven 3d facial animation*, 113–131. Springer.

Byom, Lindsey J, and Bilge Mutlu. 2013. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience* 7:1–12.

Cafaro, A., R. Gaito, and H. Vilhjálmsson. 2009. Animating idle gaze in public places. In *Intelligent virtual agents*, 250–256. Springer.

Cafaro, Angelo, Hannes Högni Vilhjálmsson, Timothy Bickmore, Dirk Heylen, Kamilla Rún Jóhannsdóttir, and Gunnar Steinn Valgarðsson. 2012. First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Intelligent virtual agents*, 67–80. Springer.

Campana, Ellen, Jason Baldridge, John Dowding, Beth Ann Hockey, Roger W Remington, and Leland S Stone. 2001. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the 2001 workshop on perceptive user interfaces*, 1–5. ACM.

Cappella, Joseph N, and Catherine Pelachaud. 2002. Rules for responsive robots. *Stability and change in relationships* 325–354.

Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999a. Embodiment in conversational interfaces: Rea. In *Proceedings of the sigchi conference on human factors in computing systems*, 520–527. ACM.

Cassell, Justine. 2000. *Embodied conversational agents*. MIT press.

———. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22(4):67–83.

Cassell, Justine, Stefan Kopp, Paul Tepper, Kim Ferriman, and Kristina Striegnitz. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics* 133–160.

Cassell, Justine, and Andrea Tartaro. 2007. Intersubjectivity in human–agent interaction. *Interaction studies* 8(3):391–410.

Cassell, Justine, Obed E. Torres, and Scott Prevost. 1998. Turn taking vs. discourse structure: How best to model multimodal conversation. In *Machine conversations*, 143–154. Kluwer.

Cassell, Justine, Obed E Torres, and Scott Prevost. 1999b. Turn taking versus discourse structure. In *Machine conversations*, 143–153. Springer.

Cassin, Barbara, Sheila Solomon, and Melvin L Rubin. 1984. *Dictionary of eye terminology*. Triad Pub. Co.

Chawla, Purnima, and Robert M Krauss. 1994. Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology* 30(6):580–601.

Chidambaram, Vijay, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction*, 293–300. ACM.

Choi, Jung Ju, Yunkyung Kim, and Sonya S Kwak. 2013. Have you ever lied?: The impacts of gaze avoidance on people's perception of a robot. In *Proceedings of the 8th acm/ieee international conference on human-robot interaction*, 105–106. IEEE Press.

Christophel, Diane M. 1990. The relationships among teacher immediacy behaviors, student motivation, and learning. *Communication education* 39(4):323–340.

Cig, Cagla, Zerrin Kasap, Arjan Egges, and Nadia Magnenat-Thalmann. 2010. Realistic emotional gaze and head behavior generation based on arousal and dominance factors. In *Motion in games*, 278–289. Springer.

Clark, Alan T, and Darren Gergle. 2011. Mobile dual eye-tracking methods: challenges and opportunities. In *Proc. of international workshop on dual eye tracking*.

Clark, A.T., and Darren Gergle. 2012. Know what i'm talking about? dual eye-tracking in multimodal reference resolution. In *Duet 2012: Dual eye tracking workshop at cscw*.

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

———. 2003. Pointing and placing. *Pointing: Where language, culture, and cognition meet* 243–268.

———. 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.

Clark, Herbert H, and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22(1):1–39.

Condon, William S, and William D Ogston. 1971. Speech and body motion synchrony of the speaker-hearer. *The perception of Language* 150–184.

Das, Dipankar, Md Golam Rashed, Yoshinori Kobayashi, and Yoshinori Kuno. 2014. Recognizing gaze pattern for human robot interaction. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction*, 142–143. ACM.

Davies, D Roy, Gerald Matthews, Rob B Stammers, and Steve J Westerman. 2013. *Human performance: Cognition, stress and individual differences*. Psychology Press.

Deng, Z., JP Lewis, and U. Neumann. 2005. Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications* 24–30.

D'Mello, Sidney, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies* 70(5):377–398.

Doherty-Sneddon, Gwyneth, and FG Phelps. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition* 33(4):727–733.

Doughty, Michael J. 2001. Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry & Vision Science* 78(10):712–725.

Downing, Paul, Chris Dodds, and David Bray. 2004. Why does the gaze of others direct visual attention? *Visual Cognition* 11(1):71–79.

Driver, Jon, Greg Davis, Paola Ricciardelli, Polly Kidd, Emma Maxwell, and Simon Baron-Cohen. 1999. Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition* 6(5):509–540.

Duncan, Starkey, Barbara G Kanki, Hartmut Mokros, and Donald W Fiske. 1984. Pseudounilaterality, simple-rate variables, and other ills to which interaction research is heir. *Journal of Personality and Social Psychology* 46(6):1335.

Ehrlichman, Howard, and Dragana Micic. 2012. Why do people move their eyes when they think? *Current Directions in Psychological Science* 21(2):96–100.

Eichner, Tobias, Helmut Prendinger, Elisabeth André, and Mitsuru Ishizuka. 2007. Attentive presentation agents. In *Intelligent virtual agents*, 283–295. Springer.

Emery, NJ. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews* 24(6):581–604.

Endsley, Mica R. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37(1):32–64.

Eriksson, Jon, Maja J Mataric, and C Winstein. 2005. Hands-off assistive robotics for post-stroke arm rehabilitation. In *Proc. ieee international conference on rehabilitation robotics (icorr'05)*, 21–24.

Evinger, Craig, Karen A Manning, John J Pellegrini, Michele A Basso, Alice S Powers, and Patrick A Sibony. 1994. Not looking while leaping: the linkage of blinking and saccadic gaze shifts. *Experimental brain research* 100(2):337–344.

