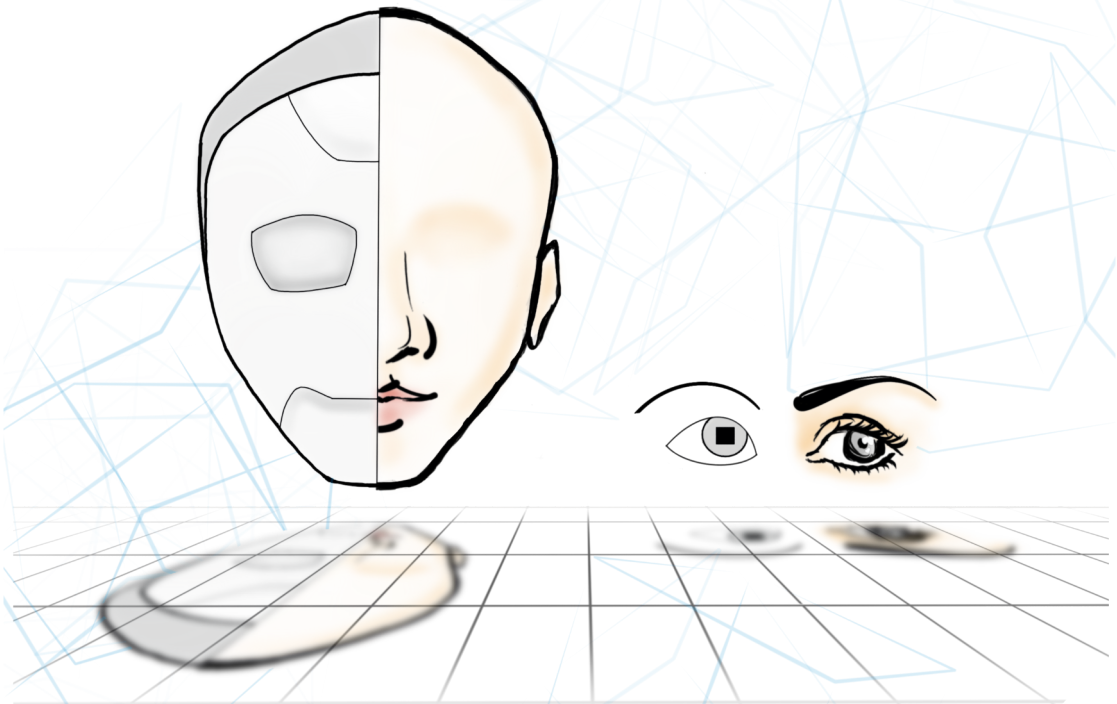
MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

The Face Says it All:

Investigating Gaze and Affective Behaviors of Social Robots



CHINMAYA MISHRA

The Face Says it All:

Investigating Gaze and Affective Behaviors of Social Robots

Chinmaya Mishra

Funding body
This research has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Grant Agreement No. 859588.

Conversational Brains (COBRA)
The educational component of the doctoral training was provided by the Conversational Brains (COBRA) consortium. COBRA is an EU funded Marie Skłodowska-Curie Innovative Training Network (COBRA ITN). The network aims to train the next generation of researchers to accurately characterize and model the linguistic, cognitive, and brain mechanisms that allow conversation to unfold in both human-human and human-machine interactions. The network includes 10 world-level academic research centers on language, cognition, and the human brain as well as 4 non-academic partners that include fast-developing SMEs and one world-level company. The partners' unique combined expertise and high complementarity will allow COBRA to offer Early-stage Researchers an excellent training program as well as very strong exposure to the non-academic sector. More information can be found at the *COBRA Website*.

The MPI series in Psycholinguistics
Initiated in 1997, the MPI series in Psycholinguistics contains doctoral theses produced at the Max Planck Institute for Psycholinguistics. Since 2013, it includes theses produced by members of the IMPRS for Language Sciences. The current listing is available at www.mpi.nl/mpi-series

ISBN: 978-94-92910-56-1
Cover design by TANMAYA MISHRA
Printed and bound by Ipskamp Drukkers, Enschede

 The research reported in this thesis was conducted at the Max Planck Institute for Psycholinguistics, in Nijmegen, the Netherlands, Furhat Robotics AB, Stockholm, Sweden and KTH Royal Institute of Technology, Stockholm, Sweden.

The Face Says it All:

Investigating Gaze and Affective Behaviors of Social Robots

Proefschrift ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de rector magnificus prof. dr. J.M. Sanders,

volgens besluit van het college voor promoties

in het openbaar te verdedigen op

woensdag 17 april 2024

om 16.30 uur precies

door

Chinmaya Mishra

geboren op 18 november 1989

te Burla (India)

Promotoren:

Prof. dr. P. Hagoort

Prof. dr. G. Skantze (Kungliga Tekniska Högskolan, Zweden)

Copromotoren:

Dr. S. Fuchs (Leibnis-Zentrum Allgemeine Sprachwissenschaft (ZAS), Duitsland)

Dr. R.G. Verdonschot (Max Planck Institute for Psycholinguistics)

Manuscriptcommissie:

Prof. dr. H. Bekkering

Dr. J.A.M. Holler

Prof. dr. G. Bailly (Université Grenoble-Alpes, Frankrijk)

The Face Says it All:
Investigating Gaze and Affective Behaviors of Social Robots

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Wednesday, April 17, 2024
at 4.30 pm

by
Chinmaya Mishra
born on November 18, 1989
in Burla (India)

Supervisors:

Prof. dr. P. Hagoort

Prof. dr. G. Skantze (KTH Royal Institute of Technology, Sweden)

Co-supervisors:

Dr. S. Fuchs (Leibniz-Centre General Linguistics (ZAS), Germany)

Dr. R.G. Verdonschot (Max Planck Institute for Psycholinguistics)

Manuscript Committee:

Prof. dr. H. Bekkering

Dr. J.A.M. Holler

Prof. dr. G. Bailly (Université Grenoble-Alpes, France)

Contents

1 | General introduction

In our interactions with others, we employ both verbal and non-verbal signals to convey our thoughts and feelings. Non-verbal signals/cues include gestures, body postures, facial expressions, eye contact, voice modulation, spatial proximity, and many other aspects. At times, non-verbal behaviors can take precedence over verbal language and communication can be exclusively non-verbal. For instance, a simple wave of the hand can signify a greeting, and a nod of the head can express agreement. The human face, in particular, plays a central role in our non-verbal communication. It is a source of rich visual cues, including facial expressions and gaze. We evolved to communicate non-verbally long before we learned to use language. This is evident in the existence of gestures that are recognized to have the same meaning across cultures and languages (Andersen, 1999). Early research on non-verbal communication emphasized its significance, famously suggesting that 90% of communication is non-verbal according to Mehrabian (1972). While this claim is exaggerated (questioned in later studies (Burgoon, 1985; Lapakko, 1997)) and the exact percentage values are very difficult to quantify, the underlying message remains clear: non-verbal behavior is integral for effective communication.

With the rapid advancements in artificial intelligence and robotics technologies, social robots are poised to have greater integration in society. These robots are designed specifically to conduct human-like interactions. So, understanding and replicating essential non-verbal cues, such as facial expressions and gaze, are essential for enhancing the effectiveness, human-likeness, and acceptance of these robotic systems. Social robots are already being employed in a variety of domains, including healthcare, education, and assistive roles, where their capacity to convey and interpret human emotions and intentions can significantly impact the quality of interactions. Modeling non-verbal behaviors on these robots would make them more capable of providing a richer user experience. For example, a social robot designed to provide companionship to the elderly could express happiness when the user is cheerful and sadness when the user is upset, enhancing emotional connection, or look directly at the user when speaking, creating a more personalized and attentive interaction.

My main goal is to investigate methods for making human-robot interactions (HRI) more seamless and human-like by modeling non-verbal behaviors on social robots. My research centered on two key areas:

- Determining optimal methods for modeling these behaviors.
- Evaluating the human perception and influence of these behaviors during HRI

This dissertation specifically focuses on modeling the eye gaze and affective behaviors of social robots. Subsequent sections offer background information on the significance of gaze and affective behavior in human-human interactions (HHI) (Section 1.1 and 1.3), an examination of existing models along with their limitations, and the resulting research questions in Sections 1.2 and 1.4. The robotic platform used in this dissertation is introduced and discussed briefly in Section 1.5. Section 1.6 outlines the structure of this dissertation and explains how the research inquiries were addressed.

1.1 Role of Eye Gaze in Communication

Gaze cues are regarded to be especially vital among non-verbal signals in HHI (Kendon, 1967). A prior investigation revealed that the human brain possesses unique, hard-wired pathways for interpreting these cues (Emery, 2000). The manner in which we focus on an object, the timing of shifting our gaze away from an interlocutor, and the duration of our gaze directed at our partner hold distinct functions in human interactions. Studies on human communication shed light on the multifaceted functions of gaze in social exchanges, encompassing conversation floor regulation (Rossano, 2012) and intimacy management (Abele, 1986). In this section, I briefly discuss several roles of gaze cues in HHI which are crucial topics, investigated in this dissertation.

Turn-taking is the process through which conversational participants coordinate and alternate their speaking roles. This coordination involves a range of non-verbal cues involving the voice and facial expressions, with gaze being a critical component (Kendrick, Holler, & Levinson, 2023; Skantze, 2021). Typically, when someone is speaking, they tend to avert their gaze from the listener, particularly during lengthy speech segments and at the outset of their turn, indicating their intention to hold the conversational floor (Kendon, 1967). In contrast, when it's time to relinquish the floor, the speaker often initiates a mutual gaze toward the conclusion of their utterance. According to Jokinen, Furukawa,

Nishida, and Yamamoto (2013), such eye gaze is crucial in determining when the speaker intends to continue speaking.

Gaze aversion entails the purposeful redirection of one's gaze away from their conversational partner. It serves three primary functions: cognitive, intimacy regulation, and facilitating turn-taking (Andrist, Tan, Gleicher, & Mutlu, 2014). Throughout a conversation, speakers frequently look away from the listener, aiding in the planning of their next statement and minimizing distractions (Argyle & Cook, 1976). Research has revealed that maintaining mutual gaze can lead to increased hesitations and false starts (Beattie, 1981). Moreover, gaze aversion is an essential cue for signaling one's intent to retain the conversational floor. Additionally, gaze aversion significantly contributes to regulating the level of intimacy within a conversation, as demonstrated by Abele (1986).

The capacity to track the direction of someone else's gaze plays a pivotal role in synchronizing attention during social interactions. When two or more individuals simultaneously focus on a shared point of interest and are mutually aware of this focus, it is typically termed **joint attention**. This concept of joint attention can be further classified into two primary components, known as **responding to joint attention (RJA)** and **initiating joint attention (IJA)** as delineated by Mundy and Newell (2007). IJA denotes the situation in which one participant in the interaction takes the lead in establishing mutual eye contact and subsequently directs their gaze toward the referential point of interest. This act of initiating joint attention is often referred to as "referential gaze," as it involves guiding the attention of the interaction partner to a specific target. On the other hand, RJA encompasses the action of observing and tracking the gaze direction of others, while also discerning the implicit signal to collectively direct attention toward a shared focal point. In essence, RJA involves the ability to follow and understand someone else's gaze cues and the underlying intention to engage in joint attention.

1.2 Gaze in HRI

In light of the substantial role that gaze behavior plays in human communication, it is reasonable to consider reproducing the same in HRI. A significant challenge in HRI pertains to effectively communicating a robot's intent to its human counterparts. Achieving this necessitates the emulation of human behavior to the greatest extent possible, which would facilitate easier perception and interpretation of a robot's cues. Consider a scenario where three cups are placed on a

table, and the robot is tasked with referring to a specific cup. Utilizing solely verbal references renders it challenging to disambiguate the exact cup to which the robot is referring. Conversely, when the robot supplements the verbal reference with a synchronized gaze directed at the cup, the identification of the referred cup becomes considerably more straightforward.

Efforts have been made by researchers to harness insights from HHI literature to develop models that govern and automate the gaze behavior of robots, denoted as Gaze Control Systems (GCS). These GCS primarily adopt two key approaches: data-driven (Andrist et al., 2014; Mutlu, Kanda, Forlizzi, Hodgins, & Ishiguro, 2012) and heuristic (Boucher et al., 2012; Mehlmann et al., 2014; Pereira, Oertel, Fermoselle, Mendelson, & Gustafson, 2019). In the data-driven approach, data obtained from HHI experiments is harnessed to train models, including deep learning models, to predict the gaze behavior for robots. Conversely, the heuristic approach involves the analysis of HHI data to formulate generalized rules governing robotic gaze behavior. While both of these approaches have reasonably succeeded in modeling human-like gaze behaviors in robots, several limitations persist.

The first limitation pertains to the lack of suitable robotic platforms equipped to replicate nuanced human-like gaze behaviors. Many existing GCS have been developed for robots like NAO[1] and Pepper[2], both of which lack independent eye and head movement capabilities. Consequently, these GCS are confined to modeling gaze behavior solely through head movements, overlooking the intricate coordination between eye and head movements that human gaze behavior entails when directing attention towards a target, as elucidated by Uemura, Arai, and Shimazaki (1980) and Stahl (1999). Additional limitations encompass the predominantly *reactive* and *static* nature of GCS. Typically, these models primarily focus on directing a robot's gaze in immediate response to events occurring during interactions, often with fixed gaze durations. However, it is known that HHI involves a lot of planning and the resulting gaze behavior is a combination of both *reactive* and *planning* components. A study by Beattie (2010) showed that gaze behavior is highly dependent on the underlying speech plan during HHI. This leads to my first research question:

- **RQ1**: *How can a comprehensive GCS be modeled for social robots?*

The other aspect that I want to focus on is the perceptual aspects of gaze behaviors exhibited by social robots and their potential impact on human behavior

[1] NAO Website
[2] Pepper website

during HRI. Are robot gaze behaviors perceived similarly to human gaze behavior? During an interaction, do robot gaze behaviors exert similar influence on human interlocutors, as human gaze behaviors? Addressing these questions is necessary as they lie at the center of designing better HRI interfaces. Whether or not we need to model human-like gaze behaviors is dependent on the answers to these questions.

Previous research has shown that human interlocutors perceive and recognize robot gaze behaviors to be intentional (Andrist et al., 2014). This is an important finding because it highlights that humans are ascribing intent to a robot's behavior and trying to interpret it. Yamazaki et al. (2008) conducted an experiment where the robot directed its gaze towards the participants at *transition relevant places*, resulting in observable influences on the participants' non-verbal behavior. Another study revealed that directing a robot's gaze when participants were being deceptive resulted in participants becoming more honest in subsequent trials (Schellen, Bossi, & Wykowska, 2021). Others have explored the impact of robot gaze on group activities, where the robot's gaze can be used to steer the conversational dynamics and improve participation equality (Gillet et al., 2021; Skantze, 2017). However, the direct influence of human gaze behavior as a result of robot gaze behavior has remained less explored. Moreover, many studies have used head motion instead of eye gaze to model robot gaze behavior due to constraints imposed by the robotic platforms in use. This limitation restricts the capacity to investigate the influence of nuanced, human-like gaze behaviors. This leads to my second research question:

- **RQ2**: *What influence does a robot's gaze have on human gaze behavior?*

1.3 Emotions in Communication

Facial expressions serve a dual purpose in human communication, functioning as a means of conveying one's emotions while also enabling the interpretation of the emotional states of others, thereby enhancing the effectiveness of interpersonal communication. This capacity to perceive and express emotions, collectively known as affective behavior, constitutes a fundamental aspect of human interaction. In our everyday interactions, the natural expression of emotions is a cornerstone for building and nurturing relationships with others (Lazarus, 2006). Facial expressions are also used to convey various meanings during interactions (Elliott & Jacobs, 2013). Elfenbein and Ambady (2002) conducted a study that underscored the universality of emotional recognition, highlighting

the significance of emotional expressions in human interactions. Furthermore, research has revealed the pivotal role that emotions play in the decision-making process (So et al., 2015).

Emotion Theories define emotions while discussing the similarities and differences between them. There are two main models of emotion that are proposed in emotion theories: *Categorical* and *Dimensional*. *Categorical* models propose that emotions elicited as a response to certain stimuli are divided into specific emotion categories such as the six basic emotions proposed by Ekman (1999). *Dimensional* models, on the other hand, suggest that the emotions experienced are influenced by fundamental dimensions, such as arousal and valence. Russell (1980) proposed the famous *Circumplex model* of emotions where the emotion categories were represented in a 2-D space, with transitions between emotions governed by variations in arousal and valence in response to stimuli (as shown in Fig. 1.1).

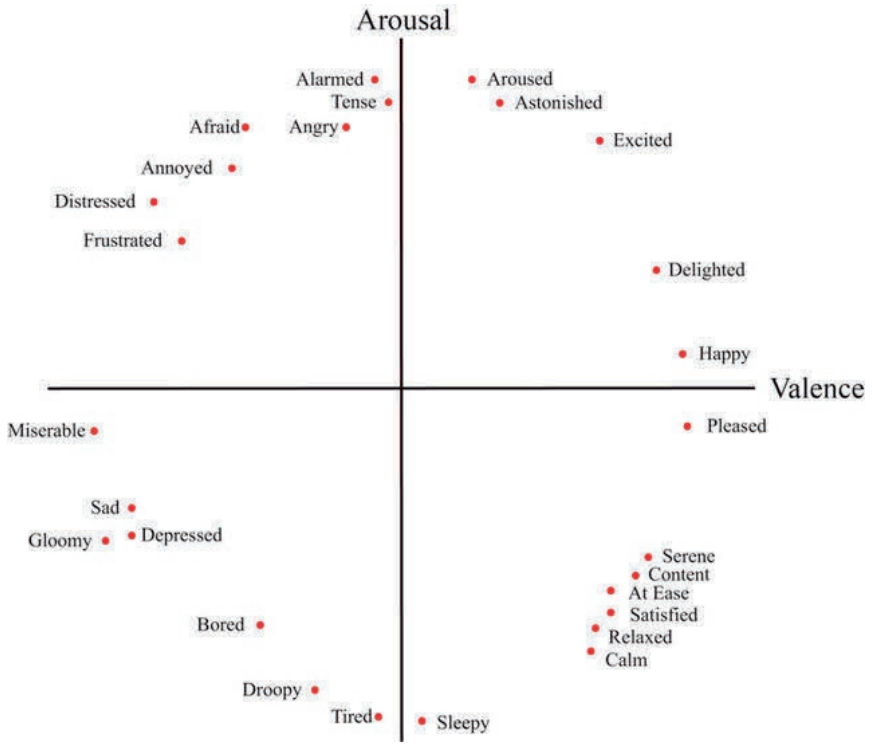


Figure 1.1: The Circumplex Model, which distributes the emotions along two dimensions; Valence and Arousal. Based on the Arousal and Valence levels in response to a stimulus, a specific emotional state can be reached. *Image adapted from* Seo and Huh (2019)

To gain a deeper comprehension of affective behavior, it is imperative to explore the factors that contribute to the generation of emotions during interactions. *Appraisal Theory*, a subset of emotion theories, seeks to elucidate this process by asserting that emotions are elicited through the appraisal of various factors, including the present situation, desired goals, and agency. According to Moors (2020),

"*Appraisal theory of emotion proposes that emotions or emotional components are caused and differentiated by an appraisal of the stimulus as mis/matching with goals and expectations, as easy/difficult to control, and as caused by others, themselves, or impersonal circumstances.*"

In essence, emotion models provide a quantitative framework for the representation and analysis of emotions, while appraisal theories offer insights into the determinants of emotional states and the underlying processes involved in their emergence.

1.4 Affective HRI

To enable robots to display affective behavior, it is essential to have the capacity to perceive various communicative signals (expressions, body posture, gaze direction, speech, spatial proximity, etc.) from human interlocutors. They must appraise these signals and subsequently exhibit an appropriate emotional response. Researchers have drawn inspiration from various models of emotions (as expounded in Section 1.3) as foundational frameworks for the emotion appraisal and emotion generation process in robots (Breazeal, 2003; Cully, Clune, Tarapore, & Mouret, 2015; Kirby, Forlizzi, & Simmons, 2010; Paplu, Mishra, & Berns, 2022; Tang et al., 2023). Breazeal (2003) introduced a framework that interprets interaction events in terms of arousal, valence, and stance dimensions, effectively mapping emotion categories in the affect space to determine the robot's appropriate emotional response. Similarly, Paplu et al. (2022) leveraged the *Circumplex model* of emotions to generate robot emotions. While these models exhibit the potential to approximate human-like emotion appraisal and generate contextually appropriate affective behaviors in robots, their development necessitates extensive effort and time due to the complexity involved in designing these architectures. Moreover, the computational intensity of such architectures poses a challenge in generating real-time emotions in HRI.

The recent surge in Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and OPT (S. Zhang et al., 2022), have

significantly impacted the field of natural language understanding and generation. LLMs are trained on very large-scale datasets comprising both dialogues and publicly available web documents. These models have showcased remarkable capabilities in solving diverse tasks that extend beyond the training data through 'few-shot' or 'zero-shot' learning. For HRI, the capacity to design interactions with 'zero-shot' chatting using LLMs (Brown et al., 2020) is of particular importance, as it significantly simplifies interaction design. Recent studies utilizing GPT-3 have demonstrated that LLMs can be employed to recognize emotions in conversations (Lammerse, Hassan, Sabet, Riegler, & Halvorsen, 2022). Given the real-time conversational capabilities of LLMs and their cloud services, this prompts my third research question:

- **RQ3**: *Can we leverage LLMs to model affective robot behavior reliably and in real-time?*

As previously mentioned, affective behavior encompasses both the perception and generation of emotions. In this section, the focus is placed on the perception of emotional behaviors, specifically facial expressions, by robots and the factors that influence this process. Existing research establishes that human brains interpret facial expressions exhibited by robots in a manner similar to how they perceive human facial expressions (Chammat, Foucher, Nadel, & Dubal, 2010). Previous studies on the impact of emotional expressions displayed by robots have shown that robots exhibiting emotional behaviors are perceived as more likable (Rhim et al., 2019), trustworthy (Cominelli et al., 2021) and intelligent (Gonsior et al., 2011). This underscores the potential of emotionally intelligent robots in enhancing user experience, fostering effective communication, and building stronger human-robot relationships. To assess the recognition of robot expressions, researchers have largely employed the method of presenting participants with images or videos of robot expressions and asking them to recognize the emotions. Cañamero and Fredslund (2001) and Breazeal (2003) reported that participants were able to recognize robot expressions similarly to human expressions. Others have investigated the recognition rates across various form factors (ranging from human-like to non-human-like) (Becker-Asano & Ishiguro, 2011; Beer, Fisk, & Rogers, 2010; Danev, Hamann, Fricke, Hollarek, & Paillacho, 2017; Lazzeri et al., 2015). However, the influence of a robot's appearance on the recognition of its emotional expressions has remained relatively uncharted.

Another aspect to consider when evaluating the recognition of robot emotions is the specific facial regions employed to convey these emotions. Prior psychological research underscores the importance of observing a full-face configuration

in emotion recognition (Baron-Cohen, Wheelwright, & Jolliffe, 1997). However, emotional information is not uniformly distributed across the entire face (Sullivan, Ruffman, & Hutton, 2007). In certain cases, the eye region alone can provide sufficient information for emotion recognition (Wegrzyn, Vogt, Kirecli-oglu, Schneider, & Kissler, 2017). Social robots come in a wide range of form factors ranging from non-humanoid (e.g., iCat) to humanoid (Ameca). This includes static face designs, devoid of facial movements, as seen in robots like Nao, to full-face designs with human-like movements, as seen in robots like Furhat (Moubayed, Skantze, & Beskow, 2013). The influence of emotions expressed in specific face regions on the recognition of robot emotions remains largely unexplored. The answers to this question would highlight the significance of particular facial regions in emotion recognition and offer guidelines for robot face designs. These factors converge to frame my fourth research question:

- **RQ4**: *How does the appearance of the robot affect the perception of emotions, and what regions of the face are important?*

1.5 Robot Platform

A common limitation that was identified in the literature for both affective and gaze behavior was the selection of an optimal robotics platform capable of enabling the replication of human-like facial expressions and nuanced gaze behaviors. The primary objective of this dissertation is to model and evaluate the affective and gaze behaviors exhibited by social robots. The Furhat robot (Moubayed et al., 2013), which is a humanoid robotic head, has been chosen as the robotic platform for this dissertation. The robot's face is an animated character that is projected onto a translucent human face-shaped mask using back-projection. Using an animated face makes it possible to model realistic characters capable of exhibiting human-like facial expressions. Moreover, its 3-DoF (Degrees of Freedom) neck facilitates nuanced and independent eye and head movements. Figure 1.2 illustrates a few examples showcasing the ability of the robot to exhibit independent eye-head movements, facial expressions, and assuming different animated characters.

1.6 Thesis Outline

This dissertation is divided into two parts; the first part is dedicated to modeling the gaze behaviors of social robots and assessing their effects on HRI. The

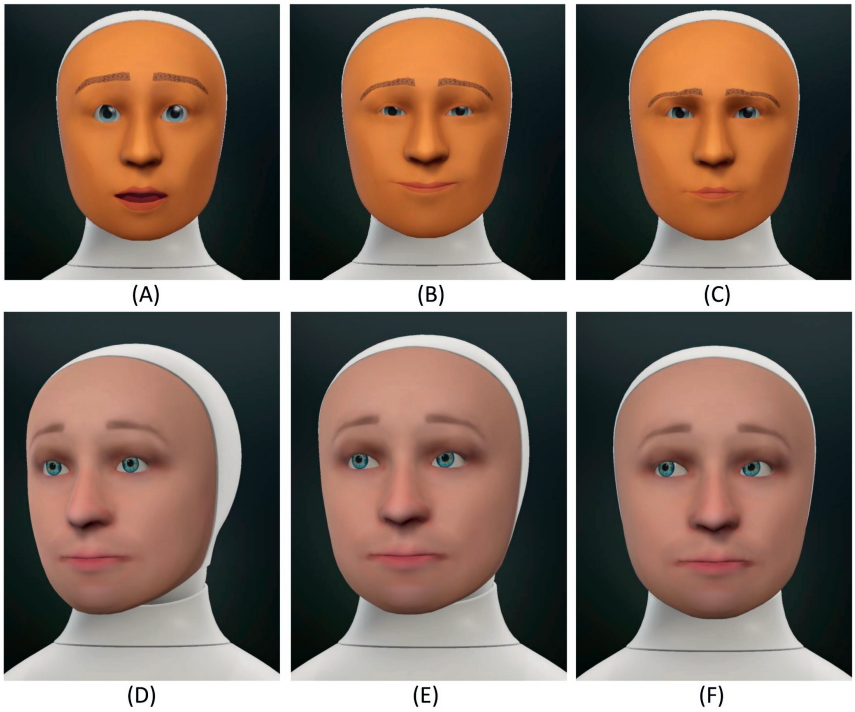


Figure 1.2: A few examples showcasing the capabilities of the Furhat robot. Sub-figures (A)-(C) depict 3 facial expressions; Surprise (A), Happy (B), and Angry (C). Sub-figures (D)-(F) show 3 examples of independent eye-head movements of Furhat; both eyes and head directed at a target (D), eyes fully directed at the target but head directed halfway (E), and only eyes directed at the target with no head movement (F)

second part investigates the affective behaviors of social robots and the factors influencing their perception by humans.

Chapter 2 focuses on **RQ1** which revolves around the development of a comprehensive GCS for social robots. As highlighted in Section 1.2), a key consideration in designing a better GCS is to overcome the *reactive* only paradigm characterizing existing models. The other aspect is to make the duration of gaze directed at targets more dynamic in nature. Finally, the GCS needs to be comprehensive, in that it should be able to model multiple communicative functions of gaze behavior (turn-taking, joint attention, gaze aversion, etc.). This chapter introduces an architecture for a comprehensive GCS that incorporates a *planning* component to pre-plan the robot's gaze behavior. The duration of gaze directed at a target is made dynamic and dependent on the plan. Additionally, the GCS

introduces a novel approach to coordinating the robot's eye-head movements during gaze shifts using the gaze plan. The architecture is evaluated through a user study involving an interactive multi-party card game. The participants played two games (one employing the proposed GCS and the other with a purely reactive GCS), with their subjective responses to a questionnaire serving as an evaluation metric.

Chapter 3 studies the influence of robot gaze behavior on human gaze behavior (**RQ2**). The study narrows its focus to the examination of gaze aversion. This particular choice is motivated by two primary factors. First, gaze aversion serves multiple communicative functions, including regulating intimacy (Abele, 1986), signaling cognitive load (Doherty-Sneddon & Phelps, 2005), and coordinating turn-taking (Ho, Foulsham, & Kingstone, 2015). Secondly, it is relatively easier to generate and perceive gaze aversion in robots. This chapter discusses the influence of robot gaze aversion on human gaze aversion behavior based on the results obtained from a user study. The study compares the gaze behavior of participants under two experimental conditions: the *Fixed Gaze* condition (which is the control condition) where the robot does not avert its gaze away from the participants and the *Gaze Aversion* condition where the robot's gaze is automated using the GCS proposed in **Chapter 2**. The analysis includes an objective examination of gaze data collected through eye-tracking glasses and subjective assessment based on questionnaire responses to draw conclusions.

Chapter 4 introduces and implements a model that leverages the capabilities of Large Language Models (LLMs) to generate context-appropriate facial expressions on a robot, aligning with **RQ3**. The chapter aims to achieve two primary objectives. Firstly, it seeks to evaluate the reliability and speed of utilizing LLMs (specifically GPT-3.5) for generating robot expressions. Secondly, it aims to assess whether people can perceive the context appropriateness of LLM-generated robot emotions. The implemented model is assessed through a user study in which participants engage in an affective-image sorting game with the robot. The game is intentionally designed to elicit emotional responses from the participants. The user study incorporates three experimental conditions: the *Neutral* condition, in which the robot displays no emotions on its face (serving as the control condition), the *Congruent* condition in which the robot exhibits emotions predicted by the LLM, and the *Incongruent* condition where the robot displays emotions opposite to those predicted by the LLM. The analysis of subjective responses to a post-interaction questionnaire and the scores for each game aids in validating the objectives of the study.

Chapter 5 focuses on identifying the factors that might influence the perception of emotions exhibited by robots (**RQ4**). The study investigates two main factors through an online user study: the impact of a human-like appearance and the specific use of the eye-region in emotion recognition of robot expressions. Recognition rates are compared across three appearance conditions: expressions by a human confederate, a human-like robot face, and a mechanical-looking robot face. Additionally, the study evaluates recognition rates between expressions displayed solely through the eye-region and those involving the full-face.

Finally, the dissertation concludes with **Chapter 6**, summarizing the results, presenting an overall discussion, and a conclusion.