Evinger, Craig, Karen A Manning, and Patrick A Sibony. 1991. Eyelid movements. mechanisms and normal data. *Investigative ophthalmology & visual science* 32(2):387–400.

Exline, Ralph, David Gray, and Dorothy Schuette. 1965. Visual behavior in a dyad as affected by interview content and sex of respondent. *Journal of Personality and Social Psychology* 95:201–209.

Fasola, Juan, and Maja J Matarić. 2012. Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE* 100(8):2512–2526.

Feil-Seifer, David, and Maja J Mataric. 2005. Defining socially assistive robotics. In *Proceedings of the 9th international conference on rehabilitation robotics*, 465–468. IEEE.

Feil-Seifer, David, and Maja J Matarić. 2009. Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In *Experimental robotics*, 201–210. Springer.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42(3):143–166.

Freedman, E.G., and D.L. Sparks. 1997. Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: evidence for a gaze displacement command. *Journal of neurophysiology* 78(3):1669.

———. 2000. Coordination of the eyes and head: movement kinematics. *Experimental Brain Research* 131(1):22–32.

Friesen, Chris Kelland, Jelena Ristic, and Alan Kingstone. 2004. Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance* 30(2):319–329.

Frischen, A., A.P. Bayliss, and S.P. Tipper. 2007. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin* 133(4):694–724.

Fry, R., and G.F. Smith. 1975. The effects of feedback and eye contact on performance of a digit-coding task. *The Journal of Social Psychology*.

Fukayama, A., T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita. 2002. Messages embedded in gaze of interface agents—impression management with agent's gaze. In *Proceedings of the sigchi conference on human factors in computing systems: Changing our world, changing ourselves*, 41–48. ACM.

Fuller, J.H. 1992. Head movement propensity. *Experimental Brain Research* 92(1):152–164.

Fullwood, C., and G. Doherty-Sneddon. 2006. Effect of gazing at the camera during a video link on recall. *Applied Ergonomics* 37(2):167–175.

Fussell, Susan R, Leslie D Setlock, and Elizabeth M Parker. 2003. Where do helpers look?: gaze targets during collaborative physical tasks. In *Chi'03 extended abstracts on human factors in computing systems*, 768–769. ACM.

Garau, M., M. Slater, S. Bee, and M.A. Sasse. 2001. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the sigchi conference on human factors in computing systems*, 309–316. ACM.

Garau, Maia, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. 2003. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the sigchi conference on human factors in computing systems*, 529–536. ACM.

Garrido-Jurado, S, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47(6):2280–2292.

Garrod, Simon, and Martin J Pickering. 2004. Why is conversation so easy? *Trends in cognitive sciences* 8(1):8–11.

Gergle, Darren, and Alan T Clark. 2011. See what i'm saying?: using dyadic mobile eye tracking to study collaborative reference. In *Proceedings of the acm 2011 conference on computer supported cooperative work*, 435–444. ACM.

Gergle, Darren, Robert E Kraut, and Susan R Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction* 28(1):1–39.

Gifford, Robert. 2006. Personality and nonverbal behavior: A complex conundrum. *Handbook of nonverbal communication* 151–179.

Giles, Howard, and Peter Powesland. 1997. *Accommodation theory*. Springer.

Glenberg, Arthur M, Jennifer L Schroeder, and David A Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition* 26(4): 651–658.

Goetz, Jennifer, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th ieee international workshop on robot and human interactive communication*, 55–60. IEEE.

Goldberg, G.N., C.A. Kiesler, and B.E. Collins. 1969. Visual behavior and face-to-face distance during interaction. *Sociometry* 43–53.

Goldring, J.E., M.C. Dorris, B.D. Corneil, P.A. Ballantyne, and D.R. Munoz. 1996. Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental brain research* 111(1):68–78.

Goossens, H.H.L.M., and A.J.V. Opstal. 1997. Human eye-head coordination in two dimensions under different sensorimotor conditions. *Experimental Brain Research* 114(3): 542–560.

Gratch, Jonathan, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *Intelligent virtual agents*, 125–138. Springer.

Greene, Peter R. 1986. Gaussian and poisson blink statistics: A preliminary study. *Biomedical Engineering, IEEE Transactions on* (3):359–361.

Gregory, Richard. 1997. *Eye and brain: The psychology of seeing*. Princeton University Press.

Griffin, Zenzi M. 2004. The eyes are right when the mouth is wrong. *Psychological Science* 15(12):814–821.

Griffin, Zenzi M, and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science* 11(4):274–279.

Griffin, Z.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 *2013 ieee/rsj international conference on intelligent robots and systems*, 4622–4629. IEEE.

Gu, Erdan, Sooha Park Lee, Jeremy B Badler, and Norman I Badler. 2008. Eye movements, saccades, and multiparty conversations. In *Data-driven 3d facial animation*, 79–97. Springer.

Guay, Frédéric, Geneviève A Mageau, and Robert J Vallerand. 2003. On the hierarchical structure of self-determined motivation: A test of top-down, bottom-up, reciprocal, and horizontal effects. *Personality and Social Psychology Bulletin* 29(8):992–1004.

Guitton, D., and M. Volle. 1987. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology* 58(3):427.

Guitton, Daniel, Raymond Simard, and François Codère. 1991. Upper eyelid movements measured with a search coil during blinks and vertical saccades. *Investigative ophthalmology & visual science* 32(13):3298–3305.

Ham, Jaap, Raymond H Cuijpers, and John-John Cabibihan. 2015. Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics* 7(4):479–487.

Hanna, Joy E, and Susan E Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language* 57(4): 596–615.

Harris, M.J., and R. Rosenthal. 2005. No more teachers' dirty looks: Effects of teacher nonverbal behavior on student outcomes. *Applications of nonverbal communication* 157–192.

Hayhoe, Mary, and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9(4):188–194.

Heck, R.M. 2007. Automated authoring of quality human motion for interactive environments. Ph.D. thesis, University of Wisconsin–Madison.

Hendrick, Clyde, and Steven R Brown. 1971. Introversion, extraversion, and interpersonal attraction.

Hendriks-Jansen, Horst. 1996. *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*. MIT Press.

Heylen, D., I. Van Es, A. Nijholt, and B. Van Dijk. 2002. Experimenting with the gaze of a conversational agent. In *Proceedings international class workshop on natural, intelligent and effective interaction in multimodal dialogue systems*, 93–100.

Heylen, Dirk, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. 2007. Searching for prototypical facial feedback signals. In *Intelligent virtual agents*, 147–153. Springer.

Hietanen, J.K. 1999. Does your gaze direction and head orientation shift my visual attention? *Neuroreport* 10(16):3443.

Hirst, Graeme, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech communication* 15(3):213–229.

Hjalmarsson, Anna, and Catharine Oertel. 2012. Gaze direction as a back-channel inviting cue in dialogue. In *Iva 2012 workshop on realtime conversational virtual agents*, vol. 9.

Hood, Bruce M, J Douglas Willen, and Jon Driver. 1998. Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science* 9(2):131–134.

Hoque, Mohammed Moshiul, Dipankar Das, Tomomi Onuki, Yoshinori Kobayashi, and Yoshinori Kuno. 2012. Robotic system controlling target human's attention. In *Intelligent computing theories and applications*, 534–544. Springer.

Hornik, Jacob, and Shmuel Ellis. 1988. Strategies to secure compliance for a mall intercept interview. *Public Opinion Quarterly* 52(4):539–551.

Huang, Chien-Ming, and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction*, 25–32. ACM.

———. 2014. Multivariate evaluation of interactive robot systems. *Autonomous Robots* 37(4):335–349.

———. 2016. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th acm/ieee international conference on human-robot interaction (hri)*, 83–90. IEEE.

Iizuka, Yuichi. 1992. Extraversion, introversion, and visual interaction. *Perceptual and motor skills* 74(1):43–50.

Imai, M., T. Kanda, T. Ono, H. Ishiguro, and K. Mase. 2002. Robot mediated round table: Analysis of the effect of robot's gaze. In *Robot and human interactive communication, 2002. proceedings. 11th ieee international workshop on*, 411–416. IEEE.

Ipeirotis, P.G. 2010. Demographics of Mechanical Turk. Tech. Rep. CeDER-10-01. Accessed on 10-Mar-2010 at http://hdl.handle.net/2451/29585.

Ito, Akira, Shunsuke Hayakawa, and Tazunori Terada. 2004. Why robots need body for mind communication-an attempt of eye-contact between human and robot. In *Robot and human interactive communication, 2004. roman 2004. 13th ieee international workshop on*, 473–478. IEEE.

Itti, L., N. Dhavale, and F. Pighin. 2006. Photorealistic attention-based gaze animation. In *2006 ieee international conference on multimedia and expo*, 521–524. IEEE.

Itti, Laurent. 2000. Models of bottom-up and top-down visual attention. Ph.D. thesis, California Institute of Technology.

Itti, Laurent, Nitin Dhavale, and Frederic Pighin. 2004. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Optical science and technology, spie's 48th annual meeting*, 64–78. International Society for Optics and Photonics.

Izard, Carroll E. 1991. *The psychology of emotions*. Springer Science & Business Media.

Jan, Dušan, David Herrera, Bilyana Martinovski, David Novick, and David Traum. 2007. A computational model of culture-specific conversational behavior. In *Intelligent virtual agents*, 45–56. Springer.

John, Oliver P, and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2: 102–138.

Johns, Murray W, Andrew Tucker, Robert Chapman, Kate Crowley, and Natalie Michael. 2007. Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers. *Somnologie-Schlafforschung und Schlafmedizin* 11(4):234–242.

Johnson, W Lewis, Jeff W Rickel, and James C Lester. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education* 11(1):47–78.

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.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering* (82):35–45.

Kang, Kyong Il, Sanford Freedman, Maja J Mataric, Mark J Cunningham, and Becky Lopez. 2005. A hands-off physical therapy assistance robot for cardiac patients. In *Proceedings of the 9th international conference on rehabilitation robotics*, 337–340. IEEE.

Kang, Sin-Hwa, Jonathan Gratch, Candy Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. 2012. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems (aamas)*, 63–70.

Karreman, Daphne, Gilberto Sepúlveda Bradford, Betsy Van Dijk, Manja Lohse, and Vanessa Evers. 2013. What happens when a robot favors someone? how a tour guide robot uses gaze behavior to address multiple persons while storytelling about art. In *Human-robot interaction (hri), 2013 8th acm/ieee international conference on*, 157–158. IEEE.

Karreman, Daphne E, Geke DS Ludden, Elisabeth MAG van Dijk, and Vanessa Evers. 2015. How can a tour guide robot's orientation influence visitors' orientation and formations? *New Frontiers in Human-Robot Interaction* 92.

Kawaguchik, Ikkaku, Hideaki Kuzuoka, and Yusuke Suzuki. 2015. Study on gaze direction perception of face image displayed on rotatable flat display. In *Proceedings of the 33rd annual acm conference on human factors in computing systems (chi '15)*, 1729–1737. ACM.

Kelley, D.H., and J. Gorham. 1988. Effects of immediacy on recall of information. *Communication Education*.

Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26(1):22–63.

Kennedy, James, Paul Baxter, and Tony Belpaeme. 2015. Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics* 7(2):293–308.

Khullar, S.C., and N.I. Badler. 2001. Where to look? automating attending behaviors of virtual human characters. *Autonomous Agents and Multi-Agent Systems* 4(1):9–23.