2 | Knowing Where to Look: A Planning-based Architecture to Automate the Gaze Behavior of Social Robots[1]

Abstract

Gaze cues play an important role in human communication and are used to coordinate turn-taking and joint attention, as well as to regulate intimacy. In order to have fluent conversations with people, social robots need to exhibit human-like gaze behavior. Previous Gaze Control Systems (GCS) in HRI have automated robot gaze using data-driven or heuristic approaches. However, these systems tend to be mainly reactive in nature. Planning the robot gaze ahead of time could help in achieving more realistic gaze behavior and better eye-head coordination. In this paper, we propose and implement a novel planning-based GCS. We evaluate our system in a comparative within-subjects user study (N=26) between a reactive system and our proposed system. The results show that the users preferred the proposed system and that it was significantly more interpretable and better at regulating intimacy.

[1]Adapted from Mishra C and Skantze G (2022) Knowing Where to Look: A Planning-based Architecture to Automate the Gaze Behavior of Social Robots". *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Napoli*, Italy, 2022, pp. 1201-1208, doi: 10.1109/RO-MAN53752.2022.9900740

2.1 Introduction

Non-verbal cues play a crucial role in realizing effective communication in Human-Human Interaction (HHI). Humans make use of many non-verbal cues such as eye gaze, facial expressions, gestures, and prosody to convey meaning during social interactions. Among these non-verbal cues, eye gaze cues are considered to be especially important, as they are interpreted using dedicated and unique hard-wired pathways in the brain (Emery, 2000). Interpreting and conveying feelings and intentions through eye gaze during a social interaction is central to HHI and comes naturally to humans. During social interactions, gaze cues are used to coordinate turn-taking (Kendon, 1967), signal cognitive effort (Argyle & Cook, 1976), and regulate intimacy (Abele, 1986), among other things.

As social robots become increasingly available in society, they are expected to be able to communicate using both verbal and non-verbal cues, similar to humans. Research has shown that the robot's gaze behavior plays a similarly important role in Human-Robot Interaction (HRI) (Imai, Kanda, Ono, Ishiguro, & Mase, 2002; Yamazaki et al., 2008). Consequently, researchers have designed architectures to control the gaze behaviors of robots to explore the impact of social gaze in HRI and exploit the many uses of gaze cues in social interactions (Admoni & Scassellati, 2017; Pereira et al., 2019). Most of these Gaze Control Systems (GCS) have generally focused on modelling specific gaze cues such as gaze aversion (Andrist et al., 2014) or turn-taking (Mutlu et al., 2012).

Even though these GCS have achieved good results in emulating human-like gaze behaviors in robots, a common limitation is that they remain mainly *reactive* in nature. Although some of these systems do plan the gaze behavior for the upcoming utterance at the onset of the utterance (e.g. (Andrist et al., 2014)), the plan does not get updated incrementally, and the plan does not really affect the gaze behavior in the current moment. Another limitation of many systems is that they are static, in the sense that they use fixed durations for gaze shifts. For example, in Pereira et al. (2019), the gaze of the robot was fixed on the relevant target for a duration of 1-5 seconds during the interaction, before moving to the target with the lowest priority. In contrast, HHI involves a lot of *planning*. Research has shown that gaze behavior is coordinated with the underlying speech plan (Beattie, 2010). Depending on how long we plan to look at something, we determine whether a quick glance would suffice or whether we need to move the head and look. In this study, we focus on bringing a planning component into GCS. More specifically, we address the following research question:

How can planning be used to generate better gaze behavior in HRI?

Our model plans the priority for each potential gaze target (e.g., users or objects) in the environment incrementally (frame-by-frame) for a future rolling time window. At each time step, various events in the conversation might update these priorities, resulting in an evolving gaze plan which produces gaze behavior that is dynamic and differs in frequency and duration based on the state of the conversation. This planning allows the robot to better coordinate eye and head movements, since it is possible to compute for how long the robot will be attending a specific target in advance. In addition, the robot can better plan when to avert the gaze to regulate intimacy. The model is comprehensive in that it encompasses turn-taking, gaze aversion (GA), referential gaze (RG) and responsive joint attention (RJA). We evaluate the proposed model and compare it with a purely reactive heuristic model, using a multi-party interaction scenario where the robot head Furhat collaborates with two human players to sort a deck of cards in the right order.

The main contributions of this paper are:

- A comprehensive gaze control architecture that accounts for turn-taking, joint attention and intimacy regulation in HRI, using a planning-based approach.
- A novel approach to make use of the planned gaze behaviors to coordinate eye-head movements of the robot.
- An evaluation done in a complex multi-party HRI setting, which shows that this system is better than a reactive version of the same system.

2.2 Related Work

Modelling approaches in HRI for GCS can be broadly categorized into *data-driven* approaches (e.g., (Andrist et al., 2014; Mutlu et al., 2012)), where HHI gaze data are used to build models that can predict the gaze of the robot, and *heuristic* approaches, where the gaze of the robot is controlled using a set of rules, based on findings from HHI literature (e.g., (Mehlmann et al., 2014; Pereira et al., 2019)).

Turn-taking refers to the process in which interlocutors coordinate and take turns while speaking (Jokinen et al., 2013; Skantze, 2021). Mutlu et al. (2012) implemented data-driven models for role-signalling, turn-taking and topic signalling gaze mechanisms based on the formal observations of human communi-

cation. It was found that the subjects were able to correctly interpret the turn-yielding signals by the robot 99% of the time.

Gaze aversion is the intentional shifting of the gaze away from the interaction partner during a conversation. Several studies have focused on modelling this and evaluate human perception of it. Andrist et al. (2014) lists three primary functions of gaze aversion: cognitive, intimacy regulation and turn-taking. They used human gaze aversion data to model gaze aversion on a NAO robot and found that gaze aversion by the robot was perceived to be intentional. Zhong, Schmiedel, and Dornberger (2019) implemented a GCS with four possible states to model mutual gaze and gaze aversion using the captured gaze data of the participants. Lala, Inoue, and Kawahara (2019) used a heuristic model to generate appropriate gaze aversion along with verbal fillers as turn-taking cues.

When the interlocutors attend to a common target during a social interaction (and are mutually aware of that), it is generally referred to as **joint attention**. Joint attention is usually split into two categories: *responding to joint attention* (**RJA**) and *initiating joint attention* (**IJA**) (Mundy & Newell, 2007). IJA refers to when the interlocutor initiates a joint attention by directing the gaze at the referent (also known as **referential gaze**). RJA refers to the act of following others' gaze direction and interpreting the need to share focus on a common point. Mehlmann et al. (2014) proposed and implemented a GCS (*Sceneflow*) that made use of the bi-directional and multimodal aspects of speech. The model implemented referential gaze, RJA and mutual gaze as a hierarchical and concurrent state-chart-based architecture. Pereira et al. (2019) focused on the effects that RJA has on people's perception of social robots. The GCS was divided into two layers: *Proactive Gaze Layer* and *Responsive Gaze Layer* which modelled RJA and IJA respectively with each module having a predefined priority used to suppress gaze shifts issued by other modules with lower priorities.

Several studies have also developed models of human gaze behavior which could then be transferred to robots. Stefanov, Salvi, Kontogiorgos, Kjellström, and Beskow (2019) used a supervised learning approach to predict eye gaze direction or head orientation of the participant in multi-party open world dialogues. A recent study modelled the robot's gaze behavior using concepts from animation instead of grounding it in human psychomotor behavior (Pan et al., 2020).

Even though data-driven approaches are potentially able to provide a more accurate representation of human gaze behavior as compared to heuristic models, they are restricted by their dependence on collecting appropriate gaze data. It

is also unclear how well they generalize to settings different from that in which the data was recorded. Another problem is that the speakers' intentions are not available in the data, which makes it difficult for data-driven models to account for planning.

Another aspect of gaze modelling is the coordination between eye and head movements during gaze shifts. Previous studies have found that human eye and head movements are coordinated based on the target angles to realize gaze shifts (Stahl, 1999; Uemura et al., 1980). Hendrikse, Llorach, Grimm, and Hohmann (2018) defined eye-head angle relationships to control the eye-head movements of a virtual avatar during gaze shifts. Gu and Su (2006) and Wijayasinghe, Das, Miller, Bugnariu, and Popa (2019) tried to model realistic eye-head coordination on humanoid robots. To the best of our knowledge, ours is the first work that incorporates a planning component into a GCS to coordinate the eye-head movement during gaze shifts and regulate intimacy during a conversation.

We summarize these previous works on modelling GCS in Table 2.1. Planning here refers to the ability of a GCS to make use of the planned gaze behavior and the executed gaze behavior to adjust the current gaze behavior. As can be seen, our proposed model is unique in its comprehensiveness and its use of planning to control the gaze behavior of the robot.

2.3 Test-bed: Card Game

Our GCS was tested with the *Card Game* scenario, which is a test-bed specifically designed for studying multi-party interactions involving joint attention to objects (Skantze, Johansson, & Beskow, 2015). The Card Game setup consists of a Furhat robot (Moubayed et al., 2013), a touchscreen and up to two players, as seen in Fig. **??**. A set of cards are shown on the touchscreen and the task is to sort the cards based on some criterion. For example, the task could be to order a set of animals from slowest to fastest based on their running speeds. Furhat and the players then collaborate with each other to arrange the cards in the right order by moving them on the touchscreen.

During the game, players are encouraged to discuss among each other and with Furhat to reach a solution. Furhat's arguments are based on a randomized belief model, which means that the players have to choose whether they trust Furhat's beliefs or not. This results in a fairly free form of multi-party conversation, and therefore constitutes a good test-bed for studying turn-taking, joint attention and gaze aversions. When players look at or move a card, Furhat can

Table 2.1: Review of Gaze Models in HRI.

Paper	Multi-party	Modelling	Planning	Gaze Aversion	RJA[2]	Referential Gaze	Turn-Taking
Mutlu et al. (2012)	Yes	Data-driven	No	No	No	No	Yes
Andrist et al. (2014)	No	Data-driven	No	Yes	No	No	Yes
Mehlmann et al. (2014)	No	Heuristic	No	Yes	Yes	Yes	Yes
Zaraki, Mazzei, Giuliani, and De Rossi (2014)	Yes	Heuristic	No	No	No	No	Yes
Andrist, Mutlu, and Tapus (2015)	No	Data-driven	No	No	Yes	No	Yes
Nakano, Yoshino, Yatsushiro, and Takase (2015)	Yes	Data-driven	No	No	No	No	Yes
Lehmann, Keller, Ahmadzadeh, and Broz (2017)	No	Data-driven	No	No	No	No	Yes
Y. Zhang, Beskow, and Kjellström (2017)	No	Heuristic	No	Yes	No	No	Yes
Zhong et al. (2019)	No	Heuristic	No	Yes	No	No	Yes
Pereira et al. (2019)	No	Heuristic	No	Yes	Yes	Yes	Yes
Lala et al. (2019)	No	Heuristic	No	Yes	No	No	Yes
Stefanov et al. (2019)	Yes	Data-driven	No	No	Yes	Yes	Yes
Pan et al. (2020)	Yes	Heuristic	No	No	No	No	Yes
Proposed Gaze model	Yes	Heuristic	Yes	Yes	Yes	Yes	Yes

[1] Responsive Joint Attention. This is set to "Yes" if the GCS was capable of responding to user's referential gaze, verbal references to objects/ interlocutors/ locations, movement of task object etc.

Figure 2.1: Third person view of the Card Game setup

display RJA behavior and when Furhat talks about the cards or the game, it can generate referential gaze.

2.4 A Comprehensive Gaze Control Architecture

Fig. 2.2 shows the overall architecture of the GCS and how it is integrated. The *Robot Platform* consists of Furhat's output interfaces (projector, neck servo motors, etc.), input devices (microphone, camera, touchscreen, etc.), as well as the software modules for automatic speech recognition (ASR), text-to-speech synthesis (TTS), face tracking, etc. All the sensory inputs, modules and actuators in the Robot Platform are mediated by the *Event System*. The *Gaze Planner* subscribes to high-level events, such as the position of the user, speech input, and location of objects on the touchscreen, to generate a *Gaze Plan*. This plan is then used by the *Gaze Controller*, to generate events that make the robot move the eyes and turn the head. Interactions are implemented using the *Skill API*, with which all the interaction specific details are defined, such as the utterances of the robot, its facial expressions and head gestures, among others.

Algorithm 1 provides an overview of how the GCS works. During the course of an interaction, the Gaze Planner identifies and maintains a set of *gaze targets* ($\mathcal{T}_i$, $i \in [1..n]$) along with their current locations in real-time. These gaze targets could be of different types, such as *users*, *task objects* or the *environment*. The

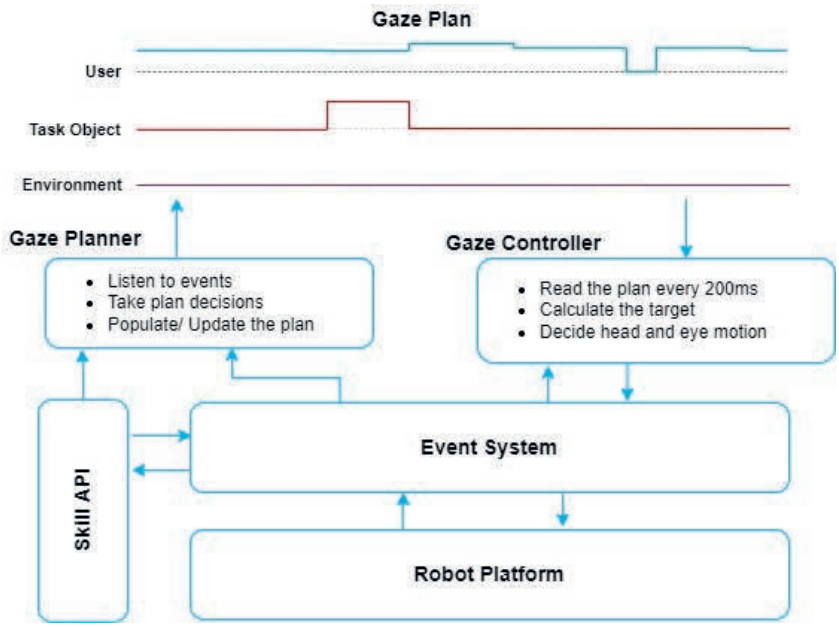


Figure 2.2: Overview of the proposed Gaze Control System

GCS continuously monitors the gaze targets to add new targets or remove targets as necessary.

We introduce a priority score $\mathscr{P} \in [0, 1]$, which determines the priority with which the GCS should be looking at a specific gaze target. The default priority is always 0. The Gaze Planner maintains a Gaze Plan (GP) which is used by the Gaze Controller to decide which gaze target to look at over a duration of time into the future. The GP stores the priority $\mathscr{P}_{i,j}$ for each target $\mathscr{T}_i$ at each future time frame j, with $j = 0$ being the immediate next time step in the future. We use a time resolution of 200ms for the plan, i.e., each time frame is 200ms long.

As can be seen in Algorithm 1, at each time step, the Gaze Planner listens to events and updates the relevant $\mathscr{P}_{i,j}$ values, as will be described in detail in the following sections. The relative priorities and specific durations chosen in this paper have either been obtained from literature or iteratively obtained after running the architecture for different scenarios. Thus, we acknowledge that these parameters are somewhat arbitrary, and that more optimal values can very likely be found. It is also possible to generate different robot gaze behaviors (e.g. introvert vs. extrovert) by tweaking the parameters. We leave this for future work.

Algorithm 1 Outline of the Gaze Control System

```
for each time step do:
  In Gaze Planner:
    updateTargets(GP)                                   ▷ Add/Remove 𝒯_i
    for each new event E do
      when E is RobotSpeaking
        checkPauses(GP)                                 ▷ see 2.4.2
        checkTurnYielding(GP)                           ▷ see 2.4.2
        checkReferentialGaze(GP)                        ▷ see 2.4.3
      when E is TargetsMoved
        attendTarget(GP)                                ▷ see 2.4.3
      when E is UserSpeaking
        checkRJA(GP)                                    ▷ see 2.4.3
        attendSpeaker(GP)                               ▷ see 2.4.2
      when E is RobotListening
        attendUser(GP)                                  ▷ see 2.4.2
    checkIntimacyRegulation(GP)                         ▷ see 2.4.4
  In Gaze Controller:                                   ▷ see 2.4.5
    GP_c = summarize(GP)
    𝒯_c, slack = getTarget(GP_c)
    headAngle = getHeadAngle(𝒯_c, slack)
    setRobotEyes(𝒯_c)
    setRobotNeck(headAngle)
  shift(GP)
```

At each time step, the Gaze Controller summarizes the current Gaze Plan, GP_c, by calculating the list of final gaze targets for the next 2 seconds into the future (10 time steps). The final gaze targets ($\mathscr{T}_{f,j}$) are calculated as the target that has the highest priority value in each frame j of the GP:

$$\mathscr{T}_{f,j} = \mathscr{T}_n, \arg\max_n(\mathscr{P}_{n,j}) \tag{2.1}$$

The current gaze target, $\mathscr{T}_c$, is then equal to the immediate next final target, $\mathscr{T}_{f,0}$. As will be described in Section 2.4.5, the rest of the GP_c is used to calculate the $slack$ value for the head movement, to achieve natural eye-head coordination.

After the Gaze Controller has executed the gaze and head movements for that time step, the GP is shifted one step, so that $\mathscr{P}_{i,j} \leftarrow \mathscr{P}_{i,j+1}$. Since the updates to the GP are done at each time step, it is possible for the Gaze Planner to overwrite $\mathscr{P}$ values in the plan, which makes it possible to dynamically update the plan depending on the events that occur during an interaction.

Fig. 2.3 shows an example of how the GCS works. Let us consider a scenario where a single user is playing the "Animal speed" Card Game, and a card with Zebra is being discussed. The $\mathscr{P}$ for each $\mathscr{T}$ in the game is plotted below the speech boxes. "Final Target" shows the calculated $\mathscr{T}_f$ for the entire example interaction. In the following subsections, we will use this example to discuss the various components of our GCS in detail.

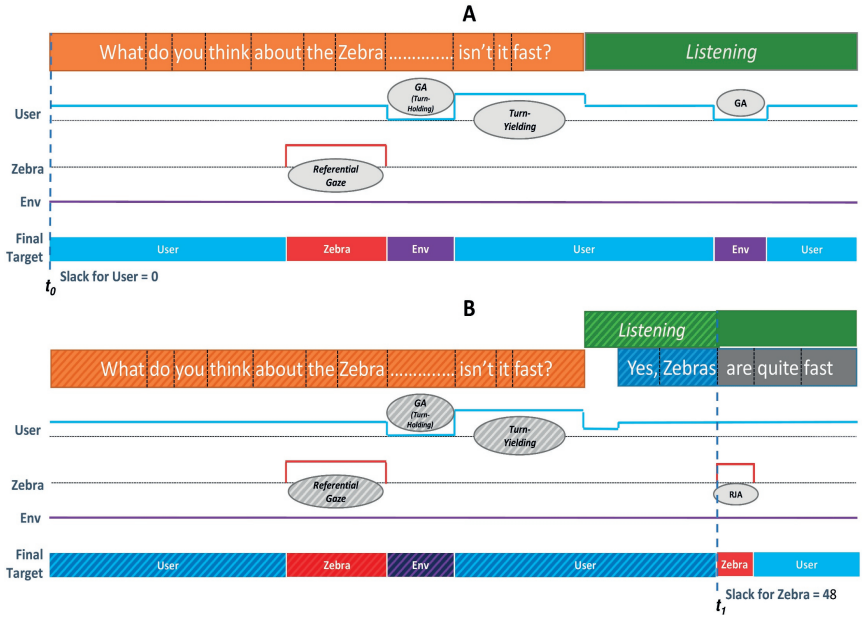


Figure 2.3: An example of gaze planning done by our GCS.**(A)** shows the plan at onset of the utterance t_0 and **(B)** shows the updated plan at time step t_1. The shaded parts show the already executed plan and the non-shaded part show the plan at the current time step. Grey speech boxes denote that the event is yet to take place.

2.4.1 Environment

As can be seen in Fig. 2.3, when the $\mathscr{P}$ value for all gaze targets in the plan are 0, the final gaze target is defaulted to the ENVIRONMENT (Env). Thus, when the user is given a low priority (e.g., due to gaze aversion) and no other target is given priority, the robot will gaze away from the user. Andrist et al. (2014) found that the distribution of eye gaze at various regions in the environment depended on the type of gaze aversion being performed. However, to keep the

model simple, we randomly select a location in the area around the currently addressed user's face as the location for the ENVIRONMENT gaze target.

2.4.2 Turn-taking

At the onset of a robot utterance, the Gaze Planner receives an event from the TTS system which gives information about the entire utterance text, as well as the phonetic transcription with precise timing information. This information can be used to plan the robot's gaze behavior related to speech production and turn-taking (as well as referential gaze, as described in the next section). During the course of the utterance (the $RobotSpeaking$ event), the $\mathscr{P}$ values of the currently addressed users are set to 0.3 which results in the robot looking at the user during the utterance, in the absence of other gaze targets with a higher $\mathscr{P}$ value. This can be seen at the beginning of the Gaze Plan (t_0) in Fig. 2.3A. This emulates **mutual gaze/ individual gaze** behavior where the speaker looks at the listener (Admoni & Scassellati, 2017).

Speakers tend to avert their gaze to signal that they are thinking or will hold the conversational floor (Andrist et al., 2014). The Gaze Planner calculates pause durations in the utterances it is about to speak (using the phoneme timings). If the pause duration is greater than 800ms, the $\mathscr{P}$ of the addressed users are set to 0 for that duration and the ENVIRONMENT becomes the $\mathscr{T}_f$ resulting in a turn-holding Gaze Aversion as seen in Fig. 2.3A.

By default, the robot will always hold the floor unless the $yielding$ flag in the $RobotSpeaking$ event is set to TRUE. This can be controlled through the Skill API (per default, $yielding$ is set to TRUE in case of a question). When $yielding$ is set to TRUE, the $\mathscr{P}$ values of the currently addressed user targets are set to 0.9, 1000ms before the end of the utterance. This results in a **turn-yielding** gaze cue (Admoni & Scassellati, 2017) as can be seen in Fig. 2.3A when the robot is asking a question. In case $yielding$ is set to FALSE, the $\mathscr{P}$ of the addressed users are set to 0 about 2000ms before the end of the utterance where we do not want the user to barge-in, and gaze aversion is a clear **turn-holding** cue (Jokinen et al., 2013).

When the robot is not speaking and instead listening to the user (the $Robot-Listening$ event), the $\mathscr{P}$ of the addressed users are set to 0.4 for the duration of the listening event. This enables the robot to keep looking at the user unless there is some higher-priority target. When a user starts to speak (the $UserSpeaking$ event), the $\mathscr{P}$ of that specific user target is increased to 0.6. In a multi-party setting, the array microphone of the robot can be used to sense

the speech direction, and thereby attribute the speech onset to the right user. This helps in directing the gaze of the robot to the active speaker and is in line with the findings in Vertegaal (1999), where it was found that the listeners always tend to spend the most time looking at the current speaker in multi-party settings.

2.4.3 Joint Attention

In the Card Game skill, the locations of the cards on the touchscreen and their order are being tracked as task object gaze targets. The $\mathscr{P}$ of the task objects can be raised if Furhat or a user talks about an object, or engages in joint attention in some other way. We do a keyword matching at the onset of the robot utterance (the *RobotSpeaking* event), to identify any references to a task object. If so, the $\mathscr{P}$ value of that task object is set to 0.9, 1000ms before the specific word is supposed to be spoken (timings are obtained from the TTS system) in order to generate **referential gaze**. This corresponds to the finding from studies of human communication, where gaze is directed at the referent about 800-1000ms before the reference is made (Mehlmann et al., 2014). The same can be seen in Fig. 2.3A, when the robot refers the Zebra card.

When a task object gaze target (i.e., a card on the touchscreen) is moved, the Gaze Planner sets gaze target's $\mathscr{P}$ to 1 for 2000ms which results in a **responsive joint attention (RJA)** and Furhat's gaze follows the card that is being moved. The current system is not capable of identifying and tracking objects other than the touchscreen locations. When Furhat is listening to user speech (the *UserSpeaking* event), we also do a keyword matching on the continuous ASR output to check for any references to the task objects. If a match is found, the corresponding task object's $\mathscr{P}$ is set to 0.7 (for a period of 800ms) after 200ms of the reference being heard. This is in line with what was reported in Mehlmann et al. (2014) and is also a form of RJA. This can be seen in Fig. 2.3B at t_1, when user refers the Zebra card.

2.4.4 Intimacy Regulation

Studies have shown that the preferred mutual gaze duration in interactions is between 3-5 seconds before the interlocutor starts to feel uncomfortable (Binetti, Harrison, Coutrot, Johnston, & Mareschal, 2016). To avoid this, the Gaze Planner also takes care of **intimacy regulating** gaze aversion. At every time step (200ms), the Gaze Planner checks the *GP* and makes sure that the $\mathscr{T}_c$ is not

assigned to a specific user for a duration longer than 3-5 seconds. In example Fig. 2.3A there is a small period of GA inserted when the planned final gaze target was User for a long duration when listening to the user, which results in intimacy regulating gaze aversions.

2.4.5 Eye-Head Coordination

While most of the previous works (see section 2.2) have made use of the target angle to coordinate the eye and head movements during gaze shifts, we propose that the planned duration of gaze also plays a role in determining the head and eye movements. If an agent knows that the gaze is planned to be directed at a specific location for a longer period of time, then it can move both the eyes and the head towards that target immediately. On the other hand, if the planned gaze duration is very short, the gaze shift should be done using only the eyes (or with little head rotation). Otherwise, rapidly shifting gaze targets could result in jerky head movements. We can mitigate this problem thanks to the planning approach of our GCS. Additionally, this allows the GCS to take the robot's intention into account when planning the gaze behavior. To the best of our knowledge, this has not been addressed in any previous GCS.

In our GCS, we introduce a control variable named $slack$ which is the angle by which the head direction is allowed to deviate from the eye gaze direction. As seen in Algorithm 1, at every time step (200ms), the Gaze Controller calculates the final gaze targets ($\mathcal{T}_f$) for 2 seconds into the future and summarises them in a list GP_c. The current gaze target ($\mathcal{T}_c$) is the first element in the GP_c. $slack$ calculation uses the frequency of the same gaze target in the final gaze plan as per equation:

$$slack = max(48 - (sameTargetFreq * 6), 0) \tag{2.2}$$

$sameTargetFreq$ is the duration of time the gaze is to be directed at a specific target. It is calculated as the frequency of having the same target as the $\mathcal{T}_f$ within a future window of 2s. This determines whether the gaze should be directed using just the eyes or using both the eyes and the head. For example, as in Fig. 2.3A, it can be seen that at t_0, the final target is planned to be User for a long duration. Thus, the $slack$ value is set to 0, and the head direction is fully aligned with the eyes. In Fig. 2.3B, at t_1, the gaze duration planned at the Zebra card is very short so the $slack$ value is set to 48. This means that only eye gaze is

directed at the Zebra card for a quick glance. The value 48 has been iteratively obtained.

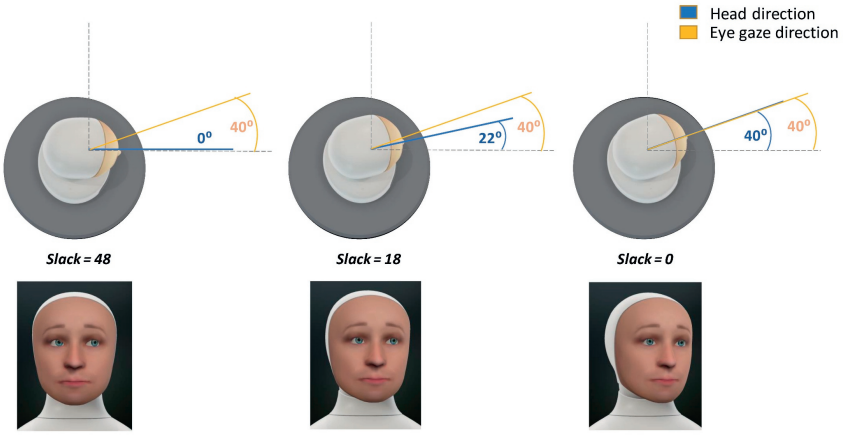


Figure 2.4: An example showing the differences in eye and head movements when looking at a target depending on the $slack$ values.

Fig. 2.4 shows an example of how $slack$ is used in coordinating the eye and head movements towards the gaze target. Since the neck movement takes some time, and the eye movement is instantaneous, the Gaze Controller first moves the eyes towards the gaze target and then the neck, while centering the eyes.

Sometimes, several gaze targets may have an equally high $\mathscr{P}$ value. For example, in multi-party interactions, two users might be addressees and would have an equally high $\mathscr{P}$ while the robot is speaking. In such cases, it is natural for the speaker's head to be directed somewhere in the middle of the gaze targets, while the eye gaze shifts rapidly between each of them (Stahl, 1999; Uemura et al., 1980). We model this behavior in our GCS by checking for any rapid shifts between two or more targets for a future time window. If so, then at every time step, the mid point between the $\mathscr{T}_c$ and the next $\mathscr{T}_f$ is calculated and head is directed at that point.

2.5 Experimental Evaluation

In order to evaluate if a GCS that takes the robot's intention (future gaze behavior) into account is perceived as better, we performed a user study to compare our GCS's performance to a purely reactive GCS.

2.5.1 Experimental Setup & Procedure

We used the Card Game scenario described in section 2.3 for the experiments. A camera was placed behind the participants so that it only captured Furhat's face and the touch screen.

There were two scenarios under which the participants played the Card Game:

- **Planned :** Our GCS with planning.
- **Reactive :** A purely reactive heuristic GCS similar to (Pereira et al., 2019) was implemented and used as a baseline for the comparative study. The gaze targets were chosen in response to the events taking place in the Card Game (e.g., when the cards were moved, when someone was speaking, etc.). The gaze of the robot was fixed on one target for a duration of 1-5 seconds (same as the original work) before moving to the next target.

Each session lasted approximately 30 minutes, during which 2 participants played 2 games (1 from each scenario) together with Furhat. The experiments followed a within-subjects design and the order of scenarios were alternated between sessions. The participants were first guided to their seats in front of the touchscreen and provided with a consent form. Then the researcher briefed the participants about the goal of the study and the way the experiment was going to be conducted. They were told that they would play two games with different versions of the system, but the nature of these two versions was not explained to them, and they were only referred to as scenario 1 and 2. Participants were encouraged to go through the questionnaire to have a better grasp of what to look out for before starting the first game. After each game, the participants filled out one part of the questionnaire. The questionnaire had two 9-point likert scales placed under each question; one for each scenario. The participants were asked to score the questions based on how they perceived the interaction in terms of the question being asked. They were asked to look for differences in the scenarios and make different judgements where applicable. The participants were instructed not to discuss the scenarios with each other before filling out the questionnaire. At the end of the session, the participants were also asked to choose which of the two interactions they preferred.

2.5.2 Data Collection and Evaluation

We recruited 28 participants to take part in the user study (14 males and 14 females) with ages ranging between 18 and 51 (mean = 32.92, SD = 8.22), and

participants were paired up. The responses from 2 participants were removed from the analysis, as they violated the instructions and discussed the scenarios with each other prior to filling out the questionnaire. No prior interaction with social robots was needed before participating in the experiment. The experiments were conducted in English.