Kidd, C.D., and C. Breazeal. 2004. Effect of a robot on user perceptions. In *Proc. iros*, vol. 4, 3559–3564.

Kidd, Cory David. 2008. Designing for long-term human-robot interaction and application to weight loss.

Kim, K.H., R. Brent Gillespie, and B.J. Martin. 2007. Head movement control in visually guided tasks: Postural goal and optimality. *Computers in Biology and Medicine* 37(7):1009–1019.

Kipp, Michael, and Patrick Gebhard. 2008. Igaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In *Intelligent virtual agents*, 191–199. Springer.

Kirchner, Nathan, Alen Alempijevic, and Gamini Dissanayake. 2011. Nonverbal robot-group interaction using an imitated gaze cue. In *Proceedings of the 6th international conference on human-robot interaction*, 497–504. ACM.

Kittur, A., E.H. Chi, and B. Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual sigchi conference on human factors in computing systems*, 453–456. ACM.

Kleinke, Chris L. 1986. Gaze and eye contact: a research review. *Psychological bulletin* 100(1):78–100.

Kokkinara, Elena, Oyewole Oyekoya, and Anthony Steed. 2011. Modelling selective visual attention for autonomous virtual characters. *Computer Animation and Virtual Worlds* 22(4): 361–369.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents*, 205–217. Springer.

Kousidis, Spyridon, and David Schlangen. 2015. The power of a glance: Evaluating embodiment and turn-tracking strategies of an active robotic overhearer. In *Proceedings of aaai spring symposium on turn-taking and coordination in human-machine interaction*.

Krenn, Brigitte, Catherine Pelachaud, Hannes Pirker, and Christopher Peters. 2011. Embodied conversational characters: Representation formats for multimodal communicative behaviours. In *Emotion-oriented systems*, 389–415. Springer.

Kuno, Yoshinori, Kazuhisa Sadazuka, Michie Kawashima, Keiichi Yamazaki, Akiko Yamazaki, and Hideaki Kuzuoka. 2007. Museum guide robot based on sociological interaction analysis. In *Proceedings of the sigchi conference on human factors in computing systems*, 1191–1194. ACM.

Lance, B.J., and S.C. Marsella. 2010a. The expressive gaze model: Using gaze to express emotion. *Computer Graphics and Applications, IEEE* 30(4):62–73.

Lance, Brent, and Stacy Marsella. 2010b. Glances, glares, and glowering: how should a virtual human express emotion through gaze? *Autonomous Agents and Multi-Agent Systems* 20(1):50–69.

Lance, Brent, Stacy Marsella, and David Koizumi. 2004. Towards expressive gaze manner in embodied virtual agents. In *Aamas workshop on empathic agents*, 194–201.

Lance, Brent, and Stacy C Marsella. 2007. Emotionally expressive head and body movement during gaze shifts. In *Intelligent virtual agents*, 72–85. Springer.

Land, Michael, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28(11):1311–1328.

Land, Michael F, and Mary Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision research* 41(25):3559–3565.

Langton, S.R.H., and V. Bruce. 1999. Reflexive visual orienting in response to the social attention of others. *Visual Cognition* 6(5):541–567.

Langton, Stephen RH, Roger J Watt, and Vicki Bruce. 2000. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences* 4(2):50–59.

Lee, J., S. Marsella, D. Traum, J. Gratch, and B. Lance. 2007. The rickel gaze model: A window on the mind of a virtual human. In *Intelligent virtual agents*, 296–303. Springer.

Lee, Kwan Min, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 2006a. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies* 64(10):962–973.

Lee, Kwan Min, and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the sigchi conference on human factors in computing systems*, 289–296. ACM.

Lee, Kwan Min, Wei Peng, Seung-A Jin, and Chang Yan. 2006b. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication* 56(4):754–772.

Lee, S.P., J.B. Badler, and N.I. Badler. 2002. Eyes alive. In *Acm transactions on graphics (tog)*, vol. 21, 637–644. ACM.

Leite, Iolanda, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of the 7th annual acm/ieee international conference on human-robot interaction*, 367–374. ACM.

Lester, J.C., S.G. Towns, C.B. Callaway, J.L. Voerman, and P.J. FitzGerald. 2000. Deictic and emotive communication in animated pedagogical agents. *Embodied conversational agents* 123–154.

Leyzberg, Daniel, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th annual conference of the cognitive science society (cogsci '12)*, 1–6. Citeseer.

Li, Zheng, and Xia Mao. 2012. Emotional eye movement generation based on geneva emotion wheel for virtual agents. *Journal of Visual Languages & Computing* 23(5):299–310.

Lippa, Richard A, and Joshua K Dietz. 2000. The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior* 24(1):25–43.

Liu, C., D.L. Kay, and J.Y. Chai. 2011. Awareness of partner's eye gaze in situated referential grounding: An empirical study. *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

Liu, Chaoran, Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Proceedings of the 7th acm/ieee international conference on human-robot interaction*, 285–292. IEEE.

Lohse, Manja, and Herwin van Welbergen. 2012. Designing appropriate feedback for virtual agents and robots. In *Position paper at ro-man 2012 workshop robot feedback in human-robot interaction: How to make a robot readable for a human interaction partner*.

Lusk, Mary Margaret, and Robert K Atkinson. 2007. Animated pedagogical agents: Does their degree of embodiment impact learning from static or animated worked examples? *Applied cognitive psychology* 21(6):747–764.

Ma, X., and Z. Deng. 2009. Natural eye motion synthesis by modeling gaze-head coupling. In *Proceedings of the 2009 ieee virtual reality conference*, 143–150. IEEE Computer Society.

Ma, X., B.H. Le, and Z. Deng. 2011. Perceptual analysis of talking avatar head movements: A quantitative perspective. *CHI'11*.

Macdonald, Ross G, and Benjamin W Tatler. 2013. Do as eye say: Gaze cueing and language in a real-world social interaction. *Journal of vision* 13(4):6–6.

Marti, Patrizia, Margherita Bacigalupo, Leonardo Giusti, Claudio Mennecozzi, and Takanori Shibata. 2006. Socially assistive robotics in the treatment of behavioural and psychological symptoms of dementia. In *Biomedical robotics and biomechatronics, 2006. biorob 2006. the first ieee/ras-embs international conference on*, 483–488. IEEE.

Mason, M.F., E.P. Tatkow, and C.N. Macrae. 2005. The look of love: Gaze shifts and person perception. *Psychological Science* 236–239.

Masuko, Soh, and Junichi Hoshino. 2007. Head-eye animation corresponding to a conversation for cg characters. In *Computer graphics forum*, vol. 26, 303–312. Wiley Online Library.

Mavridis, Nikolaos. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63:22–35.

McCluskey, M.K., and K.E. Cullen. 2007. Eye, head, and body coordination during large gaze shifts in rhesus monkeys: movement kinematics and the influence of posture. *Journal of neurophysiology* 97(4):2976.

Mehrabian, A. 1966. Immediacy: An indicator of attitudes in linguistic communication. *Journal of Personality* 34(1):26–34.

de Melo, C.M., P. Carnevale, and J. Gratch. 2011. The effect of expression of anger and happiness in computer agents on negotiations with humans. In *tenth international conference on autonomous agents and multiagent systems*.

Meltzoff, Andrew N, Rechele Brooks, Aaron P Shon, and Rajesh PN Rao. 2010. "social" robots are psychological agents for infants: A test of gaze following. *Neural Networks* 23(8): 966–972.

Meyer, Antje, Femke van der Meulen, and Adrian Brooks. 2004. Eye movements during speech planning: talking about present and remembered objects. *Visual Cognition* 11(5): 553–576.

Meyer, A.S., A.M. Sleiderink, and W.J.M. Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition* 66(2):B25–B33.

Mitake, Hironori, Shoichi Hasegawa, Yasuharu Koike, and Makoto Sato. 2007. Reactive virtual human with bottom-up and top-down visual attention for gaze generation in realtime interactions. In *Virtual reality conference, 2007. vr'07. ieee*, 211–214. IEEE.

Moon, AJung, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zeng, 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*, 334–341. ACM.

Moon, Youngme. 2000. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research* 26(4):323–339.

Moore, Chris, and Phil Dunham. 1995. *Joint attention: Its origins and role in development*. Lawrence Erlbaum Associates.

Morency, Louis-Philippe, and Trevor Darrell. 2007. Conditional sequence model for context-based recognition of gaze aversion. In *Machine learning for multimodal interaction*, 11–23. Springer.

Mumm, J., and B. Mutlu. 2011a. Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior*.

Mumm, Jonathan, and Bilge Mutlu. 2011b. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction*, 331–338. ACM.

Mutlu, B., J. Forlizzi, and J. 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, B., T. Shiwa, T. Kanda, H. Ishiguro, and N. 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*, 61–68. ACM.

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, 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.

Nakano, Tamami, Makoto Kato, Yusuke Morito, Seishi Itoi, and Shigeru Kitazawa. 2013a. Blink-related momentary activation of the default mode network while viewing videos. *Proceedings of the National Academy of Sciences* 110(2):702–706.

Nakano, Tamami, and Shigeru Kitazawa. 2010. Eyeblink entrainment at breakpoints of speech. *Experimental brain research* 205(4):577–581.

Nakano, Yukiko I, Naoya Baba, Hung-Hsuan Huang, and Yuki Hayashi. 2013b. Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In *Proceedings of the 15th acm on international conference on multimodal interaction*, 35–42. ACM.

Nakano, Yukiko I, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 1*, 553–561. Association for Computational Linguistics.

Nass, Clifford, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. 2005. Improving automotive safety by pairing driver emotion and car voice emotion. In *Chi'05 extended abstracts on human factors in computing systems*, 1973–1976. ACM.

Neider, Mark B, Xin Chen, Christopher A Dickinson, Susan E Brennan, and Gregory J Zelinsky. 2010. Coordinating spatial referencing using shared gaze. *Psychonomic bulletin & review* 17(5):718–724.

Nijholt, A., D. Heylen, and R. Vertegaal. 2000. Inhabited interfaces: Attentive conversational agents that help. In *Proceedings 3rd international conference on disability, virtual reality and associated technologies-cdvrat2000, alghero, sardinia*.

Nikolopoulos, Chris, Deitra Kuester, Mark Sheehan, Shashwati Ramteke, Aniket Karmarkar, Supriya Thota, Joseph Kearney, Curtis Boirum, Sunnihith Bojedla, and Angela Lee. 2011. Robotic agents used to help teach social skills to children with autism: the third generation. In *Ro-man, 2011 ieee*, 253–258. IEEE.

Normoyle, Aline, Jeremy B Badler, Teresa Fan, Norman I Badler, Vinicius J Cassol, and Soraia R Musse. 2013. Evaluating perceived trust from procedurally animated gaze. In *Proceedings of motion on games*, 141–148. ACM.

Novick, David G, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *Spoken language, 1996. icslp 96. proceedings., fourth international conference on*, vol. 3, 1888–1891. IEEE.

Obaid, Mohammad, Ionut Damian, Felix Kistler, Birgit Endrass, Johannes Wagner, and Elisabeth André. 2012. Cultural behaviors of virtual agents in an augmented reality environment. In *Intelligent virtual agents*, 412–418. Springer.