The questionnaire had 10 9-point Likert scale questions which were grouped into 5 dimensions, as can be seen in Table 2.2. As the goal of our study was to to compare two GCSs, we selected the dimensions and questions based on aspects that should be important for a good GCS. For each dimension, the mean score of the responses to the individual questions in that dimension was calculated. We refer to this as the *dimension score*. For a better GCS, we expected the *dimension scores* to be high for all dimensions, except for the *Intimacy* dimension, which should be lower, given how the questions were formulated. We used the statement "Furhat kept staring at me too much" as a sign of bad Intimacy regulation, since periodic GA while listening leads to making speakers more comfortable and reduces negative perceptions (Abele, 1986). While we use the term Intimacy as a short label for this dimension, this question does not of course capture all aspects of intimacy, but it was designed to be easy to interpret for the participants.

At the end of the experiment, the participants were also asked (verbally) which scenario they preferred, taking all factors into account.

2.6 Results

Fig. 2.5 shows the *dimension scores* of each dimension for both GCS versions, based on the responses from the 26 participants. As our hypothesis was that they would prefer the Planned version, we performed a one-tailed Wilcoxon signed-rank test. Since we compared five dimensions, we set $\alpha = 0.01$, after Bonferroni correction. Significant results were obtained for *Interpretation* ($p = 0.007$) and *Intimacy* ($p = 0.0013$). While the mean values for *Awareness* and *Human-Likeness* were higher for the Planned version, the differences were not statistically significant.

For the final preference question, 19 participants preferred the Planned version, whereas 4 preferred the Reactive version, and 4 could not decide. We found the results to be significat with $p = 0.0002$ and $\chi^2 = 16.66$.

Table 2.2: Questionnaire used for evaluation

Dimension	Question
Awareness	Furhat looked at the cards at the right time.
	Furhat was aware of what was happening in the game.
Interpretation	I could interpret Furhat's intention from its gaze.
	Furhat's gaze helped me understand its instructions better.
Turn-taking	I was able to understand when Furhat wanted me to speak.
	I was able to understand when Furhat wanted to keep speaking.
	I was able to understand when Furhat was talking to me.
Human-likeness	Furhat's gaze was human-like.
	The coordination between eye and head movements seemed natural.
Intimacy	Furhat kept staring at me too much.

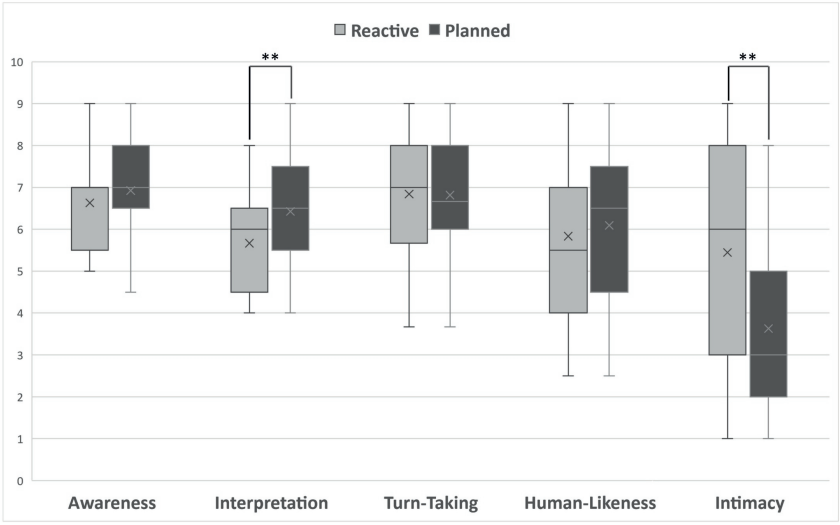


Figure 2.5: The comparison between the responses obtained for the reactive system and the proposed GCS. The grey bar in the box plot indicates the median, the × denotes the mean, the boxes show the upper and lower quartiles of the data. The bars on both ends of the vertical lines denote the maximum and minimum values in the data. ** ($p < 0.01$)

2.7 Discussion

The goal of our evaluation was to do a comparison between a purely reactive GCS and the planning-based GCS proposed in this paper. We hypothesised that using a GCS with planning can lead to improved perception of the robot's gaze behavior during the interaction. The results from the evaluation indicate that our GCS was significantly more interpretable, had better intimacy regulation, and was generally preferred over the reactive version.

When it comes to the dimensions of Awareness, Turn-taking and Human-likeness, we did not find any significant differences. One possible explanation could be that it might have been difficult for the participants to observe subtle gaze behaviors while being engaged in playing a new game. Previous studies on turn-taking models for conversational systems have shown that it is very hard for the participants to judge subtle things like turn-taking while interacting themselves (Meena, Skantze, & Gustafson, 2014). This is also in line with the *Load Theory* (Lavie, Hirst, De Fockert, & Viding, 2004), which says that when there is a higher cognitive load, the selective attention performance becomes poor. Another problem that was noticed during the experiments was that the sound source localization on the robot was not always working very well. This means that Furhat sometimes turned to the wrong participant, which clearly could have affected the perception of the Turn-taking dimension.

Another issue could be the novelty effect; most of the participants were interacting with a social robot for the first time. Attention might have been split in familiarizing themselves with the robot and its capabilities during the first scenario, which could have impacted the ratings. A potential way of mitigating this problem in future studies could be to let the participants first do a test round where they familiarize themselves with the robot. A potential follow-up study could be to show the recordings of the experiments to third-party observers and let them compare the two versions. In doing so, the participants would be able to solely focus on rating the robot's gaze during the interactions. Meena et al. (2014) reported that third-party observers could easily perceive differences between different turn-taking models, unlike the participants who were engaged in the interaction.

It should also be noted that many of the parameters chosen for the model could of course be further tuned. We can also envision a hybrid model, where certain priority scores are being data-driven, and others are rule-driven.

2.8 Conclusion

In this paper, we proposed and implemented a novel planning-based comprehensive GCS to automate the gaze behavior of social robots. The system is capable of planning the gaze behavior for a future rolling time window of fixed length, and use this plan to coordinate the eye and head movements of the robot taking the robot's intention into account. We conducted a user study to evaluate our GCS and compared it with a purely reactive GCS. The results suggest that a GCS with such type of planning is perceived to be significantly more interpretable and has better intimacy regulation. It was also found that overall, our GCS was preferred over the reactive system when users were asked to choose one. This shows that planning is an important aspect of gaze control, which has not been considered in previous works.

3 | Does a Robot's Gaze Aversion Affect Human Gaze Aversion?[1]

Abstract

Gaze cues serve an important role in facilitating human conversations and are generally considered to be one of the most important non-verbal cues. Gaze cues are used to manage turn-taking, coordinate joint attention, regulate intimacy, and signal cognitive effort. In particular, it is well established that gaze aversion is used in conversations to avoid prolonged periods of mutual gaze. Given the numerous functions of gaze cues, there has been extensive work on modelling these cues in social robots. Researchers have also tried to identify the impact of robot gaze on human participants. However, the influence of robot gaze behavior on human gaze behavior has been less explored. We conducted a within-subjects user study (N=33) to verify if a robot's gaze aversion influenced human gaze aversion behavior. Our results show that participants tend to avert their gaze more when the robot keeps staring at them as compared to when the robot exhibits well-timed gaze aversions. We interpret our findings in terms of intimacy regulation: humans try to compensate for the robot's lack of gaze aversion.

[1]Adapted from Mishra C, Offrede T, Fuchs S, Mooshammer C and Skantze G (2023) Does a robot's gaze aversion affect human gaze aversion?. *Front. Robot. AI 10:1127626*. doi: 10.3389/frobt.2023.1127626

3.1 Introduction

It is well established that gaze cues are one of the most important non-verbal cues used in Human-Human Interactions (HHI) (Kendon, 1967). Several studies have shown the many roles gaze cues play in facilitating human interactions. When interacting with each other, people use gaze to coordinate joint attention, communicating their focus of attention and perceiving their partner's focus to follow (Tomasello et al., 1995). Ho et al. (2015) also showed how people use gaze to manage turn-taking: for instance, gaze directed at or averted from one's interlocutor can indicate whether a speaker is intending to yield or hold the turn (for example when making a pause), or when the listener is intending to take the turn.

Given the importance of gaze behavior in HHI, researchers in Human-Robot Interaction (HRI) have tried to emulate human-like gaze behaviors in robots. The main motivation behind such Gaze Control Systems (GCS), or models of gaze behavior, has been to exploit the many functionalities of gaze cues in HHI and realize them in HRI. Moreover, thanks to the sophisticated anthropomorphic design of many of today's social robots (e.g., Furhat robot (Moubayed et al., 2013) or iCub robot (Metta et al., 2010)), it is possible to model nuanced gaze behaviors with independent eye and head movements. It has been established that robots' gaze behaviors are recognized and perceived to be intentional by humans (Andrist et al., 2014). Robots' gaze behaviors have also been found to play an equally important role in HRI as human gaze in HHI (Imai et al., 2002; Yamazaki et al., 2008). Thus, researchers have measured the impact of robots' gaze behavior on human behavior during HRI. In Schellen et al. (2021) participants were found to become more honest in subsequent trials if the robot looked at them when they were being deceptive. Skantze (2017) and Gillet et al. (2021) observed that robots' gaze behavior could lead to more participation during group activities. Most of these works have concentrated on human behavior in general, but not the gaze-to-gaze interaction between robots and humans. This then leads to our research question:

- *Does a robot's gaze behavior have any influence on human gaze behavior in a HRI?*

Answering this question is important because it can help in designing better GCS and interactions in HRI. Even though previous works have shown various ways in which humans perceive and respond to robot gaze behavior, whether there are changes in human gaze behavior as a direct influence of robots' gaze

behavior has remained less explored. Moreover, most of these studies have used head movements instead of eye gaze to model robot gaze behavior, due to physical constraints of the robots used (Andrist et al., 2014; Mehlmann et al., 2014; Nakano et al., 2015). While head orientation is a good approximation of gaze behavior in general, it lacks the rich information ingrained in eye gaze. Additionally, from a motor control perspective, eye gaze is much quicker than head motion and is therefore also more adaptable than moving the head. Thus, we were interested in verifying if subtle gaze cues performed by a robot are perceived by humans and if it had any influence on their own gaze behavior.

In order to verify the impact of robot gaze behavior, we narrowed our focus to gaze aversions for this study. This was mainly motivated by two considerations. First, gaze aversion has been shown to play an important role in human conversations: coordinating turn-taking (Ho et al., 2015), regulating intimacy (Abele, 1986) and signalling cognitive load (Doherty-Sneddon & Phelps, 2005). Secondly, it is an important gaze cue which is relatively easy to perceive and generate during HRI.

In this work, we designed a within-subjects user study to measure if gaze aversion exhibited by a robot has any influence on the gaze aversion behavior of participants. We automated the robot's gaze using the GCS proposed in Mishra and Skantze (2022) (more details in Section 3.3) to exhibit time- and context-appropriate gaze aversions. Participants' gaze was tracked using eye-tracking glasses throughout the interactions. Subjective responses were also collected from the participants after the experiment, using a questionnaire that asked about their impression of the interaction. Our results show that participants avert their gaze more when the robot doesn't avert its gaze as compared to when it does.

The main contributions of this paper are:

- The first study (to the best of our knowledge) that verified the existence of a direct relationship between robot gaze aversion and human gaze aversion.
- A study design to measure the influence of a robot's gaze behavior on human gaze behavior.
- An exploratory analysis of the eye gaze data, which pointed towards a potential positive correlation between gaze aversion and topic intimacy of the questions.

3.2 Background

Gaze aversion is the act of shifting the gaze away from one's interaction partner during a conversation. Speakers tend to look away from the listener more often than the other way around during a conversation. This has been thought to help plan the upcoming utterance and avoid distractions (Argyle & Cook, 1976). It has been found that holding mutual gaze significantly increases hesitations and false starts (Beattie, 1981). Speakers process visual information from their interlocutors, produce speech and plan the upcoming speech, all at the same time. Prior studies in HHI have shown that people use gaze aversions to manage cognitive load (Doherty-Sneddon & Phelps, 2005) because averting gaze reduces the load of processing the visual information. Ho et al. (2015) found that speakers signal their desire to retain the current turn, i.e., turn-holding, by averting their gaze and that they begin their turns with averted gaze. Additionally, gaze aversion has been found to have a significant contribution in regulating the intimacy level during a conversation (Abele, 1986). Binetti et al. (2016) found that the amount of time people can look at each other before starting to feel uncomfortable was 3-5 sec.

Several studies have modelled gaze aversion behavior in social robots and evaluated their impact. Andrist et al. (2014) collected gaze data from HHI and used that to model human-like gaze aversions on a NAO robot. They found that well-timed gaze aversions led to better management of the conversational floor and the robot being perceived as more thoughtful. Zhong et al. (2019) controlled the robot's gaze using a set of heuristics and found that users rated the robot to be more responsive. Subjective evaluation of the gaze system in Lala et al. (2019) showed that gaze aversions with fillers were preferred when taking turns. On the other hand, there have been a few studies that included gaze aversions as a sub-component of their GCS, but they did not measure any effects of gaze aversion (Mehlmann et al., 2014; Mutlu et al., 2012; Pereira et al., 2019; Y. Zhang et al., 2017). For example, Mehlmann et al. (2014) looked at the role of turn-taking gaze behaviors as a whole to evaluate their GCS. However, it is important to note that both Mehlmann et al. (2014) and Y. Zhang et al. (2017) used the gaze behavior of participants as feedback to manage the robot's gaze behaviors. Mehlmann et al. (2014) grounded their architecture on the findings from HHI, whereas Y. Zhang et al. (2017) relied on findings from human-virtual agent interactions.

Although it has been established that humans perceive robot gaze as similar to human gaze in many cases (Staudte & Crocker, 2009; Yoshikawa, Shinozawa,

Ishiguro, Hagita, & Miyamoto, 2006), it is still important to verify if it holds for different gaze cues and situations in an HRI setting as findings from Admoni, Bank, Tan, Toneva, and Scassellati (2011) suggest that robot gaze cues are not reflexively perceived in the same way as human gaze cues. Thus, it is crucial to investigate whether a relationship exists between robot gaze behavior and human gaze behavior, how they are related, and what are the implications of such a relationship. For example, if it is known that lack of gaze aversion by a robot makes people uncomfortable, then we might want to include appropriate gaze aversions when designing a robot for therapeutic intervention. On the other hand, we would probably do less gaze aversions when designing an interaction where a robot is training employees to face rude customers. To the best of our knowledge, this is the first work that tries to establish a direct relationship between robot gaze aversion and human gaze aversion behavior.

3.3 Automatic Gaze Aversion using GCS

To automate the robot's gaze behavior in this study, we used the GCS proposed in Mishra and Skantze (2022). It is a comprehensive GCS that takes into account a wide array of gaze-regulating factors, such as turn-taking, intimacy, and joint attention. The gaze behavior of the robot is planned for a future rolling time window, by giving priorities to different gaze targets (e.g., users, objects, environment), based on various system events related to speaking/listening states and objects being mentioned or moved. At every time step, the GCS makes use of this plan to decide where the robot should be looking and to better coordinate eye-head movements.

To model gaze aversion, the model processes the gaze plan at every time step to check if the gaze of the robot is planned to be directed at the user for a duration longer than 3-5 seconds (the preferred mutual gaze duration from HHI Binetti et al. (2016)). If that is the case, the model inserts intimacy-regulating gaze aversions into the gaze plan. This results in a quick glance away from the user for about 400ms using the eye gaze only. Additionally, when the robot's intention is to hold the floor at the beginning of an utterance or at pauses, the GCS also inserts gaze aversions at the appropriate time to model turn-taking and cognitive gaze aversions.

The parameters of the model are either taken from the literature or tuned empirically. This, combined with the novel eye-head coordination, results in a human-like gaze aversion behavior by the robot. In a subjective evaluation of

the GCS through a user study, it was found to be preferred over a purely reactive model, and the participants especially found the gaze aversion behavior to be better (Mishra & Skantze, 2022).

3.4 Hypotheses

Abele (1986) found that too much eye gaze directed at an interlocutor would induce discomfort for the speaker and that periodic aversion of gaze would result in a more comfortable interaction. The *Equilibrium Theory* (Argyle & Dean, 1965) also suggests an inverse relationship between gaze directed at and gaze averted, arguing that increased gaze at an interlocutor would be compensated with more gaze aversions by them. While the theory also discusses other factors such as proxemics, we were interested only in the gaze aspect and in verifying if there is an effect of robot gaze on human gaze behavior. Additionally, it is known that while listening, individuals tend to look more at their speaking interlocutors whereas while speaking, they tend to exhibit more gaze aversions (Argyle & Cook, 1976; Cook, 1977; Ho et al., 2015). Thus, if the robot is not averting its gaze during the interaction, we can expect the participant to produce more gaze aversion while speaking, but not necessarily while listening. Based on this, we formulate the following hypotheses:

- **H1** *Lack of gaze aversions by a robot will lead to an increase in gaze aversions by the participants when they are speaking.*
  - ***H1a:*** *Participants will avert their gaze away from the robot longer in the condition when the robot does not avert its gaze away from the participants. (see Section 3.5)*
  - ***H1b:*** *Participants will look away from the robot more often when the robot exhibits fixed gaze behavior (does not avert its gaze).*

3.5 Study Design

To investigate the effect of a robot's gaze aversion on human gaze aversion, we designed a within-subjects user study with two conditions. In the control condition, the robot constantly directs its gaze towards the participant, without averting it; we call this the *Fixed Gaze (FG)* condition. In the experimental condition (which we call the *Gaze Aversion (GA)* condition), the robot's gaze is

automated using the GCS described in Subsection 3.3 which is found to be better at exhibiting gaze aversion behavior in a subjective analysis. While the GCS is capable of coordinating individual eye and head movements, the interaction is designed in such a way that it does not require any head movements by the robot when directing its gaze. This is because the interaction involved mainly intimacy-regulating gaze aversions, which necessitate only a quick glance away from the interlocutor (see Subsection 3.3). Hence, the robot's head movements are not a factor in the study, which is in line with our aim to verify the effect of robot's eye gaze behavior on human gaze behavior.

3.5.1 Interaction Setting

We designed an interview scenario similar to that in Andrist et al. (2014), where the robot asked the participant six questions with increasing levels of intimacy (more details in Subsection 3.5.2). While Andrist et al. (2014) investigated whether appropriate gaze aversions by the robot would elicit more disclosure, we wanted to verify if gaze aversions by a robot would directly elicit lower gaze aversions by humans, signaling more comfort even with highly intimate questions (which are known to induce discomfort). To make the interaction more conversational and less one-sided, the robot also gave an answer to each question after the participant had answered it. Questions with different levels of intimacy were used in order to vary the level to which the participant might feel the need to avert their gaze.

The robot's turns were controlled by the researcher using the Wizard-of-Oz (WoZ) approach. The researcher listened to the participant's responses through a wireless microphone and controlled the robot's response by selecting one of three options, which resulted in varying flows of the conversation script. On selecting "Robot answer", the robot would answer the question that was asked to the participant before moving on to ask the next question. The option "User declined to answer" would prompt the robot to acknowledge the user's choice before moving on to the answer, and then ask the next question. The "User asked to repeat question" option was used to repeat the question. Having a WoZ paradigm enabled the researcher to control the timing of the robot's turn-taking, resulting in a smooth conversational dynamics. Additionally, it made it possible for the researcher to manage the interaction from a separate room, reducing the influence that the presence of a third-person observer might have on the participants. The robot's responses were handcrafted to be generic enough to account for most of the answers that participants might provide. They always

started with an acknowledgement of the participant's answer (e.g, "*I appreciate what you say about the weather*"). Then a response was chosen at random from previously created pool of handcrafted answers to the question and appended to the acknowledgement. In cases where the participant did not answer the question, the robot always acknowledged that by using phrases like "*That's okay*" and then appended a random response from the pool of of answers.

An example dialog where the participant answered the question has been provided below (R denotes the robot, P denotes a participant):

R: *What do you think about the weather today?*

P: *I think it is perfect. It is neither freezing nor too hot. Just the perfect balance of sunny and cool. I really don't like if it is too hot or too cold.*

R: *I appreciate what you say about the weather, but honestly, I can't relate. I never get to go outside. Maybe you didn't notice, but I don't have legs. So I never have any idea what the weather is like out in the real world. My dream is to one day see the sky. Perhaps my creators will allow me some day.*

R: *What are your views on pop music?*

We used a Furhat robot for the study, which is a humanoid robot head that projects an animated face onto a translucent mask using back-projection and has a mechanical 3-DoF neck. This makes it possible to generate nuanced gaze behavior using both eye and head movements, as well as facial expressions and accurate lip movements (Moubayed et al., 2013). For the experimental condition (Gaze Aversion; GA) the robot was named Robert and for the control condition the robot was named Marty. We wanted to give the impression that the participants were interacting with two distinct robots for each condition, but at the same time, we did not want the robots themselves to have an influence on the interaction. This led to the selection of two faces that were similar to each other from the list of characters already available in the robot. Two male voices were selected from the list of available voices based on how natural they sounded when saying the utterances for the tasks. The participants were not informed about the different gaze behaviors of the two robots.

The experiment was conducted in a closed room while restricting any outside distractions. The participants were alone with the robot during the interactions. Participants were asked to sit in a chair that was placed approximately 60-90 cm in front of the robot. The robot was carefully positioned such that it was almost at eye level and at a comfortable distance for the participants. A Tobii

Figure 3.1: Experimental setup for the interview task

Pro Glasses 2 eye-tracker was used to record the participants' eye gaze during each interaction. We also recorded the speech of the participants using a Zoom H5 multi-track microphone. A pair of Rode Wireless Go microphone systems was also used to stream the audio from the user to the Wizard. Fig. 3.1 shows an overview of the experimental setup.

3.5.2 Intimacy Rating of Questions

The questions for the task were selected from Hart, VanEpps, and Schweitzer (2021) and Kardas, Kumar, and Epley (2021), who asked their participants to rate them in terms of sensitivity and intimacy, respectively. In order to account for any influence culture and demography might have on the perceived topic intimacy levels of the questions, an online survey was conducted where residents of Stockholm rated these questions based on their perceived topic intimacy. Participants were recruited using social media forums for Stockholm residents (e.g., Facebook groups, Stockholm SubReddit). Another consideration was to avoid complex questions that would involve a lot of recalling or problem-solving (e.g., "What are your views about gun control?"). The motivation for this is that people

are known to avert their gaze when performing a cognitively challenging task (Doherty-Sneddon & Phelps, 2005). We wanted to keep the questions as simple as possible so as to restrict the influence on gaze aversions to just the robot's gaze behavior and the question's intimacy level.

A total of 28 questions were selected from the questions in Hart et al. (2021) and Kardas et al. (2021). The participants were asked to rate the questions on how intimate they felt on a 9-point Likert scale ranging from "1: Not intimate at all" to "9: Extremely intimate" (question asked: *Please indicate how intimate you find the following questions (1: not intimate at all; 9: extremely intimate). Please don't think too much about each one; just follow your intuition about what you consider personal/ intimate*). The responses from 148 participants (68 females, 76 males, one non-binary & two undisclosed), aged between 18 and 50 (mean = 29.35, SD = 6.89), were then used to order the questions based on their intimacy values. Using linear mixed models, it was verified that gender, age, nationality and L1 did not influence the intimacy ratings. We selected a total of 12 questions out of them and divided them into two sets with similar intimacy distribution which were used evenly across both the conditions (*FG* and *GA*). We tried to select simple questions that would not induce a heavy cognitive load. Table 3.1 lists the questions and their rated intimacy values from the survey (Q Set stands for Question Set).

3.5.3 Participants

We recorded eye gaze and acoustic data of 33 male participants (sex assigned at birth). The choice for male participants was methodologically and logistically motivated. Firstly, topic intimacy has been found to be perceived differently by people of different genders (Sprague, 1999). Thus, intimacy during the interaction might be affected by the participants' and robot's gender. To reduce the influence of this variable (given that it is not a variable of interest in this study), we controlled it by recruiting participants of only one gender.

In addition to the participants' gaze behavior, in future works we will use the recorded data to analyze their speech acoustics in relation to that of the robot. Since sex and gender are known to impact acoustic features of speech (Pépiot, 2014), all processing and analysis of data need to be carried out separately for males and females. This would reduce the statistical power of the acoustic analysis, leading us to choose participants from only one sex. Given the choice between female or male participants, males were chosen since they are more numerous in the institute where we collected data.

Table 3.1: Mean intimacy ratings of selected questions used in the study

Question	Mean	SD	Q Set
What do you think about the weather today?	1.192	0.558	1
What are your views on pop music?	1.976	1.372	1
How did you celebrate last Christmas?	3.023	1.758	1
Tell me about a conversation you had with another person earlier today.	4.330	2.121	1
For what in your life do you feel most grateful?	5.223	1.917	1
What is one of the more embarrassing moments in your life?	6.538	2.016	1
What did you have for breakfast this morning?	1.823	1.308	2
What season do you like the best? Why?	1.838	1.091	2
Do you have anything planned for later today? What will you do?	3.523	1.779	2
What would constitute a perfect day for you?	4.007	1.827	2
Is there something you've dreamed of doing for a long time? Why haven't you done it?	5.430	2.064	2
Can you describe a time you cried in front of another person?	7.023	1.918	2

The participants were recruited using social media, notice boards and the digital recruitment platform Accindi (https://www.accindi.se/). The participants were all residents of Stockholm. The cultural background of participants was not controlled for. Participants' ages ranged between 21 and 56 (mean = 30.54, SD = 8.07). They had no hearing or speech impairments, had normal/corrected vision (did not require the use of glasses for face-to-face interactions) and spoke English. They were compensated with a 100SEK gift card on completion of the experiment. The study was approved by the ethics committee of Humboldt-Universität zu Berlin.

3.5.4 Procedure

As described earlier, the study followed a within-subjects design. Each participant interacted with the robot under two conditions; the order of the conditions was randomized. Each set of questions (cf. Table 3.1) was also counterbalanced across the conditions. The participants were asked to give as much information as they could when answering the questions. However, they were not forced to answer any of the questions. In case they did not feel comfortable answering any questions, the robot acknowledged it and moved on to the next question.

The interaction always started with the robot introducing itself before moving on to the questions. The entire experiment took approximately 45 minutes. The experiment's procedure can be broken down into the following steps:

- **Step 1:** The participants were informed about the experiment's procedure, compensation, and data protection, both verbally and in writing. They then provided their written consent to participation.
- **Step 2:** The participants were instructed to speak freely about a prompted topic for about 2 minutes. This recording was used as the baseline speech measure for participants' speech before interacting with the robot. The speech data is not discussed in the present work.
- **Step 3:** Next, the participants were asked to put on the eye-tracking glasses, which were then calibrated. After successfully calibrating the glasses, the researcher left the room and initiated the interview task. The robot introduced itself and proceeded with the Q&A.

  The researcher kept track of the participant's responses and timed the robot's turns with the appropriate response using the wizard buttons. Once the interaction came to an end, the researcher returned to the room for the next steps.
- **Step 4:** The participants were then asked to remove the tracking glasses and were provided with a questionnaire to fill in. The questionnaire had 9-point Likert scale questions about the participant's perception of the robot and the flow of conversation (see Table 3.2).
- **Step 5:** Next, they were asked to fill out the Revised NEO Personality Inventory (NEO-PI-R) (Costa Jr & McCrae, 2008), which measures personality traits. They were also asked to take the LexTALE test (Lemhöfer & Broersma, 2012), which indicates their general level of English proficiency, on an iPad. Both of these tasks served as distractor tasks, providing a break between the two interactions and allowing the participants to focus on the second robot with renewed attention.
- **Step 6:** The participants were then asked to speak freely about another prompted topic for about 2 minutes. This served as the baseline for the second interaction before the participant interacted with the robot (data not discussed here).

- **Step 7:** After recording the free speech, the participants were asked to put on the eye-tracking glasses and the tracker was calibrated again. The researcher left the room and initiated the next interaction. The robot introduced itself again and proceeded with the Q&A.
- **Step 8:** At the end of the interaction, the researcher returned to the room and provided the participants with the last questionnaire. Apart from the 9-point Likert scale questions about the perception of the robot and the conversation flow, the questionnaire also asked about basic demographic details.

3.5.5 Measurements

In order to test **H1**, we mainly focused on the behavioral measure of gaze behavior of the participants, which was captured using the eye-tracking glasses. Our experiment had one independent variable, the *gaze aversion* of the robot which was manipulated in a within-subjects design (*GA* & *FG* condition). The order of the questions remained the same for both the *GA* and *FG* conditions, i.e., increasing intimacy with each subsequent question.

The Tobii Pro Glasses 2 eye-tracker records a video from the point of view of the participant, and provides the 2D gaze points (i.e., where the eyes are directed in the 2D frame of the video). The videos were recorded at 25fps and the eye-tracker sampled the gaze points at a 50Hz resolution. Both datasets were synchronized to obtain timestamp vs. 2D gaze point ($[ts,(x,y)]$) data for each recording, i.e, gaze location per timestamp. We used the Haar-cascade algorithm available in the OpenCV library to detect the face of the robot in the videos and obtain the timestamp vs. bounding box of face ($[ts,(X,Y,H,W)]$, X and Y - lower left corner of the bounding box, H and W - height and width of the bounding box) data.

Gaze Aversion for each time stamp was calculated by verifying if the gaze points (x,y) were inside the bounding box $[X,Y,H,W]$ or not. The parameters for Haar-cascade were manually fine-tuned for each recording to obtain the best fitting bounding boxes for detecting the robot's face. An example of non-gaze aversion detection using the algorithm can be seen in Fig. 3.2. The timing information for the robot's utterances can be obtained from the speech synthesizer. We logged the robot's responses and their time information for all interactions. This log was used to extract the participant's speaking and listening durations. When the robot is speaking, the participant is the listener and vice-versa. This

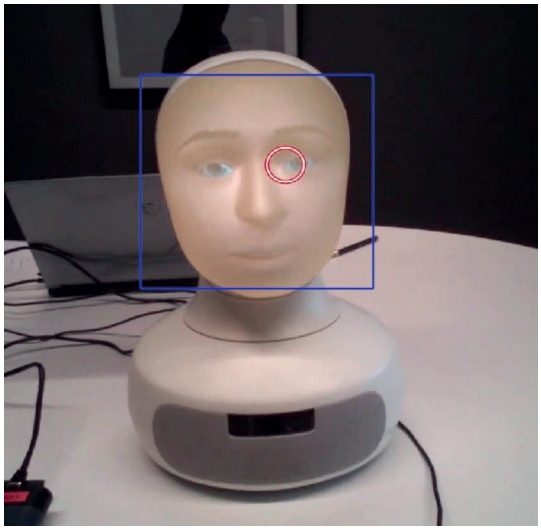


Figure 3.2: Example of Gaze Aversion detection using the algorithm. Here the gaze point (x, y) (the red circle) lies within the face's bounding box $[X, Y, H, W]$ (blue rectangle), so it is not a Gaze Aversion

information was used to extract the gaze aversion of the participant when they were *Speaking* and *Listening*.