Okumura, Yuko, Yasuhiro Kanakogi, Takayuki Kanda, Hiroshi Ishiguro, and Shoji Itakura. 2013. Infants understand the referential nature of human gaze but not robot gaze. *Journal of experimental child psychology* 116(1):86–95.

Orrill, Chandra Hawley, and David Williamson Shaffer. 2012. Exploring connectedness: applying ena to teacher knowledge. In *International conference of the learning sciences (icls)*.

Osterberg, Lars, and Terrence Blaschke. 2005. Adherence to medication. *New England Journal of Medicine* 353(5):487–497.

Otteson, J.P., and C.R. Otteson. 1980. Effect of teacher's gaze on children's story recall. *Perceptual and Motor Skills*.

Oyekoya, Oyewole, Anthony Steed, and Xueni Pan. 2011. Exploring the object relevance of a gaze animation model. In *Proceedings of the 17th eurographics conference on virtual environments & third joint virtual reality*, 111–114. Eurographics Association.

Oyekoya, Oyewole, William Steptoe, and Anthony Steed. 2009. A saliency-based method of simulating visual attention in virtual scenes. In *Proceedings of the 16th acm symposium on virtual reality software and technology*, 199–206. ACM.

Parke, F.I., and K. Waters. 2008. *Computer facial animation*. AK Peters Ltd.

Pejsa, Tomislav, Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2015. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5(1):3:1–3:34.

Pejsa, Tomislav, Bilge Mutlu, and Michael Gleicher. 2013. Stylized and performative gaze for character animation. In *Computer graphics forum*, vol. 32, 143–152. Wiley Online Library.

Pelachaud, C., and M. Bilvi. 2003. Modelling gaze behavior for conversational agents. In *Intelligent virtual agents*, 93–100. Springer.

Pelz, J., M. Hayhoe, and R. Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139(3):266–277.

Perlin, Ken. 1995. Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics* 1(1):5–15.

Peters, C. 2010. Animating gaze shifts for virtual characters based on head movement propensity. In *2010 second international conference on games and virtual worlds for serious applications*, 11–18. IEEE.

Peters, Christopher, Stylianos Asteriadis, and Kostas Karpouzis. 2010. Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces* 3(1-2):119–130.

Peters, Pamela. 2007. Gaining compliance through non-verbal communication. *Pepp. Disp. Resol. LJ* 7:87.

Phelps, Fiona G, Gwyneth Doherty-Sneddon, and Hannah Warnock. 2006. Helping children think: Gaze aversion and teaching. *British journal of developmental psychology* 24(3):577–588.

Picot, Antoine, Gérard Bailly, Frédéric Elisei, and Stephan Raidt. 2007. Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In *Intelligent virtual agents*, 272–282. Springer.

Pitsch, Karola, Anna-Lisa Vollmer, and Manuel Mühlig. 2013. Robot feedback shapes the tutor's presentation: How a robot's online gaze strategies lead to micro-adaptation of the human's conduct. *Interaction Studies* 14(2):268–296.

Poggi, I., C. Pelachaud, and F. De Rosis. 2000. Eye communication in a conversational 3d synthetic agent. *AI communications* 13(3):169–181.

Pourtois, Gilles, David Sander, Michael Andres, Didier Grandjean, Lionel Reveret, Etienne Olivier, and Patrik Vuilleumier. 2004. Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals. *European Journal of Neuroscience* 20(12):3507–3515.

Powers, Aaron, and Sara Kiesler. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction*, 218–225. ACM.

Powers, Aaron, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Proc. hri '07*, 145–152.

Queiroz, Rossana B, Leandro M Barros, and Soraia R Musse. 2007. Automatic generation of expressive gaze in virtual animated characters: From artists craft to a behavioral animation model. In *Intelligent virtual agents*, 401–402. Springer.

———. 2008. Providing expressive gaze to virtual animated characters in interactive applications. *Computers in Entertainment (CIE)* 6(3):41.

Rauthmann, John F, Christian T Seubert, Pierre Sachse, and Marco R Furtner. 2012. Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality* 46(2):147–156.

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.

Richardson, Daniel C, Rick Dale, and Natasha Z Kirkham. 2007. The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological science* 18(5):407–413.

Richardson, Daniel C, Rick Dale, and John M Tomlinson. 2009. Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science* 33(8):1468–1482.

Richter, Laurie A, and Gavriel Salvendy. 1995. Effects of personality and task strength on performance in computerized tasks. *Ergonomics* 38(2):281–291.

Rickel, J., and W.L. Johnson. 2000. Task-oriented collaboration with embodied agents in virtual worlds. *Embodied conversational agents* 95–122.

Rickel, Jeff, and W Lewis Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied artificial intelligence* 13(4-5):343–382.

Ruhland, Kerstin, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. 2015. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In *Computer graphics forum*, vol. 34, 299–326. Wiley Online Library.

Rupp, André A, Younyoung Choi, Matthew Gushta, RJ Mislevy, E Bagley, P Nash, D Hatfield, G Svarowski, and D Shaffer. 2009. Modeling learning progressions in epistemic games with epistemic network analysis: Principles for data analysis and generation. In *Proceedings from the learning progressions in science conference*, 24–26.

Rupp, André A, Matthew Gushta, Robert J Mislevy, and David Williamson Shaffer. 2010. Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology, Learning and Assessment* 8(4).

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.

Ryan, Richard M, and Edward L Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* 25(1):54–67.

Sacks, Harvey, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 696–735.

Sakamoto, Daisuke, Takayuki Kanda, Tetsuo Ono, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Android as a telecommunication medium with a human-like presence. In *Human-robot interaction (hri), 2007 2nd acm/ieee international conference on*, 193–200. IEEE.

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 2004). proceedings. 2004 ieee/rsj international conference on*, vol. 1, 846–851. IEEE.

Salvucci, Dario D, and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications*, 71–78. ACM.

Sauppé, Allison, and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction*, 342–349. ACM.

Scassellati, Brian, Henny Admoni, and Maja Mataric. 2012. Robots for use in autism research. *Annual review of biomedical engineering* 14:275–294.

Schober, Michael F. 1993. Spatial perspective-taking in conversation. *Cognition* 47(1):1–24.

Schulman, Daniel, and Timothy Bickmore. 2012. Changes in verbal and nonverbal conversational behavior in long-term interaction. In *Proceedings of the 14th acm international conference on multimodal interaction*, 11–18. ACM.

Sebanz, Natalie, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in cognitive sciences* 10(2):70–76.

Segrin, Chris. 1993. The effects of nonverbal behavior on outcomes of compliance gaining attempts. *Communication Studies* 44(3-4):169–187.

Shaffer, David Williamson, David Hatfield, Gina Navoa Svarovsky, Padraig Nash, Aran Nulty, Elizabeth Bagley, Ken Frank, André A Rupp, and Robert Mislevy. 2009. Epistemic network analysis: A prototype for 21st-century assessment of learning.

Shah, Julie, and Cynthia Breazeal. 2010. An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52(2):234–245.

Shao, Wei, and Demetri Terzopoulos. 2005. Autonomous pedestrians. In *Proceedings of the 2005 acm siggraph/eurographics symposium on computer animation*, 19–28. ACM.

Sherwood, J.V. 1987. Facilitative effects of gaze upon learning. *Perceptual and Motor Skills*.

Shiwa, Toshiyuki, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2008. How quickly should communication robots respond? In *Proceedings of the 3rd acm/ieee international conference on human robot interaction*, 153–160. IEEE.

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.

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.

Skantze, Gabriel, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication* 65:50–66.

Skantze, Gabriel, Anna Hjalmarsson, and Catherine Oertel. 2013. Exploring the effects of gaze and pauses in situated human-robot interaction. In *Proceedings of the 14th annual meeting of the special interest group on discourse and dialogue (sigdial '13)*, 375–383.

Skotte, JH, Jacob Klenø Nøjgaard, LV Jørgensen, KB Christensen, and G Sjøgaard. 2007. Eye blink frequency during different computer tasks quantified by electrooculography. *European journal of applied physiology* 99(2):113–119.

Sorostinean, Mihaela, François Ferland, Adriana Tapus, et al. 2014. Motion-oriented attention for a social gaze robot behavior. In *Social robotics*, 310–319. Springer.

Srinivasan, Vasant, and Robin R Murphy. 2011. A survey of social gaze. In *Human-robot interaction (hri), 2011 6th acm/ieee international conference on*, 253–254. IEEE.

Staudte, Maria, and Matthew W. Crocker. 2009. Visual attention in spoken human-robot interaction. In *Proc. hri*, 77–84.

Staudte, Maria, and Matthew W Crocker. 2011. Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition* 120(2):268–291.

Steptoe, W., and A. Steed. 2008. High-fidelity avatar eye-representation. In *Virtual reality conference, 2008. vr'08. ieee*, 111–114. IEEE.

Steptoe, William, Oyewole Oyekoya, and Anthony Steed. 2010. Eyelid kinematics for virtual characters. *Computer animation and virtual worlds* 21(3-4):161–171.

Stern, John A, Larry C Walrath, and Robert Goldstein. 1984. The endogenous eyeblink. *Psychophysiology* 21(1):22–33.

Strabala, Kyle, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S Srinivasa. 2012. Learning the communication of intent prior to physical collaboration. In *Ro-man 2012*, 968–973. IEEE.

Strabala, Kyle Wayne, Min Kyung Lee, Anca Diana Dragan, Jodi Lee Forlizzi, Siddhartha Srinivasa, Maya Cakmak, and Vincenzo Micelli. 2013. Towards seamless human-robot handovers. *Journal of Human-Robot Interaction* 2(1):112–132.

Takacs, Barnabas, and David Hanak. 2008. A prototype home robot with an ambient facial interface to improve drug compliance. *Journal of telemedicine and telecare* 14(7):393–395.

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.

Tanenhaus, Michael K, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217):1632–1634.

Tapus, Adriana, and Maja J Mataric. 2008. Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *Aaai spring symposium: Emotion, personality, and social behavior*, 133–140.

Tapus, Adriana, and Maja J Matarić. 2008. User personality matching with a hands-off robot for post-stroke rehabilitation therapy. In *Experimental robotics*, 165–175. Springer.

Tapus, Adriana, Andreea Peca, Amir Aly, Cristina Pop, Lavinia Jisa, Sebastian Pintea, Alina S Rusu, and Daniel O David. 2012. Children with autism social engagement in interaction with nao, an imitative robot–a series of single case experiments. *Interaction studies* 13(3):315–347.

Tapus, Adriana, Cristian Tapus, and Maja Mataric. 2009a. The role of physical embodiment of a therapist robot for individuals with cognitive impairments. In *Proceedings of the 18th ieee international symposium on robot and human interactive communication*, 103–107. IEEE.

Tapus, Adriana, Cristian Tapus, and Maja J Matarić. 2007. Hands-off therapist robot behavior adaptation to user personality for post-stroke rehabilitation therapy. In *Robotics and automation, 2007 ieee international conference on*, 1547–1553. IEEE.

Tapus, Adriana, Cristian Tapus, and Maja J Mataric. 2009b. The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In *Proceedings of the ieee international conference on rehabilitation robotics*, 924–929. IEEE.

Tartaro, A., and J. Cassell. 2006. Authorable virtual peers for autism spectrum disorders. In *Proceedings of the combined workshop on language-enabled educational technology and development and evaluation for robust spoken dialogue systems at the 17th european conference on artificial intellegence*. Citeseer.