For ***H1a***, we used the % of gaze aversion as the metric of overall gaze aversion. Each timestamp where it was possible to detect whether there was a gaze aversion or not was considered as a *gaze event*. We counted the number of gaze aversions ($gaCount$) and the total number of *gaze events* ($geTotal$) over the duration when the participants were *Speaking* and *Listening*. The % of gaze aversion ($ga\%$) is then calculated as:

$$ga\% = gaCount/geTotal \tag{3.1}$$

For ***H1b***, we identified individual gaze aversion instances, which are the number of times the participants directed their gaze away from the robot. The duration from when participants looked away from the robot until the time they returned their gaze back at the robot was counted as one gaze aversion instance.

We also collected subjective feedback from the participants for both conditions with a questionnaire. The questionnaire included the 12 questions that were used to measure the responses of the participants under three dimensions on a 9-point Likert scale (see Table 3.2).

Table 3.2: Questionnaire used for subjective evaluation

Dimension	Question
Conversation Flow (**D1**)	My conversation with the robot flowed well.
	I was able to understand when the robot wanted me to speak.
	I was able to understand when robot wanted to keep speaking.
	The robot responded to me at the appropriate time.
Human-Likeness (**D2**)	The robot's face was very human-like.
	The robot's voice was very human-like.
	The robot's behavior was very human-like.
	Throughout the conversation, I was very aware that I was talking to a robot.
Overall Impression (**D3**)	I enjoyed talking with the robot.
	I felt positively about the robot.
	I felt positively about the conversation.
	I felt comfortable while talking with the robot.

The analysis of speech data is beyond the scope of this work and will be analysed in conjugation with other variables in upcoming works.

3.6 Results

As mentioned in Subsection3.5.1, we used a WoZ approach to manage the robot's turns. While the wizard was instructed to behave in the same way for both the conditions, we wanted to make sure that the wizard did not influence the turn taking of the robot, which could in turn influence the gaze aversion behavior of the participants. We calculated the turn gaps (time between when the participant has finished speaking and the robot started to speak) from the audio recordings of the interactions. A Mann-Whitney test indicated that there was no significant difference in turn gaps between condition *FG* (N= 249, M = 1.89, SD = 2.67) and condition *GA* (N = 243, M = 1.77, SD = 1.63), W = 29848, p = 0.797. This shows that the wizard managed the turns in the same way across conditions.

Table 3.3: Mean % of Gaze Aversion per Condition

	Condition: GA		Condition: FG	
$ga\%$	Mean	SD	Mean	SD
Speaking	0.399	0.195	0.456	0.199
Listening	0.112	0.084	0.137	0.155

3.6.1 Effect of Robot's Gaze Aversion Behaviour

Of the 33 participants recorded, we excluded two participants' data from the analysis as the gaze data was corrupted due to some technical problems with the eye-tracker. Additionally, gaze data from the eye-trackers were not always available for all timestamps, due to various reasons such as calibration strength and detection efficiency. When averting gaze, participants also moved their head away from their partner's face. This varied a lot from participant to participant and led to instances where the robot's face was out of the eye-tracker's camera frame. Additionally, there were instances where Haar-cascade could not detect the robot's face for some timestamps due to various reasons. These factors resulted in instances where it was not possible to determine if there was a gaze aversion or not. We were able to capture 87.38% of gaze data (data loss = 12.62%), which is normal for eye-trackers (Holmqvist, 2017). Overall only 1.6% of data (8170 timestamps out of 504011) was affected by the technical constraints which led to the exclusion of data. Thus, instances of gaze aversion by participants where Furhat was out-of-frame (due to head movement) are not very common. Also, the amount of data lost in this way was the same across conditions so we do not believe that the excluded data had any influence on the results reported.

On average, participants averted their gaze more in the *FG* condition as compared to the *GA* condition when they were *Speaking*. A two-tailed Wilcoxon signed-rank test indicated a significant difference in gaze aversion across conditions when the participants were *Speaking* ($W = 142.0$, $p = 0.037$), as shown in Fig. 3.3. This supported ***H1a***, which predicted that participants would avert their gaze for a longer duration when there is no gaze aversion by the robot (i.e., the *FG* condition). There was no significant difference between conditions when participants were *Listening* ($W = 150.0$, $p = 0.194$) which is expected (see 3.4). The mean values of gaze aversion when participants were*Speaking* and *Listening* can be found in Table 3.3.

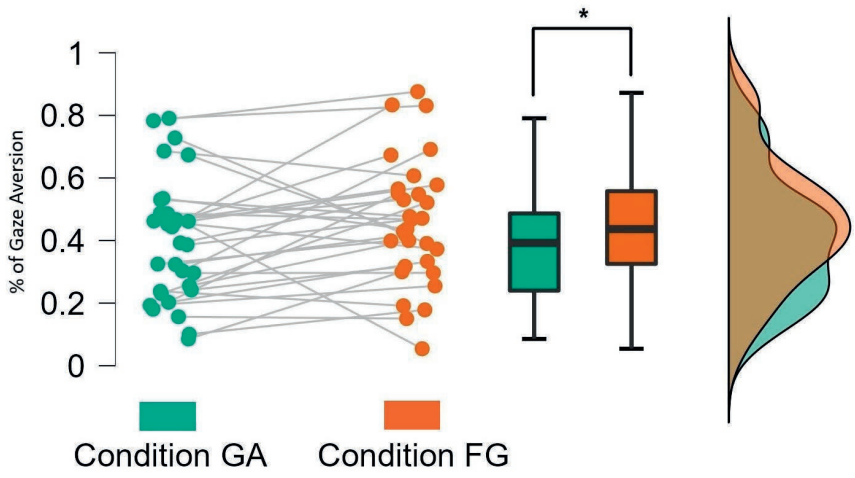


Figure 3.3: Total % of Gaze Aversion while *Speaking*

Analyzing the number of gaze aversion instances performed by the participants while *Speaking* showed that participants looked away from the robot more frequently in the *FG* condition (Mean = 91.742, SD = 60.158) as compared to the *GA* condition (Mean = 70.774, SD = 46.141). As shown in Fig. 3.4, a two-tailed Wilcoxon signed-rank test indicated a significant difference in the number of gaze aversion instances across the conditions ($W = 496.00$, $p < 0.001$). This supported ***H1b***, which predicted that participants would look away from the robot more often when there is no gaze aversion by the robot. It can be seen that the effect of robot's gaze aversion on participants' gaze behavior is stronger and more distinct when analyzing gaze aversion instances. We argue that gaze aversion instance is a better metric to verify the effect.

3.6.2 Gaze Aversion when participants were Speaking & Listening

It is already known from the HHI literature (Argyle & Cook, 1976; Cook, 1977; Ho et al., 2015) that people exhibit fewer gaze aversions while listening and more while speaking. We were interested to see if there was a similar pattern emerging from the data.

To verify this, we first calculated the % of gaze aversion when participants were listening to and answering each of the robot's questions. Since the dura-

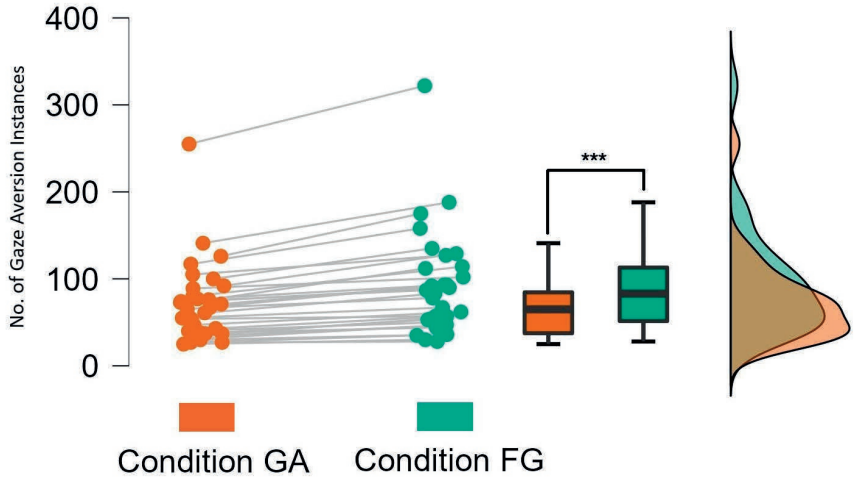


Figure 3.4: Number of Gaze Aversion Instances while participants were *Speaking*

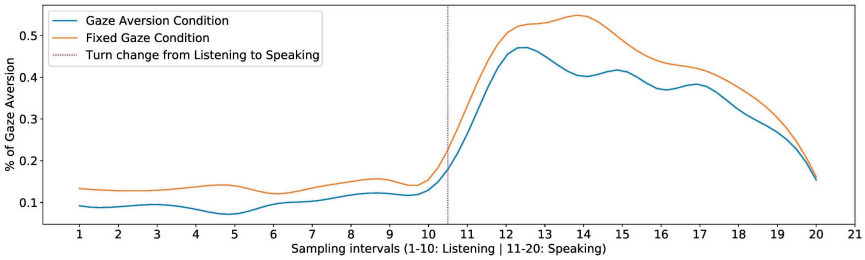


Figure 3.5: % of Gaze Aversion while participants were *Listening* and *Speaking*, for the two conditions.

tions of both *Speaking* and *Listening* varied from one participant to the other, we normalized the time into 10 intervals for *Speaking* and 10 intervals for *Listening* phase. Next, we found the aggregate % of gaze aversion for all the questions when *Speaking* and *Listening*. The resulting plot can be seen in Fig. 3.5.

We can see a clear trend emerging from the plot with the low gaze aversion during the *Listening* phase when the participants listened to the robot. However, just before taking the floor (*Speaking* phase), it can be seen that the gaze aversion starts increasing. This is in line with the findings from Ho et al. (2015), who found that speakers usually started their turns with gaze aversion and averted their gaze before taking the turn. We also notice that the gaze aversion peaks at around 20-30% of the speaker's turn, before starting to fall. Towards the end of the turn, we see a sharp decline in gaze aversion. This is consistent

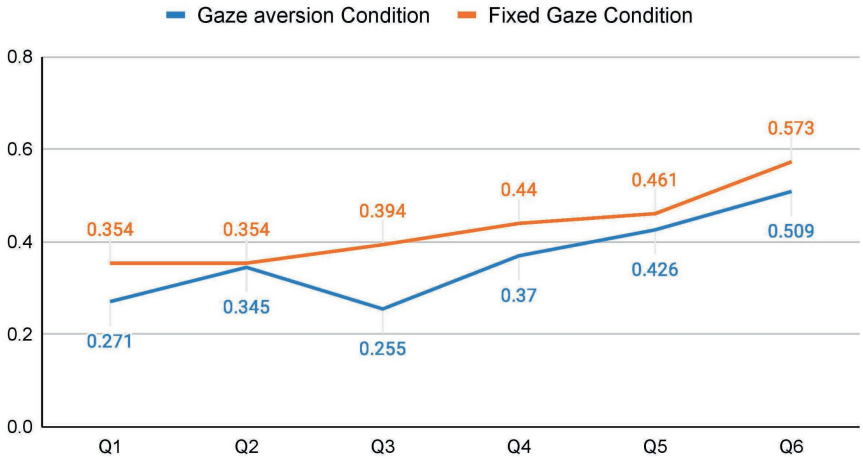


Figure 3.6: Mean Gaze Aversion per Question while participants were *Speaking*

with the findings from Ho et al. (2015), which show that people end their turns with their gaze directed at the listener. It can also be seen that even though the gaze aversion behavior of participants followed a similar pattern for both conditions, the amount of gaze aversion was lower for the *GA* condition. This further supports hypothesis **H1**.

3.6.3 Results from the Questionnaire

On analyzing the responses from the questionnaire, all three dimensions were found to have good internal reliability (Cronbach's $\alpha = 0.8$, 0.71 & 0.92 respectively). The participants found the robot under the *FG* condition to be significantly more *Human-Like* (Student's t-test, $p = 0.029$). This result was unexpected and has been further discussed in Section 3.7. We did not find any significant differences for the other two dimensions. The mean score for the LexTALE test was 80.515% (SD = 12.610) which showed that the participants had good English proficiency (Lemhöfer & Broersma, 2012).

3.6.4 Exploratory Analysis: Topic Intimacy

Apart from analyzing the data for **H1**, we were also interested in whether any trends emerged through an exploratory analysis of the topic intimacy of the questions and gaze aversion. By plotting the mean % of gaze aversion values of all participants for each question during the *Speaking* phase, we can see that there

Table 3.4: Fixed effect estimates of the GLMM model

Term	Estimate	SE	t
Intercept (Question 6)	0.396	0.030	13.138
Question 1	-0.084	0.022	-3.846
Question 2	-0.046	0.021	-2.224
Question 3	-0.071	0.019	-3.731
Question 4	0.009	0.018	-0.515
Question 5	0.047	0.023	2.085
Condition:*GA*	-0.033	0.016	-2.098

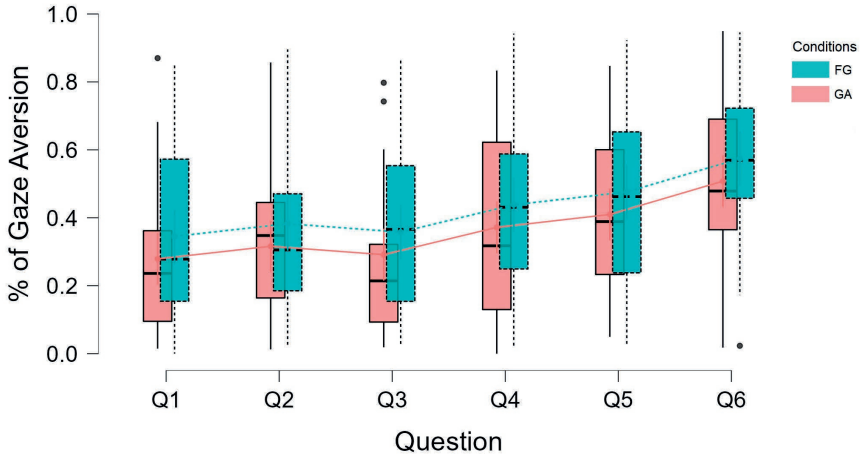


Figure 3.7: Distribution of Mean Gaze Aversion per Condition per Question Order

is an increase in gaze aversion as the intimacy values increase with the question order (cf. Fig. 3.6).

We fit a GLMM (Generalized Linear Mixed Model) with mean gaze aversion values per question of each participant as the dependent variable. The questions' order and the conditions were used as the fixed effects variables, and we included random intercepts for participants and random slopes for question order and condition per participant. The model suggested that the *gaze aversions increased as the intimacy values increased* ($\chi^2 = 41.32$, $df = 5$, $p < 0.001$). It also suggested that there was more gaze aversion in the *FG* condition as compared to the *GA* condition ($\chi^2 = 4.244$, $df = 1$, p=0.039). There were no interaction effects observed. The coefficients of the model can be found in Table 3.4.

This points in the direction of a positive correlation between topic intimacy and gaze aversion. One interpretation of this finding is that participants tend to

compensate for the discomfort caused by highly intimate questions by averting their gaze. This is in line with previous findings from HHI that suggest that any change in any of the conversational dimensions like proximity, topic intimacy or smiling would be compensated by changing one's behavior in other dimensions (Argyle & Dean, 1965). Fig. 3.7 is a visualization of how gaze aversion varied for each condition under each question. It can be seen that the gaze aversion was higher for *FG* for all the questions (except Q3), and that there is an increase of gaze aversion with the increase in question number (which in turn is the topic intimacy value for the question).

The finding here is interesting because it could mean that the participants compensated for topic intimacy with gaze aversion even when it is a robot that was asking the questions. However, since we didn't control for the order if the questions, this could also be because of other factors such as a cognitive effort and fatigue. Further studies should narrow down the factors that influenced such behavior.

3.7 Discussion

The results suggest that participants averted their gaze significantly more in the *FG* condition. Moreover, they had more gaze aversion instances in the *FG* condition. This was supported by both Wilcoxon signed-rank tests (see Section 3.6) and an exploratory GLMM (see Section 3.6.4). The results are in line with hypothesis **H1**: people compensate for the lack of robot gaze aversion by producing more gaze aversions themselves (Abele, 1986; Argyle & Dean, 1965).

We did not observe a significant difference across conditions in gaze aversion when participants were *Listening*. This could be attributed to the fact that there were too few gaze aversions during this phase to observe a significant difference, which is also suggested by prior studies in HHI (Ho et al., 2015). The gaze aversions varied between 11-14%, which meant that the participants directed their gaze at the robot for about 86-89% of the time. This is higher than the numbers reported in HHI, where listeners direct their gaze at speakers 30-80% of the time (Kendon, 1967). Our findings coincide with the findings in Yu, Schermerhorn, and Scheutz (2012), where they reported that humans directed their gaze more at a robot than at another human.

Unexpectedly, participants rated the robot in the *FG* condition as more human-like compared to the *GA* condition. A key reason for that could be the way the *GA* interaction started. The GCS used would make the robot keep looking at random

places in the environment unless the interaction is started by the researcher. This could have resulted in an unnatural behavior where the robot directs its gaze at random places even though the participant is already sitting in front of it. On the other hand, in the *FG* condition the robot kept on looking straight and only started to track the user when the interaction started. However, since the participant was sitting right in front of the robot, it would be perceived as the robot looking at the participant all the time.

While we did not find any significant differences in the other two dimensions (Conversation Flow & Overall Impression) assessed in the self-reported questionnaire, we did see a significant difference across conditions from the objective measures (i.e., gaze behavior). This could point to an effect that, even though it might not be explicitly perceived by people, a robot's gaze behavior would implicitly affect human gaze behavior. This could also be an interesting direction for further study.

A further exploratory analysis of the data reveals a positive correlation between gaze aversion and topic intimacy of the questions. Thus, more intimate questions seem to lead to a larger avoidance of eye gaze. In our study, more intimate questions occurred towards the end of the conversation. As the order of the questions was fixed, the order may of course be a confounding factor. However, we are not aware of other work showing that humans would avoid eye gaze more and more over the conversation. We argue that eye gaze is rather related to the topic (intimacy), but further work is needed that controls for this potential confound.

3.8 Limitations & Future work

The participants of our study had a rather large age span and we had only male participants. A clear limitation of this study is the lack of a balanced dataset. As the results obtained are only for male participants, these results do not necessarily generalize to other genders. The choice for male participants was methodologically and logistically motivated. Firstly, topic intimacy has been found to be perceived differently by people of different genders (Sprague, 1999). Thus, intimacy during the interaction might be affected by the participants' and robot's gender. To reduce the influence of this variable (given that it is not a variable of interest in this study), we controlled it by recruiting participants of only one gender.

In addition to the participants' gaze behavior, in future works we will use the recorded data to analyze their speech acoustics in relation to that of the robot. Since sex and gender are known to impact acoustic features of speech (Pépiot, 2014), all processing and analysis of data need to be carried out separately for males and females. This would reduce the statistical power of the acoustic analysis, leading us to choose participants from only one sex. Given the choice between female or male participants, males were chosen since they are more numerous in the institute where we collected data.

Further studies with a more diverse participant pool and female-presenting robots would be needed to verify this effect in general. However, it is interesting to note that a recent study (Acarturk et al., 2021) found no difference in gaze aversion behavior due to gender. The authors concluded that GA behavior was independent of gender and suggested "that it arises from the social context of the interaction."

It is known that culture also influences our gaze behavior on many levels such as how we look at faces (Blais, Jack, Scheepers, Fiset, & Caldara, 2008) or interpretation of mutual gaze and gaze aversions (Argyle & Cook, 1976; Collett, 1971). McCarthy, Lee, Itakura, and Muir (2006) observed that people mutual gaze and gaze aversion behaviors during thinking differed based on the culture of the individuals. However, recent studies have challenged some of aspects of cultural influences that have been reported previously (Haensel, Smith, & Senju, 2022). Nonetheless, investigating any effect culture of participants may play on their gaze behavior when interacting with a robot could also be an interesting area to look into in the future.

3.9 Conclusion

In this paper, we investigated whether a robot's gaze behavior can affect human gaze behavior during HRI. We conducted a within-subjects user study and recorded participants' gaze data along with participants' responses. The analysis of participants' eye gaze in both conditions suggests that they tend to avert their gaze more in the absence of gaze aversions by a robot. An exploratory analysis of the data also indicated that more intimate questions may lead to a larger avoidance of mutual gaze. The existence of a direct relationship between robot's gaze behavior and human gaze behavior is an original finding.

The study also shows the importance of modelling gaze aversions in HRI. In the absence of robot gaze aversions, the interaction may become more effortful

for the user while trying to avoid frequent mutual gaze with the robot. These findings go hand in hand with the Equilibrium Theory suggesting a trade-off relation between the robot's and user's interactive gaze behavior. Our findings are helpful for designing systems more capable of adapting to the context and situation by taking human gaze behavior into account.

4 | Real-time Emotion Generation in Human-Robot Dialogue Using Large Language Models [1]

Abstract

Affective behaviors enable social robots to not only establish better connections with humans, but they also serve as a tool for the robots to express their internal states. It has been well established that emotions are important to signal understanding in Human-Robot Interaction (HRI). This work aims to harness the power of Large Language Models (LLM) and proposes an approach to control the affective behavior of robots. By interpreting emotion appraisal as an Emotion Recognition in Conversation (ERC) task, we used GPT-3.5 to predict the emotion of a robot's turn in real-time, using the dialogue history of the ongoing conversation. The robot signalled the predicted emotion using facial expressions. The model was evaluated in a within-subjects user study (N = 47) where the model-driven emotion generation was compared against conditions where the robot did not display any emotions and where it displayed incongruent emotions. The participants interacted with the robot by playing a card sorting game that was specifically designed to evoke emotions. Results indicate that the emotions were generated in a reliable way by the LLM and the participants were able to perceive the robot's emotions. It was found that the robot expressing congruent model-driven facial emotion expressions was perceived to be significantly more human-like, emotionally appropriate, and elicit a more positive impression. Participants also scored significantly better in the card sorting game when the robot displayed congruent facial expressions. From a technical perspective, the study shows that LLMs can be used to control a robot's affective behavior reliably in real-time. Additionally, our results could be used in devising novel human-robot interactions, making robots more effective in roles where emotional interaction is important, such as therapy, companionship, or customer service.

[1] Adapted from Mishra C, Verdonschot R, Hagoort P and Skantze G (2023) Real-time Emotion Generation in Human-Robot Dialogue Using Large Language Models. *Front. Robot. AI 10:1271610*. doi: 10.3389/frobt.2023.1271610

4.1 Introduction

Affective behavior, the ability to perceive and express emotions, is a fundamental component in human communication. It is instrumental in building human relationships (Lazarus, 2006) and decision making (So et al., 2015). Humans use facial expressions to convey various meanings during interactions (Elliott & Jacobs, 2013). Consequently, with social robots poised to have a greater integration in society, it is prudent for these robots to be able to exhibit affective behavior. For robots to interact with humans socially, they need to be able to perceive human behaviors and the intent behind them while also expressing their understanding and intention. Facial expressions can be used by robots to signal their intentions and internal state. Research has shown that robots exhibiting emotions are more likely to be perceived as likeable (Rhim et al., 2019), intelligent (Gonsior et al., 2011), and trustworthy (Cominelli et al., 2021) by users. Emotionally responsive robots can adapt their behavior and responses based on the user's emotional states, leading to more natural and seamless interactions between humans and robots. Emotionally intelligent robots have the potential to enhance user experience, facilitate effective communication, and establish stronger rapport with humans.

However, effectively modelling emotions in robots is a challenging and active area of research. Emotions are complex, multi-dimensional phenomena that involve a combination of physiological, cognitive, and expressive components. Researchers have explored both dimensional (Mehrabian, 1995; Russell, 1980) and categorical (Ekman, 1999; Tomkins & McCarter, 1964) theories of emotions to develop models for robot emotion generation, leading to complex architectures that interpret various stimuli to generate appropriate emotional responses (Cavallo et al., 2018). While these models have shown promising results, they often require hand-crafted rules and intricate feature engineering, making them labor intensive.

The emergence of Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), has significantly transformed the landscape of natural language understanding and generation. LLMs can serve as general models for solving a multitude of tasks. For example, Lammerse et al. (2022) used GPT-3 to detect the emotions of utterances in an Emotion Recognition in Conversation (ERC) task. We aimed to harness these capabilities of LLMs to model robot emotions, specifically to generate real-time robot emotions during HRI (Human Robot Interactions). In this paper we investigate two research questions:

- *Can we use LLMs for robot emotion generation in real-time?*
- *Do people perceive the context appropriateness of a robot's emotions and what is its effect on the user?*

In this study, we implemented a model to use GPT-3.5, a state-of-the-art LLM, to control the affective behavior of a robot. We interpreted emotion appraisal as a real-time ERC task. We used GPT-3.5 to predict the emotion that the robot is likely to during real-time interactions, based on the ongoing conversation's dialogue history. The predicted emotions were then translated into facial expressions which were displayed by the robot.

To evaluate the effectiveness of our implemented model, we conducted a within-subjects user study involving 47 participants. The participants engaged in an affective image sorting game, with a robot acting as a collaborative partner. The game was designed to evoke emotional responses from the participants. The results of the study demonstrated the effectiveness of using GPT-3.5 in generating emotions in real-time.

To summarize, the main contributions of this work are:

- The first study (to the best of our knowledge) that showcases the use LLMs for emotion generation in HRI
- A novel study design to evaluate the influence of a robot's emotional expressions on human users in a collaborative setting.

4.2 Background

Emotions can be defined as "*an instantaneous affective response to an experienced event*" (Cavallo et al., 2018). Appraisal theories aim to propose a theoretical framework to understand the cognitive evaluations or appraisal of various stimuli that result in eliciting specific emotions (Ellsworth & Scherer, 2003). On the other hand, theories of emotions try to describe various emotions and discuss the similarities and differences between them. Categorical theories of emotions propose a set of specific emotion categories (e.g., Happy, Sadness, Anger, Fear, Surprise, Disgust) that are elicited due to various stimuli (Ekman, 1999; Izard, 2013; Tomkins & McCarter, 1964). Dimensional theories, on the other hand, model emotions based on certain underlying dimensions (such as arousal and valence) (Mehrabian, 1995; Plutchik, 1982; Russell, 1980).

For a robot to provide an appropriate affective response during an interaction with a human user, it needs to be able to sense and model emotions. This involves perceiving various communicative signals (body posture, facial expression, gaze, speech, etc.) from the human user and interpreting them. Many researchers have used various emotion models (Mehrabian, 1995; Russell, 1980) to interpret human emotions (Cavallo et al., 2018; Kirby et al., 2010; Paplu et al., 2022). For example, Kirby et al. (2010) developed an affective robot receptionist that mimicked human-like behavior by interpreting its interaction in terms of its emotions, mood and attitude. Paplu et al. (2022) used the circumplex model (Russell, 1980) to generate context appropriate emotions on a robot by appraising various communicative signals from the human interlocutor such as proximity, body postures, facial expressions and gestures. A recent study (Tang et al., 2023), explored the MAP-Elites (Cully et al., 2015) framework to generate emotional expressions automatically for a robotics platform they developed. While these models have shown good results in generating robot emotions, they involve building complex architectures (in some cases even hardware) that are effort and time intensive. Additionally, the models need to be fast enough to operate in real-time which is challenging in HRI. In this work, we limit the robot's emotions to a subset of the basic emotions (Ekman, 1999) (see Section 4.5.1).

Out of the many modalities of information that can be sensed and processed by a robot to generate emotions, dialogue plays a key role in providing the necessary context. The textual representation of a conversation can be analyzed using emotion classification algorithms to detect the emotions of various utterances. Emotion Recognition in Conversation (ERC) is a text classification task that aims to predict the emotions of the speakers during a conversation from the utterances. Static ERC refers to a task where a conversation has already taken place and utterance emotions are detected using both the historical and future contexts (Ghosal, Majumder, Poria, Chhaya, & Gelbukh, 2019; Lian, Liu, & Tao, 2021). On the other hand, real-time ERC refers to detecting utterance emotions relying only on the historical context (Jiao, Lyu, & King, 2020; Ma et al., 2022). Real-time ERC is very relevant in the context of HRI and can be used on-the-fly to appraise emotion of a conversation between a robot and a human. Various works have proposed to utilize ERC models for emotion recognition in HRI (Fu, Liu, Ishi, & Ishiguro, 2020; Rasendrasoa, Pauchet, Saunier, & Adam, 2022), however evaluations involving genuine interactions with robots have been notably scarce. We propose to appraise emotions as a real-time ERC task to generate emotions on a robot on-the-fly.

LLMs like GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and OPT (S. Zhang et al., 2022) have been trained on very large scale general text datasets (both dialogue and publicly available web documents). They have shown impressive capabilities in solving a variety of different tasks such as generating code (Chen et al., 2021), translation, and question-answering (Brown et al., 2020) by repurposing their learned knowledge. For example, Lammerse et al. (2022) applied GPT-3 to solve an ERC task that involved extracting emotions from interviews with children. LLMs have a great potential for application specifically in the field of HRI. The 'zero-shot' chatting capabilities of LLMs, such as GPT-3 Brown et al. (2020), have made designing interactions with robot very easy. Consequently, many works have tried to integrate LLMs to solve various HRI tasks (Axelsson & Skantze, 2023; Billing, Rosén, & Lamb, 2023; Irfan, Kuoppamäki, & Skantze, 2023). Billing et al. (2023) integrated GPT-3 as a verbal proxy on NAO and Pepper robots to model open-dialog interactions. In a recent work, Irfan et al. (2023) proposed guidelines for using LLMs to develop companion robots for older adults. Others have tried to repurpose LLMs to solve diverse HRI tasks. For example, Axelsson and Skantze (2023) developed an architecture for presenter robots (e.g., a museum guide) by using GPT-3 to access information from knowledge graphs. In this work, we use GPT-3.5 to generate robot emotions, moving beyond the domain of generating robot speech.

4.3 Emotion Generation Using LLMs

Emotion appraisal is a continuous process where humans continuously process the stimuli around them against a motivation system (Ellsworth & Scherer, 2003). Stimuli spanning across various modalities including verbal and non-verbal behaviors are processed during the appraisal process. In order to generate appropriate emotional responses for the robot in real-time, it is imperative that the computation time of the emotion appraisal process is minimized. Thus, we limited the scope of model input for this study to only the textual representation of the conversational context.

GPT-3 has been shown to perform strongly on various NLP tasks in a zero-shot fashion that need reasoning or adaptation on-the-fly (Brown et al., 2020). We wanted to harness these capabilities and generate ad-lib robot emotions. We first interpreted robot emotion appraisal as an ERC task. ERC takes the context of the conversation into account when detecting the emotions of utterances. As discussed in Section 4.2, LLMs have been shown to be effective in ERC tasks.