Tartaro, Andrea, and Justine Cassell. 2007. Using virtual peer technology as an intervention for children with autism. *Towards universal usability: designing computer interfaces for diverse user populations. Chichester: John Wiley* 231:62.

Thomas, F., and O. Johnston. 1981. *Disney animation: The illusion of life*, vol. 1. Abbeville Press New York.

Thórisson, K.R. 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in language and speech systems* 173–207.

Thórisson, Kristinn R. 1994. Face-to-face communication with computer agents. In *Aaai spring symposium on believable agents working notes*, vol. 86, 90.

Torrey, Cristen, Aaron Powers, Susan R Fussell, and Sara Kiesler. 2007. Exploring adaptive dialogue based on a robot's awareness of human gaze and task progress. In *Proceedings of the acm/ieee international conference on human-robot interaction*, 247–254. ACM.

Torrey, Cristen, Aaron Powers, Matthew Marge, Susan R Fussell, and Sara Kiesler. 2006. Effects of adaptive robot dialogue on information exchange and social relations. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction*, 126–133. ACM.

Trutoiu, Laura C, Elizabeth J Carter, Iain Matthews, and Jessica K Hodgins. 2011. Modeling and animating eye blinks. *ACM Transactions on Applied Perception (TAP)* 8(3):17.

Vala, Marco, Gabriel Blanco, and Ana Paiva. 2011. Providing gender to embodied conversational agents. In *Intelligent virtual agents*, 148–154. Springer.

Vallerand, Richard J, and Greg Reid. 1984. On the causal effects of perceived competence on intrinsic motivation: A test of cognitive evaluation theory. *Journal of Sport Psychology* 6(1):94–102.

VanderWerf, Frans, Peter Brassinga, Dik Reits, Majid Aramideh, and Bram Ongerboer de Visser. 2003. Eyelid movements: behavioral studies of blinking in humans under different stimulus conditions. *Journal of neurophysiology* 89(5):2784–2796.

Vázquez, Marynel, Aaron Steinfeld, Scott E Hudson, and Jodi Forlizzi. 2014. Spatial and other social engagement cues in a child-robot interaction: effects of a sidekick. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction*, 391–398. ACM.

Vertegaal, R. 1999. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the sigchi conference on human factors in computing systems: the chi is the limit*, 294–301. ACM.

Vertegaal, Roel, Robert Slagter, Gerrit Van der Veer, and Anton Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the sigchi conference on human factors in computing systems*, 301–308. ACM.

Vilhjálmsson, Hannes, Nathan Cantelmo, Justine Cassell, Nicolas E Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, et al. 2007. The behavior markup language: Recent developments and challenges. In *Intelligent virtual agents*, 99–111. Springer.

Vilhjálmsson, Hannes Högni, and Justine Cassell. 1998. Bodychat: Autonomous communicative behaviors in avatars. In *Proceedings of the second international conference on autonomous agents*, 269–276. ACM.

Wade, Eric, Avinash Parnandi, Ross Mead, and Maja Matarić. 2011. Socially assistive robotics for guiding motor task practice. *Paladyn* 2(4):218–227.

Wainer, Joshua, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2007. Embodiment and human-robot interaction: A task-based perspective. In *Ro-man 2007-the 16th ieee international symposium on robot and human interactive communication*, 872–877. IEEE.

Wang, N., and J. Gratch. 2010. Don't just stare at me! In *Proceedings of the 28th international conference on human factors in computing systems*, 1241–1250. ACM.

Wasserman, Stanley, and Katherine Faust. 1994. *Social network analysis: Methods and applications*, vol. 8. Cambridge university press.

Wilson, Thomas P, and Don H Zimmerman. 1986. The structure of silence between turns in two-party conversation. *Discourse Processes* 9(4):375–390.

Yamazaki, Akiko, Keiichi Yamazaki, Matthew Burdelski, Yoshinori Kuno, and Mihoko Fukushima. 2010. Coordination of verbal and non-verbal actions in human-robot interaction at museums and exhibitions. *Journal of Pragmatics* 42(9):2398–2414.

Yamazaki, Keiichi, Michie Kawashima, Yoshinori Kuno, Naonori Akiya, Matthew Burdelski, Akiko Yamazaki, and Hideaki Kuzuoka. 2007. Prior-to-request and request behaviors within elderly day care: Implications for developing service robots for use in multiparty settings. In *Ecscw 2007*, 61–78. Springer.

Yi, Weilie, and Dana Ballard. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics* 6(03):337–359.

Yngve, Victor H. 1970. On getting a word in edgewise. In *Chicago linguistics society, 6th meeting*, 567–578.

Yonezawa, Tomoko, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the 9th international conference on multimodal interfaces*, 140–145. ACM.

Yoshikawa, Y., K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. 2006. Responsive robot gaze to interaction partner. In *Proc. rss*.

Yu, Chen, Paul Schermerhorn, and Matthias Scheutz. 2012. Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1(2):13.

Zahn, Christopher J. 1984. A reexamination of conversational repair. *Communications Monographs* 51(1):56–66.

Zangemeister, WH, and L. Stark. 1982. Types of gaze movement: variable interactions of eye and head movements. *Experimental Neurology* 77(3):563–577.

Zaraki, Abolfazl, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi. 2014. Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems* 44(2):157–168.

Zbilut, Joseph P, Alessandro Giuliani, and Charles L Webber. 1998. Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A* 246(1):122–128.

Zheng, Minhua, AJung Moon, Elizabeth A Croft, and Max Q-H Meng. 2015. Impacts of robot head gaze on robot-to-human handovers. *International Journal of Social Robotics* 7(5): 783–798.

Zoric, Goranka, Rober Forchheimer, and Igor S Pandzic. 2011. On creating multimodal virtual humans—real time speech driven facial gesturing. *Multimedia Tools and Applications* 54(1):165–179.