Hence, we propose to use GPT-3 for real-time ERC, that takes the dialogue context to account when detecting emotions. We selected GPT-3.5 (an updated GPT-3 LLM (Brown et al., 2020)) with the model *'text-davinci-003'* for our study. This was the best performing model from OpenAI when the study was conducted. While ChatGPT was faster and had been trained on more recent data, we found that the behavior was not as consistent as the *davinci* models for our tasks. GPT-4 (OpenAI, 2023) was announced later and the API was not available yet during the data collection.

We wanted to adapt real-time ERC as a prediction task that predicted the emotion for the robot by taking the immediate history of the conversation into account. For example, consider the following dialogue (R denotes the robot, P denotes the participant, Ux denotes the utterance number):

P: *What do you think about picture 1? I think it looks really cool!* (U1)

R: *The picture looks like a really beautiful painting to me. Such an amazing sight.* (U2)

A real-time ERC model could for example detect the emotion following U2 as 'Happy'. In our task, we wanted to do the same using GPT-3.5, i.e., to predict what could be an appropriate emotion for U2 based on the conversation history (U1 and U2 taken together). For this study, we restricted the emotions to a subset of the six basic emotions Ekman (1999) (see Section 4.5.1 for more details).

We also introduced an emotion category 'Neutral' that the model could predict. This represents instances during the conversation where there is no need to express any emotions. We expected GPT-3.5 to be able to detect them and predict the emotion category as 'Neutral' when there was no emotion expressed in the dialogue, even though an affective artifact (such as an affective image discussed in Section 4.5.1) was being discussed as the subject of the conversation. For example, in the following conversation, let's assume that the discussion is about the positioning of an affective image in an image sorting task. The robot's emotion was predicted to be 'Neutral' by GPT-3.5 even though the subject of the conversation was an affective image (R denotes the robot, P denotes the participant):

R: *What do you think?*

P: *I think you are correct in that assessment. I will put it here.* ⟨robot's emotion⟩

We inserted a delay of about 1 sec, before the robot said the next utterance (after U2 in the example). Doing so made it so that the facial expression was

Table 4.1: Hyperparameter values set in the API call to GPT-3.5 for this study.

Hyperparameter	Set Value
Maximum Length	1
Temperature	0.0
Top P	1.0
Frequency Penalty	0.0
Presence Penalty	0.0
Stop Sequence	)

displayed between the two utterances (U2 and the upcoming utterance) and the expression felt like a continuation of what had been discussed so far while moving to the next utterance. Additionally, introducing the delay also gave the robot sufficient time to send the API call and receive the predicted emotions. We acknowledge that a delay between two sentences where the robot just displays a facial expression is perhaps unnatural. However, this helped in exaggerating the emotions the robot wanted to express (see Section 4.5.4). As the generation time by GPT-3.5 gets faster in the future, thereby reducing the latency between API calls and responses, we can adapt the model to get the emotion while the robot says an utterance removing the need of any delays.

GPT-3.5 was instructed to perform the emotion prediction for the robot as a completion task with the help of a prompt. We used zero-shot prompting (Brown et al., 2020) for the task. The prompt was divided into two sections. The first section comprised of the task description. It was asserted that the conversation was between a robot and a human. As GPT is auto-regressive, i.e., the time taken to generate response is linearly correlated to number of tokens it has to generate, we restricted the output tokens to 1. Each emotion class was assigned a number, and GPT-3.5 was asked to output only the emotion class number at the end. The first half of the prompt looked like the following:

Prompt (Part 1)

"This an emotion classifier. The following is a conversation between a human and a robot.The robot's emotion is written within the brackets ().
The emotion can be either
'Happy (1)', 'Sad (2)', 'Fear (3)', 'Anger (4)', 'Surprise (5)' or 'Neutral (6)'.
Only give the emotion class number between 1 - 6"

The second section comprised of the actual conversational data that was to be used as the historical context for the prediction. Furhat can store the utterances

during an interaction (both the user's and its own) in the $DialogueHistory$ object. Furhat's and the user's utterances were extracted to construct the turn wise dialogue in the prompt. Lammerse et al. (2022) proposed a windowing approach to control the exact number of past dialogue exchanges to be used as context in ERC task and found that a window size of 3 resulted in the best accuracy for GPT-3. We introduced a variable named $contextWindowSize$ which specified the number of turns to be included as context in the prompt. For the user study (see Section 4.5), the optimal $contextWindowSize$ was found by conducting mock sessions while iterating through various window sizes. It was found that $sontextWindowSize$ of 2 resulted in the most appropriate responses from GPT-3.5. After including the turn wise dialogue history, the final element in the prompt was the emotion prediction part for the robot's emotion. This was done by including the text "Robot: (" as the last line of the prompt. This instructed GPT-3.5 to predict the class number. The second part of the prompt looked like the following:

Prompt (Part 2)
"Human: ⟨ utterance text from $DialogueHistory$ ⟩
Robot: ⟨ utterance text from $DialogueHistory$ ⟩
Robot: ("

An example of a complete prompt with $contextWindowSize = 2$ (two turns) would look like the following:

"This an emotion classifier. The following is a conversation between a human and a robot.The robot's emotion is written within the brackets ().
The emotion can be either
'Happy (1)', 'Sad (2)', 'Fear (3)', 'Anger (4)', 'Surprise (5)' or 'Neutral (6)'.
Only give the emotion class number between 1 - 6"
"Human: What do you think about picture 1? I think it looks really cool!
Robot: The picture looks like a really beautiful painting to me. Such an amazing sight.
Robot: ("

OpenAI API provides a list of hyperparameters that can be used to control the behavior of the model during an API call. As mentioned before, since we wanted to obtain faster output from the model, we set the '*Maximum Length*' to 1. '*Temperature*' was set to 0, to obtained consistent answers and eliminate any randomness. We also used the ")" as the 'Stop Sequence' which further fine tuned the output to only generate the emotion class number as the output token. Table 4.1

lists the hyperparameter values used for this study. Another aspect to consider when using GPT-3.5 for emotion generation is to determine the instance when emotions need to be predicted during a conversation. This can differ depending on the use case/ scenario. For our user study (see Section 4.5), we sent an API call every time the human participant asked the robot to share its opinions about the affective images in the game or when the robot asked the participant to share their opinions. Figure 4.3 shows the outline of the model used to generate robot emotions in the user study. It should be noted that $contextWindowSize$ and the model hyperparameters (see Table 4.1) might need to be optimized to find the ones that fit the best for other use cases/ scenarios.

4.4 Hypothesis

Similar to Lammerse et al. (2022), we applied GPT-3.5 to detect the emotions in conversation. However, a key difference was that we predicted the emotion for the robot based on the immediate conversational history as context. In order to successfully generate contextually appropriate emotional expressions for the robot, the system has to accurately predict the appropriate emotion, as well as generating and displaying the corresponding facial expressions on the robot's face. We verify the appropriateness of the robot's expressions by evaluating whether participants are able to recognize and interpret the expressions on the robot's face in such a way that they contribute to a more positive experience of the robot. This is done by contrasting a condition where the robot's emotions are generated by our model against two other conditions, where the emotions are either incongruent with the model's predictions, or where the robot does not display any emotions at all. We hypothesise:

- **H1:** Participants will have a more positive experience of a robot displaying context appropriate facial expressions, compared to the ones that do not.

Affective behavior of the robot is known to have an influence on the behavior of human participants(Gockley, Forlizzi, & Simmons, 2006; Kaushik & Simmons, 2022; Xu, Broekens, Hindriks, & Neerincx, 2014). Kaushik and Simmons (2022) used a sorting game where task was to learn the sorting rule based on the feedback provided by a robot. It was reported that affective robot behavior improved the sorting accuracy and lowered the perceived difficulty of the task. Based on this we hypothesise that:

- **H2:** Contextually appropriate emotion expressions by the robot will increase task performance.

4.5 Study: Affective Image Sorting Game

To evaluate if emotion appraisal using GPT-3.5 was effective and if the emotions expressed by the robot could be perceived correctly by users, we designed a within-subjects user study with three conditions. In the control condition (which we call the Neutral (N) condition), the robot did not express any facial expressions at all. Two experimental conditions were created: Congruent (C) and Incongruent (I). As the name suggests, in the Congruent condition, the robot displayed facial expressions that corresponded to the emotion GPT-3.5 had predicted (for example, if GPT-3.5 predicted "Happy" then the robot displayed a happy facial expression). In the Incongruent condition, the robot displayed facial expressions opposite of the emotions predicted by GPT-3.5. If the predicted emotion was negative (*Sadness, Fear, Anger, Disgust*), then the robot displayed a positive emotion (*Happy*). Similarly, the robot displayed a negative emotion (*Sadness*) when the predicted emotion was positive (*Happy, Surprise*). Only the robot's facial expressions were varied depending on the experimental condition, its face, voice and other non-verbal behaviors remained the same across conditions.

The following requirements were taken in to consideration while designing the study:

- The setup should be able to invoke emotional responses from the participants.
- The setup should not be too immersive or challenging for the participants.
- The setup should allow for free form conversation.
- The robot's expressions should be easy for the participants to notice

Based on these requirements, we decided to adapt the *Card Game* multi-party interaction setup (Skantze et al., 2015). The *Card Game* setup is a test-bed designed for studying single and multi-party interactions between a robot and human participants. It is a collaborative game where a touchscreen is placed between the robot and the human participants, on which a set of cards are displayed. The objective of the game is for the participants to rearrange the displayed cards in a specific order, while having a free form conversation about

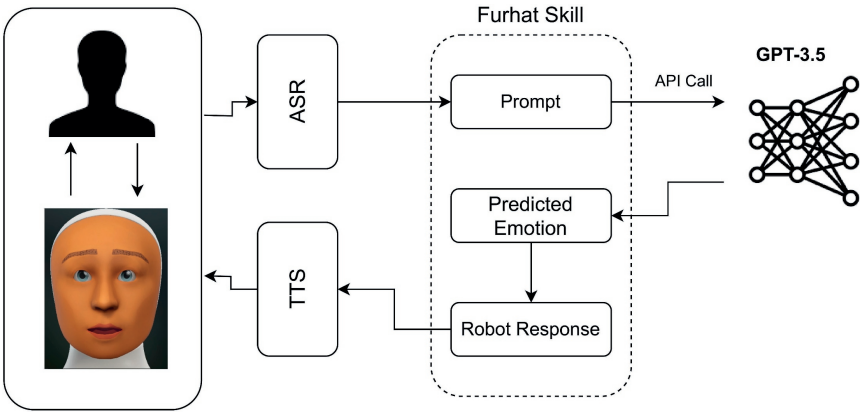


Figure 4.1: Outline of the model used in this study to generate emotions using GPT-3.5

the order and the cards both with the robot and among each other (in case of a multi-party setup).

We used a Furhat robot (Moubayed et al., 2013) for this study. It is a humanoid robot head with a back-projected face that allows it to display various facial expressions, brow movements, eye movements (e.g., , eye blinks, gaze), and head gestures (e,g., nodding, shaking). This enables the robot to convey emotions and engage in natural, human-like communication, providing a more immersive and realistic interaction experience for participants. Furhat provides a wide choice of realistic character faces and voices to choose from. For this study, we used the "default" character face (which is more cartoonish than photo-realistic) and Matthew neural TTS voice from Amazon Polly[2]. The character and voice were kept the same across experimental conditions.

A dyadic interaction setup was used where a Furhat robot and a human participant were seated face to face. A touchscreen was placed in between the robot and the participant such that the participant could move the images using their fingers and the robot could follow the images using head gestures and gaze (as shown in Figure 4.2). The interactions took place in a closed room where the participants were alone with the robot. The experimenter was present in an adjourning room where they could monitor the experiment.

In order to invoke emotional responses from the participants, a total of 45 affective images were used in the game (see Section4.5.1). The participants were tasked with sorting the images from the least positive image to the most

[2]Amazon Poly Voices

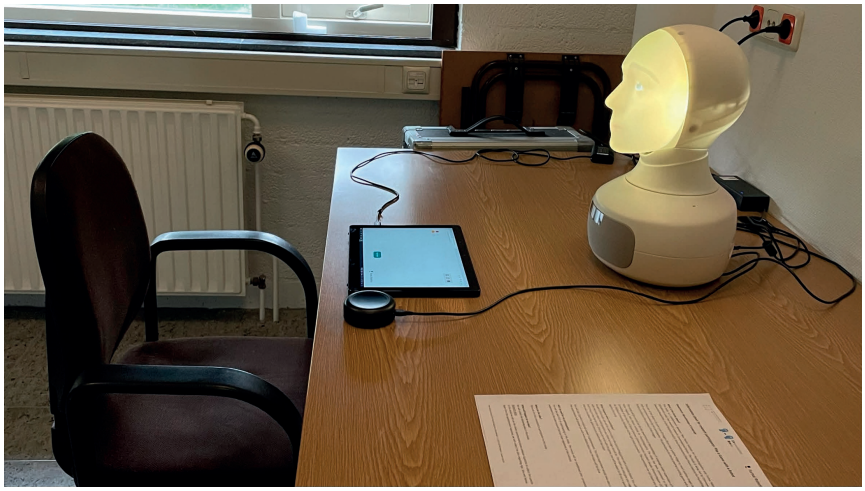


Figure 4.2: Experimental setup for the study

positive image based on the emotions they perceived from them. Each game comprised of 3 decks, with each deck having 5 affective images. The participants were instructed to play all the three decks for each game (irrespective of the order of the decks). Doing so provided more opportunities for the participants to observe the robot's behavior and counter the novelty effect of playing a game with a robot for the first time. Participants played a total of 3 games, 1 game for each experimental condition.

4.5.1 Affective Image Selection

Prior works in Psychology such as Lang, Bradley, Cuthbert, et al. (1999) have shown that emotions can be invoked in humans with the help of visual stimuli such as images. Consequently, there have been many works such as IAPS Subset (Mikels et al., 2005) and *DeepEmotion* (You, Luo, Jin, & Yang, 2016) that have developed datasets of images that are mapped to various emotions. As discussed briefly in the previous section, each deck in the game had 5 images in it and each condition had 3 decks, which means that we needed 45 images from the datasets belonging to 5 emotion categories. A key constraint was to avoid showing very disturbing images to the participants. Additionally, we wanted to have a good balance between positive and negative emotion categories in the game, so that it is easier for the participants to arrange them from least positive

to the most positive images. Thus, we decided to use the emotion categories *Happy/ Amusement, Anger, Sadness, Fear*, and *Awe/ Surprise*.

However, during the selection process, we could not find the required number of images for each category from any one dataset, either because there were not enough images for each category (for example, IAPS Subset had only 8 images for Anger) or because there were disturbing images that we could not use for our study (mainly for negative emotion categories like Fear). This led us to combine images from the IAPS Subset (Mikels et al., 2005) and *DeepEmotion* (You et al., 2016) datasets for each of the categories. We also added a few images from the internet that were suitable to be used in the experiment and were deemed to fit well for the emotion categories. From this pool of images for the 5 emotion categories, 45 images were handpicked to be used for the experiment.

4.5.2 Emotion Tagging Survey

The final pool of 45 images were a combination of images selected from the two datasets and images available online. While the images selected from the datasets for each emotion category had labels, the images from the internet were selected based on the authors perception. It is well known that the perceived emotion from visual stimuli is highly subjective in nature and varies from person to person (Machajdik & Hanbury, 2010). In order to ensure that the mapping between the emotion categories and images remained consistent, we conducted an online pilot study to map each of the selected images into an emotion category.

Qualtrics survey software was used to design the online survey. The participants were shown an image on the screen and asked to select the emotion category that matched the best with the image (exact question asked: *"Which emotion do you think the image depicts the most?"*). The 5 emotion categories were displayed as radio buttons. The order in which the images were shown to the participants were randomised to account for any order effect. Participants were recruited using notice boards and social media posts and did not take part in the later experiment with the robot.

We recorded the data from 21 participants (9 male, 11 female, 1 non-binary) with ages ranging between 19 and 48 ($M = 29.57$, $SD = \pm 7.55$). No compensation was offered for this survey. An image was assigned to an emotion category if the majority of the participants had selected that emotion for the image in the survey. There were cases where no clear selection emerged from the responses. In such cases, the images were tagged to be multi-class, i.e., belonging to multiple categories. However, for the image ordering game, it was necessary

to assign one emotion category per image. To do so, we decided the emotion category based on the original class the image belonged to as per the dataset it was taken from and the responses from the survey. For example, if image1 had "Happy" as its assigned emotion in the dataset, and the response from the survey was something like (0 participants selected Sadness, 1 Fear, 6 Anger, 7 Happy and 7 selected Surprise), then the final emotion category for image1 was selected to be "Happy".

4.5.3 Image Sorting Survey

After obtaining the emotion categories for all the 45 images, the images were divided into 9 groups which were to be used as decks for the sorting game. Each deck had one image from each of the emotion categories. Since the game assumes that there is a correct sorting order (i.e., least positive image to the most positive image), and this order is by nature very subjective, the emotion tagging survey was extended to also include an image sorting task. The outcome of this survey was used as the correct sorting order for the game.

Qualtrics survey software was used to design the sorting task. Participants were shown 5 images in the screen (1 deck) and asked to sort them from the least positive to the most positive image. The exact question asked to the participants was: *"Order the following images from Least Positive to Most Positive based on the emotion that you think is depicted in the image. You can drag & drop the images in the desired positions (1 to 5)."*. Each image position had a number displayed by the image and participants had to drag and drop the images to the right positions according to their judgement. The questions were always displayed with the 5 images placed in these positions in a random order.

The same participants who took part in the emotion tagging survey (see Section 4.5.2) were then asked to take part and complete the ordering survey. The final correct order of images in each deck was decided based on the order which most of the participants selected. These sorting orders were then used for the final scoring in the actual card sorting game that another group of participants played with the robot. The total score for the game was calculated based on the number of images that were placed in the right positions. The perfect score was 5 points where all the 5 images were placed correctly as per the results from the survey, and the lowest score was 0 (none of the images were placed in the right position).

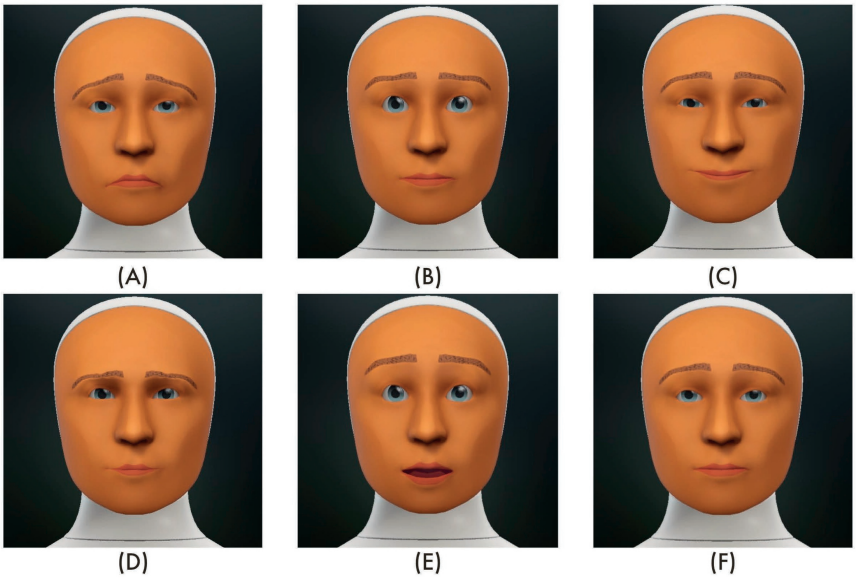


Figure 4.3: Facial expressions displayed by the robot in this study. The emotions depicted in each of the sub-figures are: (A) - Sadness, (B) - Fear, (C) - Happy, (D) - Anger, (E) - Surprise and (F) - Neutral.

4.5.4 Robot's Facial Expressions

An important consideration when designing the study was that the participants should be able to notice the robot's facial expressions easily during the game. In order to do so, two things were implemented. First, whenever the robot discussed the images or responded to what the participant had shared about the images, the cards on the display were turned translucent to make it difficult for the participants to see the images clearly. This was done to ensure that the participants' attention was not solely focused on the touchscreen during the game and that they looked at the robot's face. Second, it was decided to exaggerate the robot's facial expressions somewhat for each of the emotions. This was done to make a clear association between the facial expression displayed by the robot and the corresponding emotion category. Mäkäräinen, Kätsyri, and Takala (2014) concluded in their study that in order for humans to perceive a robot's emotion with a similar intensity as that of a human, the facial expressions should be exaggerated.

For each of the 5 emotion categories (see Section 4.5.1, the facial expressions of the robot ware implemented using the FACS (Facial Action Coding sys-

Table 4.2: Mapping of FACS Action Units (AU) to emotion categories used in the study

Emotion	Action Units (AU)
Amusement/ Happy	6 + 12
Sadness	1 + 4 + 15
Anger	4 + 5 + 7 + 24
Awe/ Surprise	1 + 2 + 5 + 26
Fear	1 + 2 + 4 + 5

tem) (Ekman & Friesen, 1978). FACS is a system developed to assign a common nomenclature to the the individual or group of muscles in the face that are fundamentally responsible for various facial expressions. These muscles were named Action Units (AUs) which are identified by a number in FACS. Ekman and Friesen (1978) provided a list of AUs mapped their corresponding muscle/ muscle group in the face. EMFACS (Emotional FACS) (Friesen, Ekman, et al., 1983) proposed a mapping between AUs and the six basic emotions (Ekman, Sorenson, & Friesen, 1969). There have been many works in HRI that have used FACS to interpret and generate communicative non-verbal behaviors such as facial expressions related to emotions (Auflem, Kohtala, Jung, & Steinert, 2022; Rossi, John, Taglialatela, & Rossi, 2022; Wu, Butko, Ruvulo, Bartlett, & Movellan, 2009). Furhat uses Apple's ARKit for its face model, so the corresponding ARKit parameters to FACS AUs were modified to generate the emotional facial expressions on the robot. Table 4.2 lists the mapping of AUs to emotions used for this study (adopted from E. A. Clark et al. (2020)). All the parameters were set to the maximum (i.e., 1) in order to exaggerate the expressions. Figure 4.3 shows the facial expressions for each emotion category used in this study.

4.5.5 Participants

We collected data from a total of 47 participants (22 males and 25 females). The responses from 4 participants were excluded from the analysis. One participant was 65 years old, which was beyond the predetermined age range of our experiment (18 - 60). The age of the participant was not known until after the experiment. The other three participants did not follow the instructions and focused only on the touchscreen throughout the experiment. The decision to exclude their responses was taken after observing their behavior during the experiment (from a separate room) and post experiment questions. The post ex-

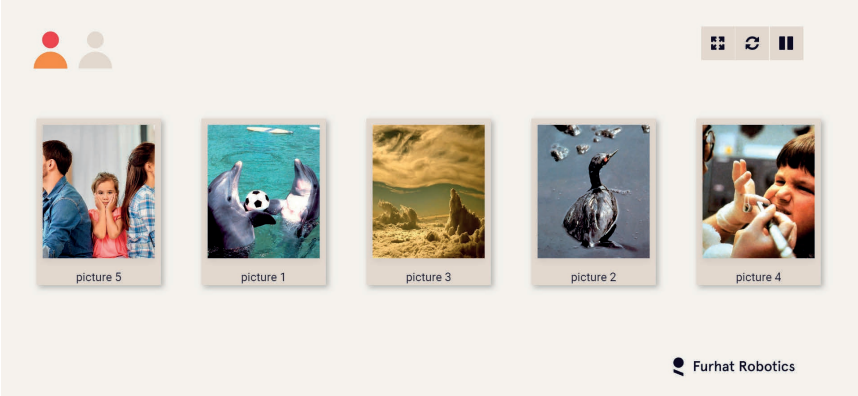


Figure 4.4: An example of a deck of affective images shown to the participants during the game

periment questions revealed that they had not been able to observe any behaviors on the robot's face in any of the conditions. The final pool of 43 participants (24 females, 19 males), whose responses were included in the analysis, had ages ranging from 20 to 59 ($M = 31.83$, $SD = \pm 9.91$).

Data collection took place in the labs at two places: Max Planck Institution for Psycholinguistics, Nijmegen (MPI) and KTH Royal Institute of Technology, Stockholm. For the data collection at MPI, the participants were recruited using the Max Planck institute's participant database[3]. A total of 22 participants (17 females and 5 males) were recruited at MPI. They were compensated €15 on completion. The recruitment at Stockholm was done using the participants recruitment website Accindi[4] and university notice boards. 21 participants (7 females and 14 males) participated in the study at Stockholm and were compensated with 100 SEK gift vouchers for their participation. All the participants spoke English. The study has received the approval by the ethics committee of the Faculty of Science, Radboud University, Nijmegen (reference no. ECSW-LT-2023-3-13-98066).

4.5.6 Process

As discussed earlier, the study followed a within-subjects paradigm. Each participant played 3 games with the robot, each game corresponding to one of the three experimental conditions (see Section 4.5). Each game comprised of 3 decks of

[3]MPI NL Participant Database
[4]Accindi

affective images. Participants were asked to play all the three decks (the order of decks was left for the participant to decide). Each affective image had a picture name displayed under it as shown in Figure 4.4. The participants could move the images by dragging them on the touchscreen. The order of games (experimental conditions) were balanced across participants. At the beginning of the experiment, while describing the experiment to the participants, the experimenter informed them about the technical limitations of the interaction, a few of which have been listed below:

- The robot could not hear the participants while it was speaking. The participants had to wait for the robot to finish speaking before they could speak.
- The participants had to use the exact names indicated below the images for the robot to understand which image they were referring to.

The experiment took approximately 45 minutes to finish. The experiment followed the steps given below:

1. The participants were given a description of the experiment, data management and compensation by the experimenter. They were also provided with an information sheet containing the same information. They were informed that the robot would provide them with the instructions on how to play the game and that the robot was a collaborator. The participants were instructed to have a discussion with the robot about their opinions regarding the positioning of the affective images. The robot's opinions may or may not be correct and that they were welcome to disagree with the robot

   A few examples were provided to give the participants an impression about the capabilities of the robot. For example, they were informed that they could ask the robot to comment on a specific image or compare two images. Additionally, they were informed that the robot could only discuss the images shown in the touchscreen.

2. The participants were informed that their task was to observe the behavior of the robot when it was discussing with them. They were asked to focus more on the robot during the interaction and not pay too much attention to the images on the touchscreen. Once they felt that the images had been arranged to their satisfaction, they could ask the robot to show the scores.

It was clarified that the scores were subjective and that they should not worry about the scores. This was done to ensure that the participants did not feel pressured to score better, as that could take their focus away from observing the robot's behavior during the game. The participants then provided their informed consent to participate in the experiment and data collection.

3. The experimenter left the room and initiated the game. The experimenter observed the participant through the robot's camera feed.

4. After the participant had finished playing all the three decks (1 game), the experimenter returned to the room and provided the participant with the questionnaire on an iPad. The questionnaire asked about the participant's impression of the interaction and the behavior of the robot. It comprised of 12 9-point Likert scale questions (see table 4.3). The order of questions presented to each participant was randomized to account for any order effect.

5. Once they had filled out the questionnaire, the experimenter collected the iPad and initiated the 2nd game, repeating step 3 and 4.

6. The same process was followed for the 3rd game as well. In addition to the 12 9-point Likert scale questions, the questionnaire also asked about basic demographic details such as age, gender and native language.

7. Finally, the participants were asked verbally to choose which game they thought was the best among the three games, and to provide a motivation for their choice. The exact question asked was *"Which game did you like the most out of the 3? Why did you like it?"*

4.5.7 Measurements

H1 pertained to the perception of robot's emotions through its facial expressions by the participants. To evaluate this, we collected subjective questionnaire data (Table 4.3) from the participants that asked them about their impression of the interaction with robot and the robot's behavior. The questionnaire had 12 9-point Likert scale questions that were further grouped into 3 dimensions (4 questions per dimension). *Positive Impression* ***D1*** comprised of questions that asked the participants about how positively they felt about their conversation with the robot. The questions under the *Emotion Perception* ***D2*** dimension tried

Table 4.3: Questionnaire used for subjective evaluation

Dimension	Question
Positive Impression (**D1**)	I enjoyed talking with the robot. My conversation with the robot flowed well. I felt positively about my interaction with the robot. I felt comfortable while talking to the robot.
Emotion Perception (**D2**)	The robot understood what *I* was talking about. The robot understood what *it* was talking about. The robot was able to understand and share my feelings. The robot felt emotions.
Human-likeness (**D3**)	The robot's face was human-like. The robot's behavior was human-like. Throughout the conversation, I felt like I could have been talking to a human. Throughout the conversation, robot's expressions were human-like.

to measure the perception of robot's emotion expressions by the participants. Finally, the *Human-likeness **D3*** dimension asked questions pertaining to how human-like the robot's behavior was. The responses were analyzed for each of the dimensions to see if one experimental condition was preferred over the others. The verbal responses of the participants for their preferred game was also included in the analysis.

To test **H2**, which predicted that congruent emotions would positively affect the task performance of the participants, we used the final score for each deck in the sorting game as a measure to evaluate task performance across the experimental conditions. The correct order for the affective images in each of the decks was obtained through the image ordering survey (see Section 4.5.3). During the sorting game, after each deck was sorted by the participants, the final order was scored between 0 to 5 and saved to a log file. A score of 5 (the perfect score) signified that the participant had arranged the images presented in the deck in same the exact order as the one obtained from the survey. A score of 0 signified that not a single image position arranged by the participants coincided with the image positions obtained from the survey.

4.6 Results

4.6.1 Questionnaire Data analysis

The responses to the 12 questions were analyzed to check the internal reliability of the questionnaire for the three dimensions. Cornbach's alpha was calculated to be 0.90, 0.88 and 0.93 for dimensions **D1**, **D2** and **D3** respectively, signalling good internal consistency. The responses were then analyzed for each of the dimensions to to see if participants rated one condition better than the others.

For dimension **D1**, the responses were analyzed through the use of an ANOVA test (using JASP (JASP Team, 2023)) to compare the effect of the experimental condition on the mean ratings. Results indicated a significant effect of experimental condition on the mean ratings by the participants ($F(2,513) = 11.40$, $p < 0.001$). *Post-hoc* Tukey's test were performed to obtain pair-wise comparisons of scores under each condition. It was found that participants rated the Congruent condition significantly higher than the Incongruent condition ($t = 4.67$, $SE = \pm 0.205$, $p < 0.001$). We did not find any significant difference between Neutral and Congruent conditions ($t = 1.47$, $SE = \pm 0.205$, $p = 0.305$). Participants also rated the Neutral condition higher than the Incongruent condition ($t = 3.20$, $SE = \pm 0.205$, $p = 0.004$).

Dimension **D2** asked questions that tried to measure the perception of the robot's emotions by the participants. ANOVA test results revealed a significant effect of the experimental conditions on the mean ratings by the participants ($F(2,513) = 17.24$, $p < 0.001$). Pair-wise comparisons using *post-hoc* Tukey's test showed that participants rated the Congruent condition significantly higher than both the Neutral ($t = 4.26$, $SE = \pm 0.234$, $p < 0.001$) and Incongruent ($t = 5.63$, $SE = \pm 0.205$, $p < 0.001$) conditions. This showed that participants were able to perceive the context appropriateness of the robot's facial expressions. We did not find any significant difference between the mean ratings for Neutral and the Incongruent conditions ($t = 1.36$, $SE = \pm 0.234$, $p < 0.36$).

Finally, dimension **D3** asked about the human-likeness of the robot's behaviors. An ANOVA test was conducted, which showed significant effect of the conditions on the ratings ($F(2,513) = 13.13$, $p < 0.001$). Using *post-hoc* Tukey's test it was found that participants perceived the robot as more human-like under the Congruent condition as compared to the Neutral ($t = 2.77$, $SE = \pm 0.216$, $p = 0.016$) and the Incongruent ($t = 5.14$, $SE = \pm 0.216$, $p < 0.001$) conditions. Neutral condition was also rated higher than the Incongruent condition ($t = 2.37$, $SE = \pm 0.216$, $p = 0.048$).

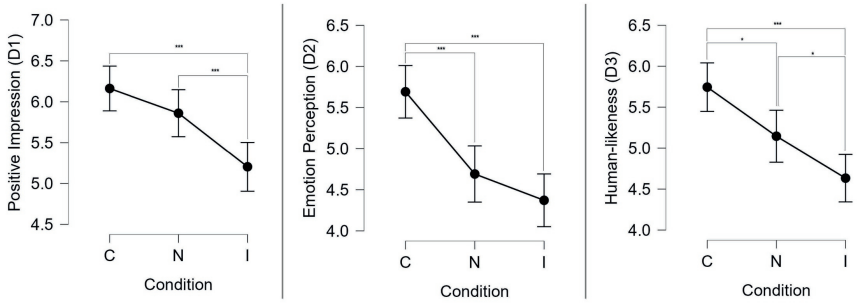


Figure 4.5: Mean ratings by the participants per condition for all the three dimensions in the questionnaire. *** denotes $p < 0.001$ and * denotes $p < 0.05$

A comparison of the mean ratings per condition for each of the dimensions is shown in Figure 4.5. The results supported hypothesis **H1**, which predicted that the participants would perceive a robot displaying context appropriate emotions as better than one that does not display emotions or one that displays incongruent emotions. To summarize the results from the questionnaire:

- The conversation left a more positive impression in the Congruent condition compared to the Incongruent condition.
- The emotions expressed by the robot were perceived to be significantly better in the Congruent condition compared to the other conditions.
- The robot's behaviors were perceived to be significantly more human-like in the Congruent condition compared to the other conditions.

We also analyzed the verbal responses from the participants to the post experiment question (see Section 4.5.7). Of the 43 participants recorded, 23 said that they preferred the Congruent condition, 15 preferred the Neutral condition, 3 preferred the Incongruent condition and 2 could not decide.

4.6.2 Sorting Task Score Analysis

As mentioned in Section 4.4, the robot's emotional expressions are known to have a influence on the final task performance. To verify this, we analyzed the scores participants obtained during the sorting game. For each participant, the sorting scores were retrieved for each experimental condition from the log files. An ANOVA test was performed to compare the effect of the three experimental

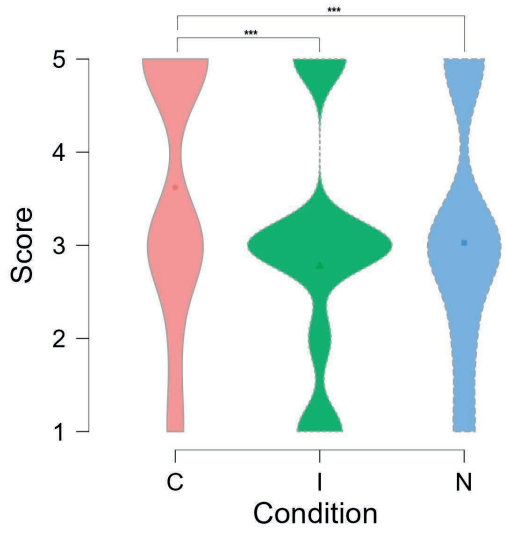


Figure 4.6: Sorting scores under different experimental conditions. *** denotes $p < 0.001$

conditions on the final sorting scores. Results indicated that there was a significant effect of experimental conditions on the mean sorting scores ($F(2, 448) = 14.53$, $p < 0.001$). *Post-hoc* Tukey's test revealed that the mean score in the Congruent condition was significantly higher than both the mean scores in the Neutral condition ($t = 3.67$, $SE = \pm 0.162$, $p < 0.001$) and the Incongruent condition ($t = 5.25$, $SE = \pm 0.161$, $p < 0.001$), as shown in Figure 4.6. We did not find any significant differences between the mean scores under the Neutral and Incongruent conditions ($t = -1.55$, $SE = \pm 0.162$, $p < 0.266$). This showed that task performance was positively affected by the contextual appropriateness of the robot's facial expressions, supporting **H2**.

4.6.3 Exploratory Analysis

We also wanted to see if any trends emerged through an exploratory analysis of the questionnaire response data. Additionally, we were interested to analyze the GPT-3.5 predictions during the interactions.

Effect of Condition Order

To evaluate the overall perception of the participants towards the robot's facial expressions, a GLMM (Generalized Linear Mixed Model) was fitted. The partic-

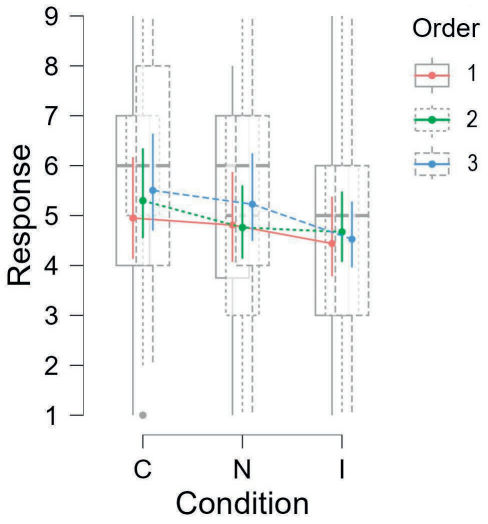


Figure 4.7: Participants' mean ratings per condition depending on the order they were presented.

ipants' ratings for all the question were used as the dependent variable. Experimental conditions and the order they were presented to each participant were used as the fixed effects variables. The participant id:s along with the question numbers were used as the random effects grouping factors. Inverse Gaussian family was used as the model family.

The model showed a significant main effect of experimental condition on the user ratings ($\chi^2(2) = 59.94$, $p < 0.001$). *Post-hoc* pairwise comparisons using Bonferroni correction showed that participants rated the Congruent (C) condition significantly higher than both Neutral (N) ($t = 4.94$, $SE = \pm 0.128$, $p < 0.001$) and Incongruent (I) ($t = 8.81$, $SE = \pm 0.128$, $p < 0.001$) conditions. Ratings for N were also significantly higher than the ratings for I ($t = 3.87$, $SE = \pm 0.128$, $p < 0.001$). This further supported hypothesis **H1** that predicted that participants will perceive a robot with context appropriate facial expressions better than the others.

We also observe an interaction effect between condition and order on the question ratings ($\chi^2(4) = 15.63$, $p = 0.004$). This suggested that the participants' ratings under each condition varied depending on the order in which the conditions were presented to them. Figure 4.7 shows the difference in participants' ratings per condition depending on the order. The order in which the conditions were presented to the participants followed the following sequence:

- Order 1 : $C \rightarrow N \rightarrow I$
- Order 2 : $N \rightarrow I \rightarrow C$
- Order 3 : $I \rightarrow C \rightarrow N$

It can be observed in Figure 4.7, that for *Order 1* and *Order 3*, the mean ratings were highest for Congruent condition, followed by the Neutral and Incongruent conditions. Where as, in *Order B*, even though Congruent condition was rated the highest, Neutral and Incongruent conditions did not have much difference. This might be attributed to the fact that in *Order 2*, participants first interacted under the Neutral condition where there were no facial expressions displayed by the robot. That followed with the Incongruent condition which had mismatched facial expressions so the ratings were still similar as compared to Neutral. Finally, the ratings increased when the robot expressed context appropriate expressions under the Congruent condition, which further shows that participants were able to perceive the robot's emotions and preferred the Congruent condition.

Impact of Location or Gender?

Since the data collection took place in two locations, Stockholm and Nijmegen (see Section 4.5.5), we were curious to see if location had any effect on the subjective ratings provided by the participants. A GLMM was fitted with participants' ratings as the dependent variable, and experimental conditions, order and the location as the fixed effects variables. The participant id:s and the question numbers were used as the random effects grouping factors. The inverse Gaussian family was used as the model family.

As expected, the model showed a significant main effect of experimental condition on the user ratings ($\chi^2(2) = 60.53$, $p < 0.001$). In addition to the interaction effect between condition and order, the model also showed interaction effect between condition and location ($\chi^2(2) = 6.91$, $p = 0.032$). This suggested that the participants' ratings under each condition also varied depending on the location where the experiment took place as shown in Figure 4.8. However, on further analyzing the participant distribution between the two locations, we observed that gender distribution at both the locations was very extreme. In Nijmegen, out of the 22 participants recorded, there were 17 females and 5 males. Where as, in Stockholm, out of the 21 participants recorded, there were 7 females and 14 males. This led us to wonder if the interaction effect that we observed earlier was due to gender instead of location.

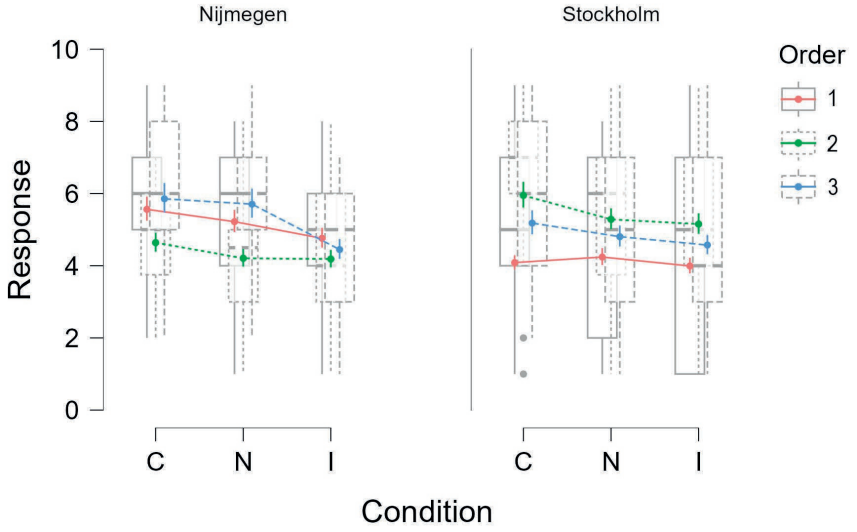


Figure 4.8: Effect of condition and location on subjective responses by the participants.

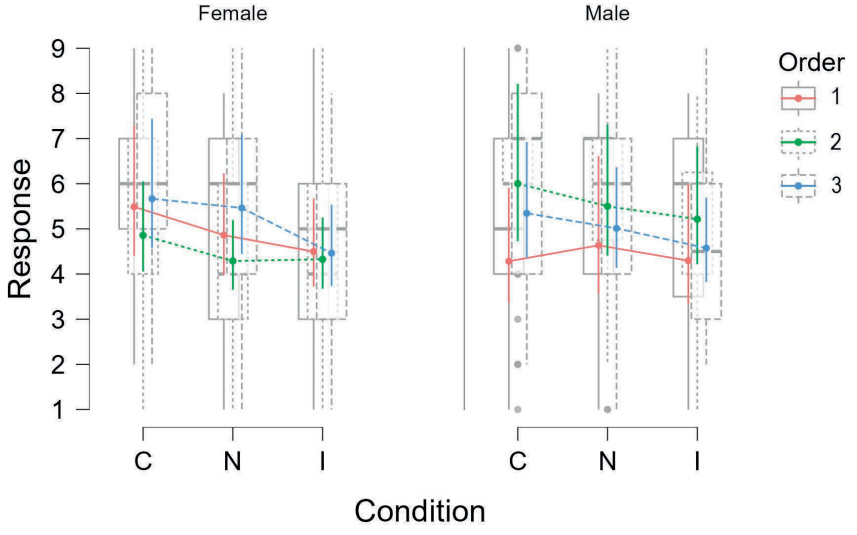


Figure 4.9: Effect of gender on the subjective ratings per condition

To investigate this, we fitted another GLMM with the same variables as the previous one, except that we swapped location with gender as a fixed effects variable. The model showed significant main effect of condition on the ratings as expected ($\chi^2(2) = 60.39$, $p < 0.001$), and also an interaction effect of condition and order. However, we also found out an interaction effect between condition and gender ($\chi^2(2) = 13.56$, $p = 0.001$). This indicates that ratings per condition were influenced by the gender of the participants as well, as shown in Figure 4.9. This was an interesting finding as it has been observed in prior studies that gender has an influence on the perception of emotional intelligence in robots (Chita-Tegmark, Lohani, & Scheutz, 2019). However, since we did not control for either gender or location, there might have been other factors that might have influenced this behavior. Further studies are needed to narrow down and verify any effect of gender or location on the perception of robot emotions.

GPT-3.5 Emotion Prediction

Results indicated that participants were able to perceive the context appropriateness of the robot's model-driven facial expressions. This implied that GPT-3.5 was able to reliably predict the emotions for the robot. In addition, we wanted to analyze the emotion predictions made by GPT-3.5 during the interactions, compared to the ground truth label for the picture being discussed (see Section 4.5.2). It should be stressed that this analysis is limited, given that the emotion appraisal label was not based on the image itself, but the preceding dialogue. Thus, the dialogue might in many cases express a different emotion or be neutral. Nevertheless, this analysis might give an overall idea of how often the emotion of the picture and the emotion appraisal aligned.

A prediction confusion matrix was calculated for each emotion category using the predicted vs. the actual image labels (ground truth), as shown in Figure 4.10. It can be seen that GPT-3.5 predicted aligned emotion categories consistently, with the best performance for 'Surprise' (65%) and worst for Anger (41%). Overall, GPT-3.5 predicted the emotion category to be 'Neutral' for about 17.6% of the cases.

4.7 Discussion & Limitations

The results suggest that the GPT-3.5 model was able to accurately predict the emotions for the robot's utterances across all the experimental conditions. This highlights the model's capability in generating contextually appropriate emo-

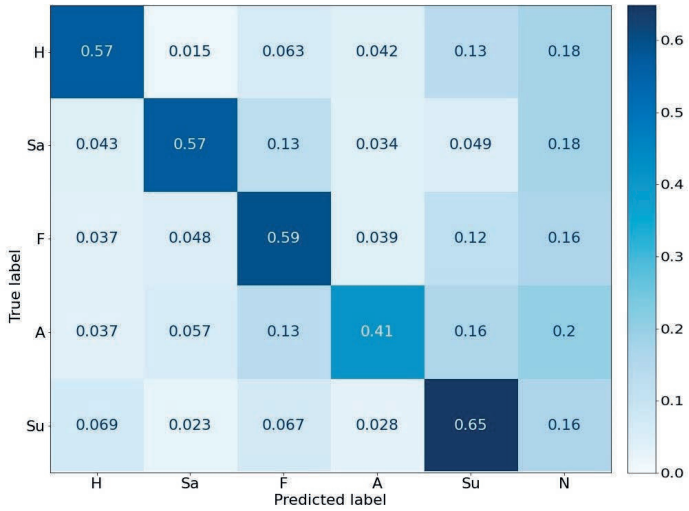


Figure 4.10: Normalized Confusion matrix between actual and predicted emotions by GPT-3.5. H - 'Happy', Sa - 'Sadness', F - 'Fear', A - 'Anger', 'Su' - 'Surprise', N - 'Neutral'

tional responses, which is crucial for effective and engaging human-robot interactions. Analysis of the questionnaire responses indicated that participants favored the Congruent condition over the other experimental conditions, as expected. The exploratory analysis of the responses further corroborated these findings. This preference for the Congruent condition suggests that emotional congruency between the robot's expressions and its verbal responses enhances user experience and perceived emotional authenticity, contributing to more positive interaction outcomes, which supports **H1**. Furthermore, ANOVA results revealed that participants achieved the highest sorting scores in the Congruent condition, followed by the Neutral and Incongruent conditions. This indicates that appropriate robot expressions positively influence task engagement and overall performance (**H2**), underscoring the significance of emotion-appropriate responses in facilitating effective human-robot collaboration.

We did not find any significant differences between the Neutral and the Incongruent conditions from the questionnaire responses. The post experiment verbal responses showed that participants occasionally attributed more complex meanings to the robot's emotions. For example, in the Incongruent condition when the robot displayed a happy facial expression when discussing a sad picture, one of the participants commented "I think the robot was feeling so sad that it was covering it by smiling. I do the same". In some cases participants also inferred

the facial expressions beyond the basic emotions used in this study (e.g., interpreting happy expression in the Incongruent condition as sarcasm). On the other hand, in the Neutral condition, due to lack of any facial expressions by the robot, there were no conflicting stimuli for the participants, which was perceived as appropriate behavior for a robot. We believe that these factors might have led to the lack of significant differences between the Neutral and the Incongruent experimental conditions. A recent study by H. H. Clark and Fischer (2023) argued that social robots are perceived by humans as a depiction of social agents. The emotions that the robot displays are perceived as not being felt by the robot, but by the character that the robot is portraying. This aspect warrants further exploration to better understand the human tendency to anthropomorphize robots and its implications on the perception of robots' emotions.

A technical limitation was that we occasionally observed a slight lag in the robot's expressions during the interaction. This was attributed to the API call during emotion generation. While the typical response time from the GPT-3.5 service was $\leq$ 1sec, it could in some cases take more than 2-4 seconds to receive a response, due to server lags, which delayed the emotion generation on the robot's face. In rare cases, GPT-3.5 was unable to return any response due to server overload. As cloud services continue to improve, such delays and errors are expected to diminish, leading to more seamless and natural interactions in real-time.

Even though GPT-3.5 predicted the emotions for the robot reliably, fine-tuning a model on more specific datasets may yield even better contextually relevant emotional responses. Additionally, while we restricted the emotions in this study to the basic emotions, participants attributed emotions beyond these basic categories to the robot's expressions. Future studies should incorporate a broader range of emotions to better align with human emotional complexity and facilitate more nuanced interactions. Another limitation is that the model could not generate long term emotional responses due to its context window size being restricted to just 2 past turns. While a larger window size could have taken more turns (there by more information) into the context, GPT-3.5 has a limit of 4097 tokens per prompt. This makes it very difficult to keep track of the events that have taken place during a prolonged interaction and use it to generate any long term emotions that may arise over time.

Finally, the current model utilized only the textual representation of the conversational speech for emotion generation in the robot. To develop a more holistic and multimodal emotion generation system, future research should consider

integrating other modalities, such as facial expressions and body language into the architecture. This would of course need advancement in LLMs that take multi-modal information as input. For example, GPT-4(OpenAI, 2023) is the latest model from OpenAI that is capable of taking text and images as inputs to generate text. As LLMs advance further, their applicability in modelling multi-modal emotion generation systems would become easier and more effective.

4.8 Conclusion

In this paper, we implemented a model to leverage LLMs for real-time robot emotion generation in HRI. By framing emotion appraisal as an ERC task, we utilized GPT-3.5 to accurately predict the emotions of a robot based on ongoing dialogue history. We conducted a within-subject user study to evaluate the effectiveness of the implemented model. The study was specifically designed to elicit emotional responses from the participants which made it possible to have an affective HRI. GPT-3.5 was found to be able to reliably predict context appropriate emotions for the robot. Results showed that participants perceived the Congruent condition to be significantly more human-like, emotionally appropriate and positive than the others, indicating that alignment between the robot's expressions and verbal responses significantly enhances the perceived emotional authenticity and overall positive interaction outcomes. Additionally, it was found that the participants scored the highest under the Congruent condition, further supporting the significance of emotion-appropriate responses in fostering effective human-robot collaboration.

This research explored the possibility of using LLMs in real-time HRI tasks beyond generating robot speech. Using cloud services and leveraging powerful pre-trained models to address complex HRI problems may be the next step forward. As language models and robotics technologies continue to evolve, our work contributes to the broader pursuit of creating more empathetic, socially-aware, and emotionally connected robots that seamlessly integrate into human environments, ultimately enhancing our everyday lives.

5 | The Influence of Human-likeness and Facial Regions on the Perception of Social Robot Emotions [1]

Abstract

The increased interest in developing next-gen social robots has raised questions about the factors affecting the perception of robot emotions. This study investigates the impact of robot appearances (human-like, mechanical) and face regions (full-face, eye-region) on human perception of robot emotions. A between-subjects user study (N = 305) was conducted where participants were asked to identify the emotions being displayed in videos of robots, as well as a human baseline. The results showed a positive correlation between human-likeness and better emotion recognition, suggesting the benefit of a human-like face for social robots. The recognition rates from eye-region were found to be comparable to full-face. These results offer insights for effective social robot face design in Human-Robot Interaction (HRI).

[1] Adapted from Mishra C, Skantze G, Hagoort P and Verdonschot R (2024) The Influence of Human-likeness and Facial Regions on the Perception of Social Robot Emotions. *12th International Conference on Affective Computing and Intelligent Interaction (ACII 2024)*. (Under Review)

5.1 Introduction

There has been a surge in the development of next-generation social robots. Numerous commercial entities have proposed their versions of general purpose robots, such as Optimus[2], GR-1[3], and Ameca[4]. While many new robots maintain a humanoid body design akin to NAO[5] and Pepper[6], the robot faces exhibit significant diversity, ranging from a highly human-like face in Ameca to a blank face design in Optimus. This calls for more research investigating how the design of the face affects the perception of social robots, and consequently the interaction humans will have with them.

Social robots, by definition, are designed to conduct human-like interactions (Hegel, Muhl, Wrede, Hielscher-Fastabend, & Sagerer, 2009). A key component of human communication is facial expressions which are used to convey meaning (Elliott & Jacobs, 2013), build relationships (Lazarus, 2006), and help in decision making (So et al., 2015). Prior studies suggest that our brains perceive robot facial expressions similarly to human expressions (Chammat et al., 2010; Craig, Vaidyanathan, James, & Melhuish, 2010). Thus, social robots must not only recognize human emotions but also be able to convey them. Modelling appropriate robot emotions is an active field of research. It has been found that robots expressing emotions are perceived as more intelligent (Gonsior et al., 2011) and trustworthy (Cominelli et al., 2021). However, it is equally important to investigate the factors influencing the perception of robot emotions. Identifying these factors would help design social robots that are easier to understand and interact with.

Researchers have investigated how robot facial expressions are perceived by humans based on robot form and appearance. An early study on emotion recognition with the Feelix robot found that adults recognize emotions in still images of the robot similarly to human faces (Cañamero & Fredslund, 2001). Breazeal (2003) obtained similar results, indicating that individuals were able to interpret the robot's facial expressions from both images and videos. Other studies have explored emotion recognition rates across various robot form factors, ranging from human-like (Becker-Asano & Ishiguro, 2011; Danev et al., 2017; Lazzeri et al., 2015) to non-humanoid (Beer et al., 2010; Cohen, Looije, & Neerincx, 2011). This leads to the first research question:

[2] https://en.wikipedia.org/wiki/Optimus_(robot)
[3] https://robots.fourierintelligence.com/
[4] https://www.engineeredarts.co.uk/robot/ameca/
[5] https://www.aldebaran.com/it/nao
[6] https://www.aldebaran.com/en/pepper

R1: *Does having a human-like face improve the recognition of a robot's emotions?*

The answer to this question is not clear from the literature. Beer et al. (2010) investigated this query in virtual agents, comparing recognition rates for human faces, synthetic human faces, and a non-human-like virtual agent (iCat). Their results indicated a higher recognition for the human face, followed by the synthetic human face, and lastly, the virtual agent. Chevalier, Martin, Isableu, and Tapus (2015) also reported that emotions in a female humanoid virtual agent (Mary) were better recognized than in Nao and Zeno. Lazzeri et al. (2015) and Becker-Asano and Ishiguro (2011) assessed humanoid android faces against human faces. Lazzeri et al. (2015) found robot facial expressions were on par with human expressions, while Becker-Asano and Ishiguro (2011) noted human emotions surpassed those of the Geminoid F robot. While these trends suggest that greater human-likeness enhances emotion recognition, this remains uncertain for human-like robot faces. Moreover, studies involving robots do not compare recognition between human-like and mechanical-looking robot faces, hindering clarity on human-likeness impact. Thus, we propose our first hypothesis:

H1: *Human-like robot faces yield better emotion recognition compared to mechanical-looking robot faces*

Another aspect to consider is the role of specific face regions in emotion recognition. This stems from the broad variation in robot face designs, resulting in diverse implementations of facial regions. For instance, robots like Nao and Pepper feature static faces devoid of human-like movements, while others, such as Fuahat (Moubayed et al., 2013) and Ameca, possess full-face designs with human-like movements across all facial regions. This leads to our second research question:

R2: *Is it necessary to model the entire robot face with intricate human-like movements, or could we focus solely on certain regions, like the eyes?*

This question not only sheds light on the significance of distinct facial regions in emotion recognition but also offers a chance to simplify robot emotion generation by reducing complexities. Previous studies in psychology show the significance of seeing full-face over specific facial regions in emotion recognition Baron-Cohen et al. (1997); Sullivan et al. (2007), however they also point to the fact that information for emotion recognition is not distributed evenly across the entire face. For example, studies suggest that the eye region alone provides sufficient information for emotion recognition (Baron-Cohen et al., 1997; Wegrzyn et al., 2017). Baron-Cohen et al. (1997) compared the emotion recognition from

pictures of the eye-region, mouth region and the full face. Their results indicated that eye-region was as informative as the full face for complex emotions. Real-world examples include animated characters like those in the movie WALL-E (e.g., WALL-E, EVE, MO), which use minimalistic eye expressions to convey emotions and meanings[7].

Insights from human emotion recognition studies form the basis to investigate modelling specific face regions (like eye-region) instead of the full face for social robot design. However, this aspect remains less explored in the literature, possibly due to limited platforms with capabilities for human-like facial and eye movements. For instance, social robots like Pepper and Nao feature static eyes-only designs, precluding comparisons of emotion recognition between eye-only and full-face expressions. Some studies have tried to evaluate emotion recognition from robots' eye expressions (Barrett, Weimer, & Cosmas, 2019; Kang & Park, 2021) and find the best ways to model them (Barrett et al., 2019; Chumkamon, Masato, & Hayashi, 2014; Greczek, Swift-Spong, & Mataric, 2011; Pollmann, Tagalidou, & Fronemann, 2019). In a study (Danev et al., 2017) on "animated faces" for the MASHI robot, researchers compared emotion recognition rates between full-face and eye-region expressions, finding that while the full-face yielded better recognition, eye-region expressions remained acceptable. However, these studies have been limited to either virtual characters or robots with limited expressive capabilities, such as Nao or Pepper. This leads us to our second hypothesis:

H2: *Full face expressions will lead to better emotion recognition compared to eye-region only*

To explore the impact of robot appearance and facial regions on emotion recognition, we conducted a between-subjects user study, comprising two online experiments. One experiment centered on full-face emotion recognition, while the other focused solely on the eye region. In both studies, participants were tasked with identifying emotions conveyed in video recordings featuring a human, a human-like robot, and a mechanical-looking robot.

[7] https://www.pixar.com/feature-films/walle

5.2 Materials and Methods

5.2.1 Robot Platform

We used the Furhat robot (Moubayed et al., 2013) for this study, a humanoid robot head featuring a 3D animated face projected onto a translucent mask via back projection. This setup enables the robot to adopt diverse appearances, spanning from realistic human-like to mechanical characters. Furthermore, Furhat can perform nuanced facial movements, resulting in human-like expressions. For the experiment, we chose two pre-installed characters: Hayden with a realistic human-like appearance, and Titan with a mechanical look (see Fig. 5.1). Titan gets its mechanical look from the square pupils, lack of eyebrows, white face color, and lines on the face that give the impression of its face comprising of different modular parts. Apart from these differences, Titan is able to express emotions similarly to the human-like face Hayden, as both of them share the same face model. This is in contrast to the mechanical faces that have been used in prior studies which had static eyes and mouths like Nao and Pepper. Thus, it is possible to directly compare recognition of the emotions expressed by the human-like and mechanical looking face using the Furhat robot.

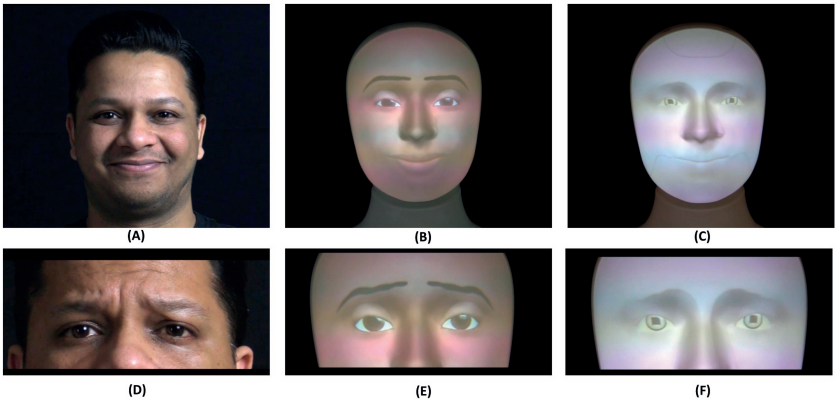


Figure 5.1: Emotional expressions displayed by the three characters. The first row depicts the full-face expressions of *Happy* by the human confederate (A), the human-like robot character Hayden (B), and the mechanical-looking character Titan (C). The second row depicts the eye-region expressions for *Sad* by the human confederate (D), Hayden (E), and Titan (F)

5.2.2 Robot Emotions

Facial Action Coding System (FACS) is a comprehensive and widely used system for describing and categorizing facial expressions based on the movement of individual facial muscles (Ekman & Friesen, 1978). The muscle movements, called Action Units (AUs), are assigned numerical codes to represent different facial expressions and emotions. FACS has been widely used in modelling robots' expressions in HRI (Barrett et al., 2019; Beer et al., 2010; Lazzeri et al., 2015; So et al., 2015; Stock-Homburg, 2022).

For this study, we modelled the six basic emotions (Ekman et al., 1969) on Furhat. Since Furhat employs Apple's ARKit parameters for its face model, we mapped the FACS AUs to their corresponding ARKit parameters to generate facial expressions. Table 5.1 shows the AUs used to generate the basic emotions (adopted from E. A. Clark et al. (2020)).

Table 5.1: Mapping of FACS AUs to emotion categories used in the study

Emotion	Action Units (AU)
Amusement/ Happy	6 + 12
Sadness	1 + 4 + 15
Anger	4 + 5 + 7 + 24
Awe/ Surprise	1 + 2 + 5 + 26
Fear	1 + 2 + 4 + 5

5.2.3 Experiment Setup

To evaluate our hypotheses (refer to Section 5.1), we conducted a between-subjects user study with a 3-way ANOVA design (2 face regions $\times$ 3 appearances $\times$ 7 emotions). The two face regions studied were full-face expressions and eye-region-only expressions. Three appearance conditions were defined: one with expressions by a human confederate (referred to as H) serving as the control, and the other two featuring emotions expressed by the robot characters Hayden (Ha) and Titan (Ti). These robot characters represented varying degrees of human likeness, allowing us to examine their impact on emotion recognition. Instead of images, short videos were used as stimuli for the user study. This was because still images capture only a snapshot of the emotion being expressed (Beer et al., 2010) and contain very little information about expressive posturing (Breazeal,

2003). We recorded videos of 6 emotions, *Happy, Sad, Anger, Surprise, Disgust,* and *Fear*, for each of the face types (H, Ha, and Ti), with a baseline *Neutral* expression (42 videos).

For **H1**, which pertained to the influence of the human-likeness of the robot's face on emotion recognition, we compared the recognition of emotions expressed by a human confederate, the human-like-looking robot character Hayden, and the mechanical-looking robot character Titan. The expressions of the confederate and robots were recorded in high resolution using the Canon HF-G30 video camera at the Max Planck Institute for Psycholinguistics, Nijmegen. The confederate was shown examples of images and videos of facial expressions using FACS before the recording. A total of 21 videos were recorded for all the emotions and appearance conditions (7 emotions $\times$ 3 appearance types).

H2 aimed to compare the recognition rates between the full-face and eye-region-only expressions. The video recordings of the full-face expressions for all three appearance types; human, Hayden, and Titan were cropped to the eye-region only. The cropped region and the proportion of the visible eye-region were kept consistent for all the videos. A total of 21 eye-region videos were extracted from the original full-face recordings.

Two online experiments were designed using the survey software Qualtrics. In the first experiment (full face condition), participants were shown short videos of the robots and the confederate on the screen and asked to select the matching emotion from the options provided on the screen. The experiment adopted a forced-choice paradigm, requiring participants to choose one of the 7 emotions displayed as radio buttons below the video. The exact question asked was: *"What emotion is being expressed in the video below?"*. Video presentation order was randomized, with a constraint to ensure that consecutive videos of the same appearance type did not occur more than twice in a row. The second experiment (eyes-only condition) followed a similar design but used the cropped eye-region videos as the stimuli. Each of the experiments took roughly 7-8 minutes to finish.

5.2.4 Participants and Procedure

We recruited a total of 305 participants via the online survey platform Prolific (`https://www.prolific.com/`). The first experiment involved 153 participants (77 males, 74 females, 2 non-binary, and 1 undisclosed), aged 18 to 59 ($M = 30.05, SD = \pm 8.23$). The second experiment collected data from 152 participants (76 males, 74 females, 2 non-binary) aged 19 to 54 ($M = 27.70, SD = \pm 6.66$), with no overlap between participants in both experiments. We imple-

mented two manipulation check questions; failing either resulted in automatic discarding of the survey response. Participants received 1 GBP upon successful experiment completion. The study has received the approval by the ethics committee of the Faculty of Science, Radboud University, Nijmegen (reference no. ECSW-LT-2023-3-13-98066).

5.3 Results

A response was counted as correct if it matched the intended emotion expressed in the video. JASP 0.17.3 software (JASP Team, 2023) was used for the statistical analysis. A three-way ANOVA was conducted to assess the impact of facial regions, appearances, and emotions on the correctness of participants' responses. Results indicated a significant main effect of the face regions ($F(1, 6279) = 114.28$, $p < 0.001$), appearances ($F(2, 6279) = 27.63$, $p < 0.001$), and emotions ($F(6, 6279) = 310.47$, $p < 0.001$) on the responses. We also observed significant interaction effects between face regions and appearance ($F(2, 6279) = 10.41$, $p < 0.001$), face regions and emotions ($F(6, 6279) = 46.87$, $p < 0.001$), and, appearances and emotions ($F(12, 6279) = 63.50$, $p < 0.001$) on the responses.

5.3.1 Effect of Appearances

Post-hoc Tukey's tests were performed to obtain pair-wise comparisons of recognition under each appearance type. It was found that participants recognized the emotions significantly better in the human face compared to the mechanical-looking Titan ($t = 5.76$, $SE = \pm 0.01$, $p < 0.001$). Emotion recognition was significantly better in Hayden than in the Titan face ($t = 6.95$, $SE = \pm 0.01$, $p < 0.001$). However, there were no significant differences between the recognition rates in the human face vs. Hayden ($t = 1.194$, $SE = \pm 0.01$, $p = 0.456$). Taken together, these results seem to support **H1**: A more human-like appearance leads to better emotion recognition.

However, on further analyzing the results for the interaction between appearances and face regions, it was found that these differences only held for the eye-region conditions. When looking specifically at this condition, there was a significant decrease in the recognition rate from human face to Titan ($t = 6.99$, $SE = \pm 0.01$, $p < 0.001$) and Hayden to Titan ($t = 7.53$, $SE = \pm 0.01$, $p < 0.001$), again in line with **H1**. However, we did not find any significant dif-

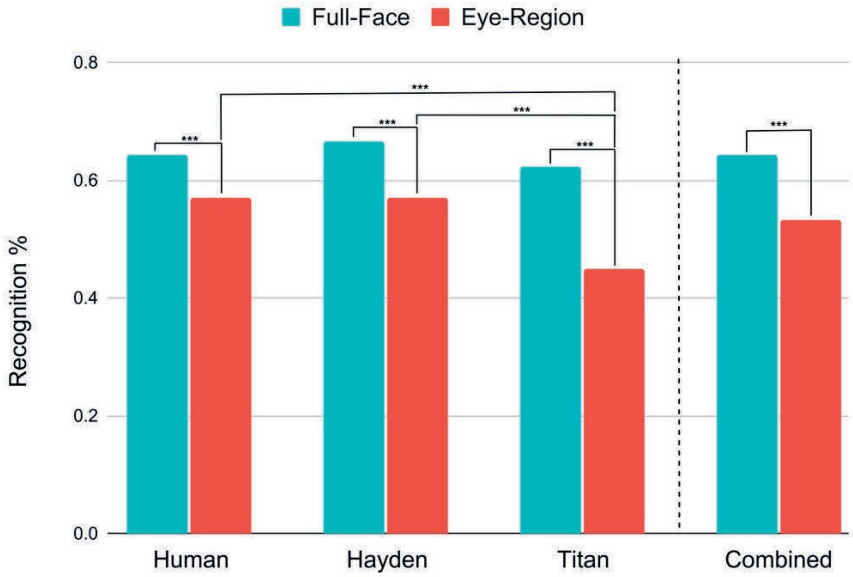


Figure 5.2: Emotion recognition score for each of the three appearances and the combined results for both the face region conditions: full-face and eye-region. *** indicates a significant difference with $p < 0.001$

ferences between the recognition rates for the full-face data based on the appearances (see Fig. 5.2). This indicates that the significant main effect of appearance is driven by the eye-region condition.

5.3.2 Effect of Facial Regions

Post-hoc Tukey's test revealed that participants recognized the emotions significantly better when they were shown the full-face videos as compared to the eye-region videos ($t = 10.69$, $SE = \pm 0.01$, $p < 0.001$). Overall, participants were able to recognize 64.4% of the emotions correctly when shown the full face of the robots. Recognition was 51.3% when only the eye-region videos of the robot were shown. Additionally, the eye-region recognition was higher than in a previous study with similar stimuli (49.1% in Barrett et al. (2019)). This supports **H2**, which predicted that emotion recognition from a full face should be better than just the eye region.

Further pair-wise comparisons between face regions and emotions were conducted using Tukey's tests. Figure 5.3 shows the confusion matrix with recog-

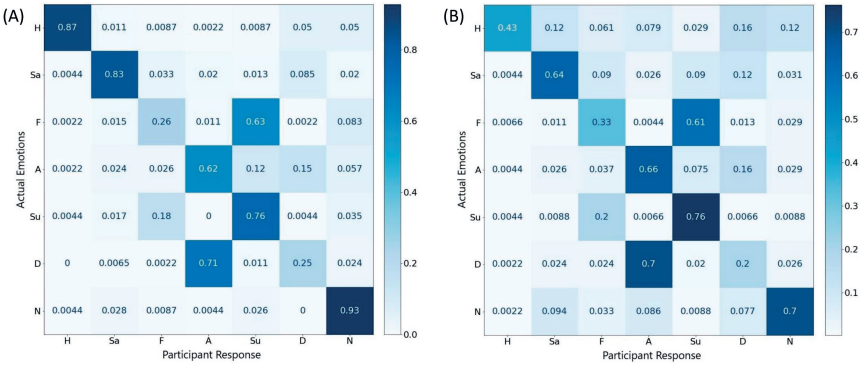


Figure 5.3: Normalized confusion matrix between actual and selected emotions by the participants under both the face region conditions. Sub-figure (A) depicts the confusion matrix for the full-face condition and (B) depicts the confusion matrix for the eye-region condition. Emotion abbreviation in the figure: H - *Happy*, Sa - *Sad*, F - *Fear*, A - *Anger*, Su - *Surprise*, D - *Disgust*, N - *Neutral*

nition accuracy for both face region types. It was observed that the recognition rates for *Fear, Anger, Surprise,* and *Disgust* were similar for both full-face and eye-region videos. Significant differences were found for *Happy* ($t = 16.33$, $SE = \pm 0.02$, $p < 0.001$), *Sad* ($t = 6.73$, $SE = \pm 0.03$, $p < 0.001$), and *Neutral* ($t = 8.37$, $SE = \pm 0.03$, $p < 0.001$), with higher recognition rates in full-face videos.

5.4 Discussion

This study aimed to explore how robot appearances and facial regions affect robot emotion recognition. ANOVA results highlighted the significant effects of appearances and facial regions on emotion recognition. *Post-hoc* analysis supported **H1**, indicating that greater human-likeness correlated with improved emotion recognition. Notably, emotion recognition rates showed no significant differences between the human and human-like robot faces (Hayden) in full-face videos, which further strengthens the advantage of a human-like face design. However, this difference only holds when the perception is limited to the eye region. This could be because, even though Titan is mechanical-looking, it still has a very expressive mouth region. In comparison with mechanical-looking robots like iCub, Furhat's Titan character appears more human-like when the full face is viewed.

On the other hand, a significant difference was found in emotion recognition favoring human-likeness for eye-region videos. This can be attributed to the video cropping, which obscured human-like features and emphasized mechanical aspects (see Section 5.2.1). The significant decrease in emotion recognition between full-face and eye-region for Titan's emotions is indicative of the same. A comparison of Hayden and Titan's eye-region videos sheds light on how mechanical appearance impacts emotion recognition. This raises questions about the role of eyebrows or pupil shape in the difficulty in recognizing Titan's emotions and whether having a mouth mitigates these effects, as full-face emotion recognition did not differ significantly. These questions could be investigated in a broader study. These questions warrant further exploration in broader studies. Our findings underscore the potential of human-like robot appearance for enhancing emotion recognition. Future research can explore diverse robot embodiments to address appearance variability among robot designs. In line with **H2**, the recognition rate for full-face was significantly higher than for the eye-region videos. *Post-hoc* analysis also supported this hypothesis, with significantly higher recognition for 3 emotions (*Happy, Sad, Neutral*) in the full-face videos. This is in line with previous findings which reported better recognition with full-face stimuli (Danev et al., 2017). However, it is worth noting that the difference in recognition for four emotions between eye-region and full-face responses was not statistically significant. This could point to the capability of the eye-region to express emotions sufficiently. Nonetheless, omitting a full face may result in the loss of valuable additional cues that could greatly help in emotion recognition. For example, we observe a significant decrease in the recognition of *Happy* when moving from full-face to eye-region. This could be attributed to the fact that the major cues for happiness lie in the mouth region (Wegrzyn et al., 2017). This needs to be kept in mind when deciding whether or not to model the full face when designing a social robot's face.

It was found that participants struggled to recognize *Fear* and *Disgust* for both facial region types (see Fig. 5.2), consistent with findings from a study using a virtual eye region model (Barrett et al., 2019). Additionally, participants often confused *Fear* with *Surprise* and *Disgust* with *Anger*. This could be explained by the Perceptual-Attentional Limitation Hypothesis which posits that the confusion between these emotions arises due to their shared muscle movements and visual similarities (Hendel, Gallant, Mazerolle, Cyr, & Roy-Charland, 2023; Roy-Charland, Perron, Young, Boulard, & Chamberland, 2015).

5.5 Conclusion

In this study, we investigate the influence of appearance and facial region on robot emotion recognition, with specific attention to the human-likeness of their appearance and the role of the eye-region. A comprehensive between-subjects user study was conducted with 305 participants. Results indicated that human-likeness improved participants' ability to recognize emotions in robots. Additionally, recognition rates from the eye-region, while not as effective as full-face, were found to be within a comparable range. However, it is essential to acknowledge that foregoing the modeling of the full face may result in the loss of crucial cues for certain emotions, as exemplified by the significance of mouth cues in recognizing happiness. Our study provides insights into the design principles of social robots and underscores the importance of considering human-like features for effective emotion communication.

6 | Discussion and Conclusion

This dissertation aims to model both the gaze and affective behaviors of social robots while examining how humans perceive them. This concluding chapter provides a concise overview of the results from the studies presented in the dissertation, discusses their limitations, and reflects on their significance. Additionally, it also outlines potential future directions and offers overall conclusion.

6.1 Summary of Results and Discussions

Chapter 2 primarily focused on modeling a comprehensive Gaze Control System (GCS) for social robots, addressing **RQ1** (see Section 1.2). A Planning-based GCS was proposed which planned the gaze behavior of the robot into the future, effectively coordinating its eye-head movements during gaze shifts. The gaze plan evolved as the conversation progressed as opposed to having a fixed plan decided at the beginning of an utterance. This resulted in dynamic gaze behavior contingent on the ongoing conversation. The proposed architecture was evaluated by comparing it to a reactive GCS in a user study. The study involved a multi-party card sorting game where two participants collaborated with the robot in a sorting activity. This setup encouraged spontaneous interaction between the participants and the robot. The results indicated that participants found the planning-based GCS significantly more *interpretable*, better at *intimacy regulation*, and *preferred* over the reactive GCS. These findings underscored the advantages of planning a robot's gaze behavior and dynamically determining the duration of the robot's gaze at specific targets, aspects that had not been previously addressed in GCS. Additionally, a recent study by (Haefflinger et al., 2023) corroborated our findings on the benefit of independently controlling the eye and head movements of the robot during an interaction.

However, no significant differences were observed in dimensions related to *awareness*, *turn-taking*, and *human-likeness*. This may be attributed to three influencing factors: the impact of cognitive load, limited sound source localization capabilities on the robot, and the novelty effect. Consistent with the *Load theory* (Lavie et al., 2004), participants might have encountered difficulties in perceiv-

ing subtle gaze behaviors, especially in cognitively demanding situations. Erroneous sound source localization led to the robot directing its gaze towards the wrong speaker, which would have impacted the perception of *turn-taking* dimension. Moreover, since many participants interacted with a robot for the first time, it would have split their attention in acclimatizing themselves to interacting with a robot, potentially affecting their ratings.

To gain a better understanding of the perception under these dimensions, future studies could involve showing the video recordings of the games to third-party observers. Alternatively, a different approach might involve designing simpler interactions that specifically target individual dimensions. Such measures would alleviate cognitive load and facilitate the perception of subtle gaze cues during HRI. The proposed GCS followed a heuristic approach to modeling robot gaze behavior; however, a data-driven approach might offer a more accurate representation of human gaze, provided that suitable data is available (which presents a challenge). An architecture that could integrate both data-driven and heuristic approaches could, in theory, result in more finely tuned and human-like gaze behavior for robots. A potential approach to implementing such an architecture could involve high-level decision-making through a heuristics approach (similar to the proposed GCS) with control subsequently transitioning to a data-driven approach (e.g., a deep learning model), governing low-level decisions such as eye-head movements.

Chapter 3 centered on investigating **RQ2**, which pertained to the influence a robot's gaze behavior might have on human gaze behavior. A within-subjects user study featuring two experimental conditions, namely, *Fixed Gaze* and *Gaze Aversion* (as detailed in Section 3.5), was designed to specifically examine gaze aversion behavior. The robot's gaze aversion behavior was automated using the GCS implemented in **Chapter 2**. Analysis of the gaze data collected from participants using an eye tracker revealed that participants averted their gaze for longer durations and more frequently when the robot did not avert its gaze from them. This observation aligned with the *Equilibrium theory* (Argyle & Dean, 1965), which posited that an increased gaze directed towards an interlocutor would be counterbalanced by increased gaze aversion by the interlocutor. Further analysis of the data indicated that participants exhibited increased gaze aversion just prior to speaking while showing lower levels of gaze aversion while speaking. Both of these findings carry significance as they provide evidence that human gaze behavior in HRI is influenced in a manner similar to HHI. It is important to ascertain whether the various gaze cues exhibited by robots exert a comparable

influence on humans as human gaze cues, as this understanding can inform the design of HRI that can better adapt to human behavior. An exploratory analysis of the data indicated the potential influence of topic intimacy on participants' gaze aversion, where participants averted their gaze more as the topic intimacy of the questions increased. This was interesting because this change in participants' gaze behavior was in response to questions posed by a robot. However, because the study did not control for the order of questions nor for the topic intimacy, a future study could explore these areas further.

Subjective analysis of the questionnaire responses indicated that participants perceived the *Fixed Gaze* condition as more human-like. This might be attributed to an initial unnatural gaze behavior displayed by the robot due to a technical limitation, where it directed its gaze randomly even when the participant was in front. Future studies could be designed to mitigate this issue effectively. Another limitation of the study pertained to the lack of gender balance in the dataset; all participants were male. Additionally, the robot characters used in the study were also male. While a recent study suggested that gaze aversion behavior might be gender independent (Acarturk et al., 2021), further investigations with a more diverse participant pool and a variety of robot characters would be essential to validate this effect in general. Another avenue for exploration is the potential influence of culture on gaze aversion during HRI. Does a robot's ethnic appearance influence the interpretation of perceived gaze behavior by humans? Do individuals from different cultures respond differently to a robot's gaze? These aspects could serve as potential subjects for future studies.

Chapter 2 and **Chapter 3** focused on the gaze behavior of social robots from both the modeling and the perceptual aspects respectively. The proposed GCS planned the robot's gaze behavior using various inputs from the conversational context. A future study could extend the capabilities of the GCS by incorporating human interlocutors' gaze aversion behavior as one of the inputs for planning robot gaze behavior. This would result in the generation of a more adaptive and context-appropriate gaze behavior that not only capitalizes on explicit cues like user speech, pointing gestures, and the movement of objects of interest, but also takes into account involuntary behaviors such as gaze aversion.

Chapter 4 explored two aspects of affective behavior in robots. The first objective was to examine the feasibility of harnessing Large Language Models (LLMs) in reliably modeling a robot's affective behavior, addressing **RQ3**. This entailed utilizing GPT-3.5 to assess emotions based on ongoing dialogue history and subsequently generating corresponding robot emotions. The second objective was

to verify if humans are able to discern the context-appropriateness of expressions exhibited by robots. To do this, a within-subjects user study was designed to investigate the reliability and context-appropriateness of emotions generated by GPT-3.5. The study involved participants engaging in an affective-image sorting game with the robot, comprising of three experimental conditions. In the *Neutral* condition, which served as the control condition, the robot did not exhibit any facial expressions. The robot displayed the emotions predicted by GPT-3.5 in the *Congruent* condition while the opposite emotions were displayed in the *Incongruent* condition. The game was specifically designed to evoke emotional responses from the participants. The analysis of subjective questionnaire responses indicated that participants found the emotions expressed by the robot to be significantly better in the *Congruent* condition. Participants also found the conversation to be significantly more human-like and leave a positive impression in the *Congruent* condition. These findings imply participants' capability to perceive context-appropriate robot expressions and the reliability of GPT-3.5 in predicting them. Moreover, participants achieved the highest scores in the *Congruent* condition, showing the positive influence of context-appropriate robot emotions on the effectiveness of collaborative tasks in HRI.

An interesting finding from the user study was the absence of significant differences in ratings between the *Neutral* and *Incongruent* conditions. This may have been because, at times, the participants attributed more complex meanings to the robot's emotions. For example, a smile displayed during a sad topic (incongruent) was interpreted as a masking smile intended to conceal the robot's underlying sadness. This underscored the constraints of using only the basic emotions Ekman (1999) in the user study, as human interactions involve many complex emotions. A key limitation of the implemented model was its sole reliance on textual representations of conversational speech for robot emotion generation. The emotion appraisal process in human communication entails the assessment of multi-dimensional inputs during a conversation. Future studies could explore the utilization of multi-dimensional LLMs, such as GPT-4, for robot emotion generation. For example, the current study could be extended by using the affective image being discussed as an input to GPT-4 along with the conversation text. Additionally, participants' facial expressions could also be integrated as an input parameter. Another aspect to consider could be the prospect of generating the robot's responses using LLMs, rather than crafting them manually.

While, using LLMs such as GPT-3.5 or GPT-4 offers a suitable alternative to real-time emotion appraisal using their cloud services, it is essential to bear in mind that these models operate as black boxes, lacking the guarantee of consistently producing the same outcomes. Implementing LLMs for Natural Language Generation tasks, such as engaging in direct conversations with users in an HRI context, subjects the system to the same limitations inherent to LLMs in text-based chats (e.g., hallucinations and bias due to training data). Until the inconsistencies and limitations intrinsic to LLMs are adequately addressed, the integration of LLMs in HRI systems should be a decision made after careful consideration of the advantages and drawbacks.

Building upon the insights from **Chapter 4**, **Chapter 5** addresses **RQ4** and delves deeper into human perception of robot expressions by investigating the factors influencing the recognition of robot expressions. Specifically, the influence of a robot's appearance and distinct facial regions on how humans recognize the emotions conveyed by the robot was studied with the help of a between-subjects online user study. The study involved the presentation of short videos to the participants featuring two robot characters displaying basic emotions: one characterized by a realistic, human-like face, and the other with a more mechanical and artificial appearance. These videos were recorded to ascertain whether the human-likeness of a robot's appearance positively affected the recognition of the robot's facial expressions. As a baseline for comparing emotion recognition rates, video recordings of a human confederate displaying the same emotions were used. To verify the impact of facial regions on emotion recognition, the videos were cropped to solely highlight the eye-regions. The results indicated a significant effect of both the robot's appearance and facial region on the recognition of emotions. It was found that the emotions of the human-like robot character were better recognized than that of the mechanical-looking robot character. Moreover, no significant differences were found in the recognition rates between the expressions by a human-like robot character and the human confederate. These observations point out the positive correlation between the human-likeness of a robot's appearance on the recognition of its expressed emotions. Moreover, it was found that full-face videos resulted in better emotion recognition rates as compared to the videos showing only the eye-region.

A more in-depth examination of the data unveiled that the significant difference in emotion recognition linked to human-likeness primarily stemmed from the data associated with the eye-region. The emotions conveyed through the eye-region videos of the human-like robot character were notably better recognized

than those of the mechanical-looking one. This observation suggests several potential factors that may have influenced the outcomes. The mouth region of the mechanical-looking robot face was equally as expressive as the human-like robot face, which likely facilitated emotion recognition. In contrast, in the eye-region videos, the mechanical-looking robot face displayed certain distinct features, such as the absence of eyebrows and unnatural square-shaped pupils, which might have contributed to reduced emotion recognition. This raises the question of whether the reduced recognition rate resulted from the unconventional pupil shape, the absence of eyebrows, or a combination of both factors. Furthermore, it prompts consideration of whether possessing an expressive mouth is adequate to counterbalance the adverse effects of missing eyebrows and unconventional pupil shapes. On a different note, although the analysis revealed a significant increase in the recognition rates when participants viewed full-face videos compared to eye-region videos, no significant differences were observed in the recognition rates for the four emotions. This could indicate that designing robot faces with just expressive eyes might be sufficient to convey emotions. However, it is essential to acknowledge that numerous facial expressions rely significantly on cues situated around the mouth region, which would be forfeited in the absence of a full-face design. These aspects warrant further exploration to establish robust design principles for modeling the faces of social robots.

6.2 General Conclusion

The four research studies presented in this dissertation provide insights into the modeling and perception of gaze and affective behaviors in HRI. In response to the research questions posed within this dissertation, these studies collectively address various aspects of HRI, shedding light on the significance of planning in gaze control, the relationship between robot and human gaze behavior, the potential of LLMs for real-time emotion generation, and the influence of appearance and facial features on emotion recognition in robots. The first study introduced a novel planning-based GCS, showcasing its significant advantages in terms of interpretability and intimacy regulation when compared to reactive systems. The findings underscored the importance of considering planning as a crucial aspect of gaze control that has been previously overlooked. The second study investigated the influence of robot gaze behavior on human gaze behavior, showcasing the trade-off relationship between robot and user gaze behavior.

This study offered an original contribution by identifying a direct relationship between robot gaze and human gaze behavior.

The third study focused on leveraging the capabilities of LLMs and integrating them into an appraisal system for real-time robot emotion generation. The findings demonstrated that aligning the robot's expressions and verbal responses significantly enhanced emotion perception and overall interaction outcomes, highlighting the potential of LLMs to extend beyond generating robot speech in HRI tasks. Finally, the investigation into the role of appearance and facial regions in robot emotion recognition revealed the impact of human-likeness in improving emotion recognition in robots. While recognizing emotions from the eye-region proved effective, this study underscored the importance of considering a full-face in designing social robots.

These collective findings contribute to the broader objective of creating empathetic, socially aware, and emotionally connected robots capable of exhibiting human-like gaze and affective behaviors. Robots with such human-like non-verbal behaviors would make the interactions significantly better and richer. It is worth noting that non-verbal behaviors, such as eye contact, facial expressions, and gestures, are deeply ingrained in human communication. These behaviors allow us to convey emotions, intentions, and understanding in a nuanced and intricate manner. When robots are equipped with the ability to perceive and reciprocate these non-verbal cues, they bridge the gap between human and robot interaction, facilitating a more natural and seamless integration of robots into human environments. To summarize, the four main contributions of this dissertation are:

- A novel GCS to plan the robot's gaze into the future and coordinate its eye-head movements.
- The first to verify a direct influence of a robot's gaze behavior on human gaze behavior.
- The first implementation to leverage LLMs in generating a robot's affective behavior.
- One of the first studies to investigate the influence of a robot's appearance and facial regions on the recognition of its emotions.

References

Abele, A. (1986). Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior*, *10*(2), 83–101.

Acarturk, C., Indurkya, B., Nawrocki, P., Sniezynski, B., Jarosz, M., & Usal, K. A. (2021). Gaze aversion in conversational settings: An investigation based on mock job interview. *Journal of Eye Movement Research*, *14*(1).

Admoni, H., Bank, C., Tan, J., Toneva, M., & Scassellati, B. (2011). Robot gaze does not reflexively cue human attention. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, pp. 1983–1988). Boston, MA: . Austin, TX: Cognitive Science Society.

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, *6*(1), 25–63.

Andersen, P. A. (1999). Nonverbal communication: Forms and functions. *(No Title)*.

Andrist, S., Mutlu, B., & Tapus, A. (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* (pp. 3603–3612). ACM.

Andrist, S., Tan, X. Z., Gleicher, M., & Mutlu, B. (2014). Conversational gaze aversion for humanlike robots. In *Proceedings of the 9th acm/ieee international conference on human-robot interaction (hri)* (pp. 25–32). IEEE.

Argyle, M., & Cook, M. (1976). Gaze and mutual gaze. *British Journal of Psychiatry*, *165*, 848–850.

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*(3), 289–304.

Auflem, M., Kohtala, S., Jung, M., & Steinert, M. (2022). Facing the facs—using ai to evaluate and control facial action units in humanoid robot face development. *Frontiers in Robotics and AI*, *9*, 887645.

Axelsson, A., & Skantze, G. (2023). Do you follow? a fully automated system for adaptive robot presenters. In *Proceedings of the 2023 acm/ieee international conference on human-robot interaction* (pp. 102–111).

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a “language

of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. *Visual cognition*, *4*(3), 311–331.

Barrett, S., Weimer, F., & Cosmas, J. (2019). Virtual eye region: development of a realistic model to convey emotion. *Heliyon*, *5*(12).

Beattie, G. W. (1981). A further investigation of the cognitive interference hypothesis of gaze patterns during conversation. *British Journal of Social Psychology*, *20*(4), 243–248.

Beattie, G. W. (2010). Sequential temporal patterns of speech and gaze in dialogue. In A. Kendon (Ed.), *Nonverbal communication, interaction, and gesture* (pp. 297–320). De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110880021.297 doi: doi:10.1515/9783110880021.297

Becker-Asano, C., & Ishiguro, H. (2011). Evaluating facial displays of emotion for the android robot geminoid f. In *2011 ieee workshop on affective computational intelligence (waci)* (pp. 1–8).

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2010). Recognizing emotion in virtual agent, synthetic human, and human facial expressions. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 54, pp. 2388–2392).

Billing, E., Rosén, J., & Lamb, M. (2023). Language models for human-robot interaction. In *Acm/ieee international conference on human-robot interaction, march 13–16, 2023, stockholm, sweden* (pp. 905–906).

Binetti, N., Harrison, C., Coutrot, A., Johnston, A., & Mareschal, I. (2016). Pupil dilation as an index of preferred mutual gaze duration. *Royal Society Open Science*, *3*(7). (id: 160086)

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PloS one*, *3*(8), e3022.

Boucher, J.-D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., ... Ventre-Dominey, J. (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.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International journal of human-computer studies*, *59*(1-2), 119–155.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burgoon, J. K. (1985). Nonverbal signals. *in M. L Knapp & G. R. Miller (Eds.),*

Handbook of interpersonal communication, 344–390.

Cañamero, L., & Fredslund, J. (2001). I show you how i like you-can you read it in my face?[robotics]. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, *31*(5), 454–459.

Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., & Dario, P. (2018). Emotion modelling for social robotics applications: a review. *Journal of Bionic Engineering*, *15*, 185–203.

Chammat, M., Foucher, A., Nadel, J., & Dubal, S. (2010). Reading sadness beyond human faces. *Brain Research*, *1348*, 95–104.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., ... others (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chevalier, P., Martin, J.-C., Isableu, B., & Tapus, A. (2015). Impact of personality on the recognition of emotion expressed via human, virtual, and robotic embodiments. In *2015 24th ieee international symposium on robot and human interactive communication (ro-man)* (pp. 229–234).

Chita-Tegmark, M., Lohani, M., & Scheutz, M. (2019). Gender effects in perceptions of robots and humans with varying emotional intelligence. In *2019 14th acm/ieee international conference on human-robot interaction (hri)* (pp. 230–238).

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... others (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Chumkamon, S., Masato, K., & Hayashi, E. (2014). The robot's eye expression for imitating human facial expression. In *2014 11th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ecti-con)* (pp. 1–5).

Clark, E. A., Kessinger, J., Duncan, S. E., Bell, M. A., Lahne, J., Gallagher, D. L., & O'Keefe, S. F. (2020). The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. *Frontiers in psychology*, *11*, 920.

Clark, H. H., & Fischer, K. (2023). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, *46*, e21.

Cohen, I., Looije, R., & Neerincx, M. A. (2011). Child's recognition of emotions in robot's face and body. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 123–124).

Collett, P. (1971). Training englishmen in the non-verbal behaviour of arabs:

An experiment on intercultural communication 1. *International Journal of Psychology*, *6*(3), 209–215.

Cominelli, L., Feri, F., Garofalo, R., Giannetti, C., Meléndez-Jiménez, M. A., Greco, A., ... Kirchkamp, O. (2021). Promises and trust in human–robot interaction. *Scientific reports*, *11*(1), 9687.

Cook, M. (1977). Gaze and mutual gaze in social encounters: How long—and when—we look others" in the eye" is one of the main signals in nonverbal communication. *American Scientist*, *65*(3), 328–333.

Costa Jr, P. T., & McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, *2*, 179–198.

Craig, R., Vaidyanathan, R., James, C., & Melhuish, C. (2010). Assessment of human response to robot facial expressions through visual evoked potentials. In *2010 10th ieee-ras international conference on humanoid robots* (pp. 647–652).

Cully, A., Clune, J., Tarapore, D., & Mouret, J.-B. (2015). Robots that can adapt like animals. *Nature*, *521*(7553), 503–507.

Danev, L., Hamann, M., Fricke, N., Hollarek, T., & Paillacho, D. (2017). Development of animated facial expressions to express emotions in a robot: Roboticon. In *2017 ieee second ecuador technical chapters meeting (etcm)* (pp. 1–6).

Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: A response to cognitive or social difficulty? *Memory & cognition*, *33*(4), 727–733.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*, 45-60.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, *128*(2), 203.

Elliott, E. A., & Jacobs, A. M. (2013). Facial expressions, emotions, and sign languages. *Frontiers in psychology*, *4*, 115.

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In R. J. Davidson, K. R. Sherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (p. 572-595). Oxford: Oxford University Press.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, *24*(6), 581–604.

Friesen, W. V., Ekman, P., et al. (1983). Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, *2*(36), 1.

Fu, C., Liu, C., Ishi, C. T., & Ishiguro, H. (2020). Multi-modality emotion recognition model with gat-based multi-head inter-modality attention. *Sensors*, *20*(17), 4894.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Gillet, S., Cumbal, R., Pereira, A., Lopes, J., Engwall, O., & Leite, I. (2021). Robot gaze can mediate participation imbalance in groups with different skill levels. In *Proceedings of the 2021 acm/ieee international conference on human-robot interaction* (pp. 303–311). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3434073.3444670

Gockley, R., Forlizzi, J., & Simmons, R. (2006). Interactions with a moody robot. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction* (pp. 186–193).

Gonsior, B., Sosnowski, S., Mayer, C., Blume, J., Radig, B., Wollherr, D., & Kühnlenz, K. (2011). Improving aspects of empathy and subjective performance for hri through mirroring facial expressions. In *2011 ro-man* (pp. 350–356).

Greczek, J., Swift-Spong, K., & Mataric, M. (2011). Using eye shape to improve affect recognition on a humanoid robot with limited expression: University of southern california. *Comp. Sci. Department*.

Gu, L., & Su, J. (2006). Gaze control on humanoid robot head. In *Proceedings of the 6th world congress on intelligent control and automation* (Vol. 2, pp. 9144–9148).

Haefflinger, L., Elisei, F., Gerber, S., Bouchot, B., Vigne, J.-P., & Bailly, G. (2023). On the benefit of independent control of head and eye movements of a social robot for multiparty human-robot interaction. In *International conference on human-computer interaction* (pp. 450–466).

Haensel, J. X., Smith, T. J., & Senju, A. (2022). Cultural differences in mutual gaze during face-to-face interactions: A dual head-mounted eye-tracking study. *Visual Cognition*, *30*(1-2), 100–115.

Hart, E., VanEpps, E. M., & Schweitzer, M. E. (2021). The (better than expected)

consequences of asking sensitive questions. *Organizational Behavior and Human Decision Processes*, *162*, 136–154.

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G. (2009). Understanding social robots. In *2009 second international conferences on advances in computer-human interactions* (pp. 169–174).

Hendel, E., Gallant, A., Mazerolle, M.-P., Cyr, S.-I., & Roy-Charland, A. (2023). Exploration of visual factors in the disgust-anger confusion: the importance of the mouth. *Cognition and Emotion*, 1–17.

Hendrikse, M. M., Llorach, G., Grimm, G., & Hohmann, V. (2018). Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Communication*, *101*, 70–84.

Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one*, *10*(8), e0136905.

Holmqvist, K. (2017). Common predictors of accuracy, precision and data loss in 12 eye-trackers. In *The 7th scandinavian workshop on eye tracking*.

Imai, M., Kanda, T., Ono, T., Ishiguro, H., & Mase, K. (2002). Robot mediated round table: Analysis of the effect of robot's gaze. In *Proceedings. 11th ieee international workshop on robot and human interactive communication* (pp. 411–416).

Irfan, B., Kuoppamäki, S.-M., & Skantze, G. (2023). Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults.

Izard, C. E. (2013). *Human emotions*. Springer Science & Business Media.

JASP Team. (2023). *JASP (Version 0.17.2)[Computer software]*. Retrieved from `https://jasp-stats.org/`

Jiao, W., Lyu, M., & King, I. (2020). Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 8002–8009).

Jokinen, K., Furukawa, H., Nishida, M., & Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(2), 1–30.

Kang, J., & Park, Y. E. (2021). A preliminary study on reading the mind in the eyes of the robot. In *2021 30th ieee international conference on robot & human interactive communication (ro-man)* (pp. 839–843).

Kardas, M., Kumar, A., & Epley, N. (2021). Overly shallow?: Miscalibrated

expectations create a barrier to deeper conversation. *Journal of Personality and Social Psychology*, *122*(3), 367–398.

Kaushik, R., & Simmons, R. (2022). Affective robot behavior improves learning in a sorting game. In *2022 31st ieee international conference on robot and human interactive communication (ro-man)* (pp. 436–441).

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*, 22–63.

Kendrick, K. H., Holler, J., & Levinson, S. C. (2023). Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, *378*(1875), 20210473.

Kirby, R., Forlizzi, J., & Simmons, R. (2010). Affective social robots. *Robotics and Autonomous Systems*, *58*(3), 322–332.

Lala, D., Inoue, K., & Kawahara, T. (2019). Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *Proceedings of the 2019 international conference on multimodal interaction* (pp. 226–234). ACM.

Lammerse, M., Hassan, S. Z., Sabet, S. S., Riegler, M. A., & Halvorsen, P. (2022). Human vs. gpt-3: The challenges of extracting emotions from child responses. In *2022 14th international conference on quality of multimedia experience (qomex)* (pp. 1–4).

Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (1999). International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*.

Lapakko, D. (1997). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education*, *46*(1), 63–67.

Lavie, N., Hirst, A., De Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of experimental psychology: General*, *133*(3), 339.

Lazarus, R. S. (2006). Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping. *Journal of personality*, *74*(1), 9–46.

Lazzeri, N., Mazzei, D., Greco, A., Rotesi, A., Lanatà, A., & De Rossi, D. E. (2015). Can a humanoid face be expressive? a psychophysiological investigation. *Frontiers in bioengineering and biotechnology*, *3*, 64.

Lehmann, H., Keller, I., Ahmadzadeh, R., & Broz, F. (2017). Naturalistic conver-

sational gaze control for humanoid robots-a first step. In *Proceedings of the 9th international conference on social robotics* (pp. 526–535). Springer.

Lemhöfer, K., & Broersma, M. (2012). Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior research methods*, *44*(2), 325–343.

Lian, Z., Liu, B., & Tao, J. (2021). Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 985–1000.

Ma, H., Wang, J., Lin, H., Pan, X., Zhang, Y., & Yang, Z. (2022). A multi-view network for real-time emotion recognition in conversations. *Knowledge-Based Systems*, *236*, 107751.

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th acm international conference on multimedia* (pp. 83–92).

Mäkäräinen, M., Kätsyri, J., & Takala, T. (2014). Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley? *Cognitive Computation*, *6*, 708–721.

McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2006). Cultural display rules drive eye gaze during thinking. *Journal of cross-cultural psychology*, *37*(6), 717–722.

Meena, R., Skantze, G., & Gustafson, J. (2014). Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech and Language*, *28*(4), 903–922.

Mehlmann, G., Häring, M., Janowski, K., Baur, T., Gebhard, P., & André, E. (2014). Exploring a model of gaze for grounding in multimodal hri. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 247–254).

Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, *121*(3), 339–361.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., ... others (2010). The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, *23*(8-9), 1125–1134.

Mikels, J. A., Fredrickson, B. L., Larkin, G. R., Lindberg, C. M., Maglio, S. J., & Reuter-Lorenz, P. A. (2005). Emotional category data on images from the international affective picture system. *Behavior research methods*, *37*,

626–630.

Mishra, C., & Skantze, G. (2022). Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In *2022 31st ieee international conference on robot and human interactive communication (ro-man)* (pp. 1201–1208). Naples, Italy: IEEE. doi: 10.1109/RO-MAN53752.2022.9900740

Moors, A. (2020). Appraisal theory of emotion. In *Encyclopedia of personality and individual differences* (pp. 232–240). Springer.

Moubayed, S. A., Skantze, G., & Beskow, J. (2013). The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*, *10*(01). (Id: 1350005)

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, *16*(5), 269–274.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *1*(2), 1–33.

Nakano, Y. I., Yoshino, T., Yatsushiro, M., & Takase, Y. (2015). Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *5*(4), 1–23.

OpenAI. (2023). *Gpt-4 technical report.*

Pan, M. K., Choi, S., Kennedy, J., McIntosh, K., Zamora, D. C., Niemeyer, G., … Christensen, D. (2020). Realistic and interactive robot gaze. In *Proceedings of the 2020 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 11072–11078).

Paplu, S. H., Mishra, C., & Berns, K. (2022). Real-time emotion appraisal with circumplex model for human-robot interaction. *arXiv preprint arXiv:2202.09813*.

Pépiot, E. (2014). Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in parisian french and american english speakers. In *Speech prosody 7* (pp. 305–309). Dublin, Ireland: HAL CCSD.

Pereira, A., Oertel, C., Fermoselle, L., Mendelson, J., & Gustafson, J. (2019). Responsive joint attention in human-robot interaction. In *Proceedings of the 2019 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 1080–1087). IEEE.

Plutchik, R. (1982). *A psychoevolutionary theory of emotions.* Sage Publications.

Pollmann, K., Tagalidou, N., & Fronemann, N. (2019). It's in your eyes: Which

facial design is best suited to let a robot express emotions? In *Proceedings of mensch und computer 2019* (pp. 639–642).

Rasendrasoa, S., Pauchet, A., Saunier, J., & Adam, S. (2022). Real-time multimodal emotion recognition in conversation for multi-party interactions. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 395–403).

Rhim, J., Cheung, A., Pham, D., Bae, S., Zhang, Z., Townsend, T., & Lim, A. (2019). Investigating positive psychology principles in affective robotics. In *2019 8th international conference on affective computing and intelligent interaction (acii)* (pp. 1–7).

Rossano, F. (2012). Gaze in conversation. In *The handbook of conversation analysis* (p. 308-329). John Wiley & Sons, Ltd. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118325001.ch15` doi: https://doi.org/10.1002/9781118325001.ch15

Rossi, A., John, N. E., Taglialatela, G., & Rossi, S. (2022). Generating emotional gestures for handling social failures in hri. In *2022 31st ieee international conference on robot and human interactive communication (ro-man)* (pp. 1399–1404).

Roy-Charland, A., Perron, M., Young, C., Boulard, J., & Chamberland, J. A. (2015). The confusion of fear and surprise: a developmental study of the perceptual-attentional limitation hypothesis using eye movements. *The Journal of Genetic Psychology*, *176*(5), 281–298.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Schellen, E., Bossi, F., & Wykowska, A. (2021). Robot gaze behavior affects honesty in human-robot interaction. *Frontiers in artificial intelligence*, *4*, 51.

Seo, Y.-S., & Huh, J.-H. (2019). Automatic emotion-based music classification for supporting intelligent iot applications. *Electronics*, *8*(2), 164.

Skantze, G. (2017). Predicting and regulating participation equality in human-robot conversations: Effects of age and gender. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction* (pp. 196–204).

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction : A Review. *Computer Speech & Language*, *67*, 1–26.

Skantze, G., Johansson, M., & Beskow, J. (2015). A collaborative human-robot

game as a test-bed for modelling multi-party, situated interaction. In *Proceedings of the 15th international conference on intelligent virtual agents* (pp. 348–351).

So, J., Achar, C., Han, D., Agrawal, N., Duhachek, A., & Maheswaran, D. (2015). The psychology of appraisal: Specific emotions and decision-making. *Journal of Consumer Psychology*, *25*(3), 359–371.

Sprague, R. J. (1999). The relationship of gender and topic intimacy to decisions to seek advice from parents. *Communication Research Reports*, *16*(3), 276–285.

Stahl, J. S. (1999). Amplitude of human head movements associated with horizontal saccades. *Experimental brain research*, *126*(1), 41–54.

Staudte, M., & Crocker, M. W. (2009). Visual attention in spoken human-robot interaction. In *2009 4th acm/ieee international conference on human-robot interaction (hri)* (pp. 77–84). La Jolla, CA, USA: IEEE.

Stefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., & Beskow, J. (2019). Modeling of human visual attention in multiparty open-world dialogues. *ACM Transactions on Human-Robot Interaction (THRI)*, *8*(2), 1–21.

Stock-Homburg, R. (2022). Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research. *International Journal of Social Robotics*, *14*(2), 389–411.

Sullivan, S., Ruffman, T., & Hutton, S. B. (2007). Age differences in emotion recognition skills and the visual scanning of emotion faces. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *62*(1), P53–P60.

Tang, B., Cao, R., Chen, R., Chen, X., Hua, B., & Wu, F. (2023). Automatic generation of robot facial expressions with preferences. In *2023 ieee international conference on robotics and automation (icra)* (pp. 7606–7613).

Tomasello, M., et al. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, *103130*, 103–130.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? some evidence for a theory. *Perceptual and motor skills*, *18*(1), 119–158.

Uemura, T., Arai, Y., & Shimazaki, C. (1980). Eye-head coordination during lateral gaze in normal subjects. *Acta Oto-Laryngologica*, *90*(1-6), 191–198.

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* (pp. 294–301).

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Map-

ping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS one*, *12*(5), e0177239.

Wijayasinghe, I. B., Das, S. K., Miller, H. L., Bugnariu, N. L., & Popa, D. O. (2019). Head-eye coordination of humanoid robot with potential controller. *Journal of Intelligent & Robotic Systems*, *94*(1), 15–27.

Wu, T., Butko, N. J., Ruvulo, P., Bartlett, M. S., & Movellan, J. R. (2009). Learning to make facial expressions. In *2009 ieee 8th international conference on development and learning* (pp. 1–6).

Xu, J., Broekens, J., Hindriks, K. V., & Neerincx, M. A. (2014). Robot mood is contagious: effects of robot body language in the imitation game. In *Aamas* (pp. 973–980).

Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., & Kuzuoka, H. (2008). Precision timing in human-robot interaction: coordination of head movement and utterance. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 131–140).

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., & Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Robotics: Science and systems* (pp. 37–43). Philadelphia, USA: Robotics: Science and systems. doi: 10.15607/RSS.2006.II.037

You, Q., Luo, J., Jin, H., & Yang, J. (2016). Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 30). doi: "https://doi.org/10.1609/aaai.v30i1.9987"

Yu, C., Schermerhorn, P., & Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *1*(2), 1–25.

Zaraki, A., Mazzei, D., Giuliani, M., & De Rossi, D. (2014). Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems*, *44*(2), 157–168.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... others (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Y., Beskow, J., & Kjellström, H. (2017). Look but don't stare: Mutual gaze interaction in social robots. In *Proceedings of the 9th international conference on social robotics* (pp. 556–566). Springer.

Zhong, V. J., Schmiedel, T., & Dornberger, R. (2019). Investigating the effects of gaze behavior on the perceived delay of a robot's response. In *Proceedings*

of the 11th international conference on social robotics (pp. 54–63). Springer.

Research Data Management

This section gives a brief overview of the collection and management of the data used in this dissertation.

Data Collection

The user studies across all four chapters gathered questionnaire responses. Table 7.1 provides an overview of the number of questionnaires per chapter, the participant counts, and the data collection methods.

Chapter	No. of Questionnaires	No. of Participants	Mode of Data Collection
2	1	28	In-person
3	2	181	In-person & Online
4	3	68	in-person & Online
5	2	305	Online

Table 7.1: Details about the collected questionnaire responses per chapter.

In **Chapter 3**, responses were initially obtained from 148 participants via an online survey to assess topic intimacy ratings for selected questions. Subsequently, responses to post-interaction questionnaires were collected from 33 participants who participated in the in-person sessions with the robot. **Chapter 3** also involved the collection of gaze and audio data. The gaze dataset comprised gaze information from 66 sessions (33 sessions for each experimental condition), with transcripts of the robot's utterances, including timing data, stored in *.json* files. The audio dataset consisted of 132 *.wav* format audio recordings (33 sessions X by 2 conditions X by 2 baselines, with each session commencing with a baseline voice recording followed by the interaction).

In **Chapter 4**, two instances of data collection occurred. The first involved two online surveys with 21 participants (the pilot) to determine emotion tags and deck orders for affective images. The second instance comprised questionnaire responses collected during in-person interactions with the robot, constituting the primary user study with 47 participants.

Ethics Approval

The user study in **Chapter 3** was approved by the ethics committee of the Faculty of Language, Literature and Humanities at the Humboldt-Universität zu Berlin. The user studies in **Chapter 4** and **Chapter 5** received approval from the ethics committee of the Faculty of Science, Radboud University, Nijmegen (reference no. ECSW-LT-2023-3-13-98066).

Informed Consent

For each of the user studies conducted, participants completed informed consent forms as a prerequisite for their participation. These consent forms were in compliance with GDPR regulations and met the specific requirements of the institutions that granted ethical approvals for the research. The signed consent forms were securely stored at Furhat Robotics AB, Stockholm and the Max Planck Institute for Psycholinguistics, Nijmegen.

Data Privacy

No personally identifiable information was gathered from the participants in any of the experiments. To ensure anonymity, each participant was assigned a unique alphanumeric code, with no records linking their names to these participant codes.

Data Storage and Sharing

Questionnaire responses from the user studies in **Chapter 2**, **Chapter 3**, and **Chapter 5** were securely stored at Furhat Robotics AB, Stockholm. Gaze, audio, and questionnaire response data from the study in **Chapter 3** were archived in the GDPR-compliant online repository, *HU-Box*, which is managed by Humboldt-Universität zu Berlin. Given that the data collection for the user study in **Chapter 4** occurred in Stockholm and Nijmegen, the corresponding questionnaire responses were stored at Furhat Robotics AB, Stockholm and Max Planck Institute for Psycholinguistics, Nijmegen.

Participants in all these studies provided their consent for sharing their data for research purposes. The data is accessible to the researchers involved in the project and members of the COBRA consortium. Any requests from other researchers for data access will be accommodated in accordance with the guidelines established by the respective institutions where the data is stored.

English Summary

In our day-to-day interactions, we utilize verbal and non-verbal cues (like gestures, body language, facial expressions, eye contact, tone of voice, and spatial proximity) to express our thoughts and emotions. Sometimes, we don't even have to use verbal language to communicate with others and rely solely on non-verbal cues. For example, when meeting a friend a simple hand wave can convey a greeting without the necessity of saying "hello". Similarly, a head nod can indicate agreement or understanding without needing to verbally signal the same. The human face holds particular importance in non-verbal communication, offering a plethora of visual cues such as facial expressions and eye contact. Early research suggested that a significant portion of communication is non-verbal, though the exact percentage is debated. Despite the debates, the fundamental message remains: non-verbal behavior is vital for effective communication.

With the rapid advancements in artificial intelligence and robotics technologies, social robots are poised to have greater social integration. These robots are designed specifically to conduct human-like interactions. So, understanding and replicating essential non-verbal cues, such as facial expressions and gaze, are essential for enhancing these robotic systems' effectiveness, human-likeness, and acceptance. Social robots are already being employed in a variety of domains, including healthcare, education, and assistive roles, where their capacity to convey and interpret human emotions and intentions can significantly impact the quality of interactions. Modeling non-verbal behaviors on these robots would make them more capable of providing a richer user experience. For example, imagine a social robot designed for companionship. When engaging with a person, the robot's facial expressions can reflect warmth and compassion, while maintaining appropriate eye contact can convey active listening and emotional support. These non-verbal cues can help alleviate feelings of loneliness and promote a sense of connection between the patient and the robot.

This research investigates methods for making human-robot interactions (HRI) more seamless and human-like by modeling non-verbal behaviors on social robots and is centered on two key areas:

- Developing architectures to model the eye gaze and emotional behaviors of social robots.
- Evaluating the human perception and influence of these behaviors during HRI

In **Chapter 1**, I introduce the topic and the underlying background that is needed to understand the research better. I first discuss the significance of gaze and affective behaviors in human communication before moving on to Human-Robot Interactions (HRI). Next, I examine the existing models to automate the gaze and affective behaviors of robots and their limitations, which lead to the research questions addressed in this dissertation.

Chapter 2 and **Chapter 3** focused on the gaze behavior of social robots. A comprehensive Gaze Control System (GCS) was proposed and implemented in **Chapter 2** which was used to automate the gaze behavior of a social robot when it is interacting with others. The GCS tried to automate the gaze behavior by considering questions like "*Where should a robot look at during a conversation and why/ how/ when?*". Findings from various fields such as Psychology, Cognitive Sciences, and HRI were studied and used to design the GCS. A user study was carried out to evaluate the GCS which revealed that participants perceived the robot to be more interpretable and preferred when it exhibited human-like gaze behavior. **Chapter 3** investigated the perception of the human-like behavior exhibited by the robot based on the GCS by human participants, more specifically the question "*Does a robot's gaze behavior have any influence on human gaze behavior?*". A user study was conducted to observe the gaze behavior of the participants when they interacted with a robot that kept staring at them vs. a robot that averted its gaze in a human-like manner using the GCS developed in **Chapter 2**. Results from the study showed a direct influence of the robot's gaze aversion behavior on the gaze behavior of the participants. It was observed that participants averted their gaze more when the robot did not avert its gaze at all, as compared to when it averted the gaze in a human-like manner. This showed that in the absence of gaze aversions by a robot, the interaction may become more effortful for the user while trying to avoid frequent mutual gaze with the robot.

I focused on the affective behavior of social robots in **Chapter 4** and **Chapter 5**. **Chapter 4** investigated the possibility of leveraging Large-Language Models (LLM) such as GPT-3 from OpenAI to automate the emotional expressions on a social robot. Recently, LLMs have gained a lot of attention and shown significant capabilities to solve a multitude of problems. I aimed to harness these

capabilities to generate real-time emotional expressions on a robot's face when it is interacting with a human. I implemented a model to use GPT-3.5 (the predecessor of ChatGPT) to predict the emotion that the robot is likely to have during an interaction, based on the ongoing conversation's dialogue history as the context. A user study with an interactive image-sorting task was conducted to evaluate the model and see if the participants could perceive the robot's emotions. Results from the study showed that the participants found the robot to be significantly more human-like, emotionally appropriate, and positive when it exhibited context-appropriate emotional expressions using GPT-3.5. Additionally, it was also found that the participants scored highest in the task when the robot exhibited context-appropriate emotions, showcasing the significance of emotion-appropriate responses in fostering effective human-robot collaboration.

In **Chapter 5** I investigated the questions on how the face should be modeled on a robot for us to better recognize its facial expressions: "*Do we need to model the entire face on a robot or is having just the eyes enough?*" and "*Does the face have to look like a human or would mechanical looking face also be okay?*". These questions are important because they not only shed light on the significance of different facial regions in emotion recognition but also an opportunity to reduce the complexity of generating robot emotions. Results from an online user study indicated that people were able to recognize the emotions exhibited by a robot better when it looked more human-like as compared to a more mechanical look. Additionally, people recognized the expressions of a robot within an acceptable accuracy range when only the eyes were visible as opposed to the full face.

Lastly, in **Chapter 6**, I summarize the overall findings from Chapters 2 to 5 and discuss them as a whole. The combined findings align with the broader goal of developing robots that are empathetic, socially aware, and capable of establishing emotional connections by displaying human-like gaze and affective behaviors. Robots exhibiting such non-verbal behaviors similar to humans have the potential to greatly enhance the quality and depth of interactions. It is crucial to recognize that non-verbal behaviors, encompassing elements like eye contact, facial expressions, and gestures, are deeply ingrained in human communication. These behaviors enable us to convey emotions, intentions, and understanding with subtlety and complexity. By enabling robots to recognize and reciprocate these non-verbal cues, a bridge is formed between human and robot interaction, fostering a more natural and seamless integration of robots into human environments.

Nederlandse samenvatting

In onze dagelijkse interacties gebruiken we verbale en non-verbale signalen (zoals gebaren, lichaamstaal, gezichtsuitdrukkingen, oogcontact, toon van de stem en ruimtelijke nabijheid) om onze gedachten en emoties uit te drukken. Soms hoeven we niet eens verbale taal te gebruiken om met anderen te communiceren en vertrouwen we alleen op non-verbale signalen. Wanneer we bijvoorbeeld een vriend ontmoeten, kan een eenvoudige handbeweging een begroeting overbrengen zonder dat we “hallo” hoeven te zeggen. Op dezelfde manier kan een knikje met het hoofd wijzen op instemming of begrip zonder dat dit verbaal hoeft te worden gesignaleerd. Het menselijk gezicht is bijzonder belangrijk in non-verbale communicatie en biedt een overvloed aan visuele signalen zoals gezichtsuitdrukkingen en oogcontact. Eerder onderzoek suggereerde dat een aanzienlijk deel van de communicatie non-verbaal is, hoewel het exacte percentage wordt betwist. Ondanks de discussies blijft de fundamentele boodschap overeind: non-verbaal gedrag is van vitaal belang voor effectieve communicatie.

Met de snelle vooruitgang in kunstmatige intelligentie en robotica zijn sociale robots klaar voor een grotere sociale integratie. Deze robots zijn speciaal ontworpen om mensachtige interacties uit te voeren. Het begrijpen en nabootsen van essentiële non-verbale signalen, zoals gezichtsuitdrukkingen en kijkgedrag, is dus essentieel om de effectiviteit, menselijkheid en acceptatie van deze robotsystemen te verbeteren. Sociale robots worden al ingezet in verschillende domeinen, waaronder gezondheidszorg, onderwijs en ondersteunende rollen, waar hun vermogen om menselijke emoties en bedoelingen over te brengen en te interpreteren de kwaliteit van interacties aanzienlijk kan beïnvloeden. Het modelleren van non-verbaal gedrag op deze robots zou ze beter in staat stellen om een rijkere gebruikerservaring te bieden. Stel je bijvoorbeeld een sociale robot voor die ontworpen is om iemand gezelschap te houden. Wanneer de robot met een persoon omgaat, kan zijn gezichtsuitdrukking warmte en medeleven uitstralen, terwijl het onderhouden van passend oogcontact een actief luisterend oor en emotionele steun kan uitstralen. Deze non-verbale signalen kunnen gevoelens van eenzaamheid helpen verlichten en een gevoel van verbondenheid tussen de patiënt en de robot bevorderen.

In het onderzoek van dit proefschrift worden methoden onderzocht om mens-robot interacties (HRI) naadlozer en menselijker te maken door het modelleren van non-verbaal gedrag op sociale robots:

- Ontwikkeling van architecturen om de oogopslag en het emotionele gedrag van sociale robots te modelleren.
- Evalueren van de menselijke perceptie en invloed van deze gedragingen tijdens HRI.

In **Hoofdstuk 1** introduceer ik het onderwerp en de onderliggende achtergrond die nodig is om het onderzoek beter te begrijpen. Ik bespreek eerst het belang van kijk- en affectief gedrag in menselijke communicatie voordat ik overga op Mens-Robot Interacties (HRI). Vervolgens onderzoek ik de bestaande modellen om het kijk- en affectieve gedrag van robots te automatiseren en hun beperkingen, wat leidt tot de onderzoeksvragen die in dit proefschrift aan de orde komen.

Hoofdstuk 2 en **Hoofdstuk 3** richtten zich op het kijkgedrag van sociale robots. In **Hoofdstuk 2** werd een uitgebreid Gaze Control System (GCS) voorgesteld en geïmplementeerd dat werd gebruikt om het kijkgedrag van een sociale robot te automatiseren tijdens interactie met anderen. Het GCS probeerde het blikgedrag te automatiseren door vragen te overwegen als "Waar/Wanneer moet een robot naar kijken tijdens een gesprek en Waarom/Hoe?". Bevindingen uit verschillende vakgebieden zoals psychologie, cognitieve wetenschappen en HRI werden bestudeerd en gebruikt om het GCS te ontwerpen. Er werd een gebruikersonderzoek uitgevoerd om het GCS te evalueren, waaruit bleek dat deelnemers de robot als beter interpreteerbaar beschouwden en de voorkeur gaven aan het vertonen van mensachtig kijkgedrag. **Hoofdstuk 3** onderzocht de perceptie van het mensachtige gedrag dat de robot vertoont op basis van de GCS door menselijke deelnemers, meer specifiek de vraag "Heeft het kijkgedrag van een robot enige invloed op het menselijke kijkgedrag?". Er werd een gebruikersonderzoek uitgevoerd om het kijkgedrag van de deelnemers te observeren wanneer ze interageerden met een robot die hen bleef aanstaren versus een robot die zijn blik op een mensachtige manier afwendde met behulp van de GCS die in **Hoofdstuk 2** werd ontwikkeld. De resultaten van het onderzoek toonden een directe invloed van het afkerige kijkgedrag van de robot op het kijkgedrag van de deelnemers. Er werd waargenomen dat deelnemers hun blik meer afwendden wanneer de robot zijn blik helemaal niet afwendde, vergeleken met wanneer de robot de blik op een mensachtige manier afwendde. Dit toonde aan

dat bij afwezigheid van afkerende blikken door een robot, de interactie voor de gebruiker moeizamer kan verlopen wanneer hij veelvuldige wederzijdse blikken met de robot probeert te vermijden.

In **Hoofdstuk 4** en **Hoofdstuk 5** heb ik me gericht op het affectieve gedrag van sociale robots. **Hoofdstuk 4** onderzocht de mogelijkheid om gebruik te maken van Large-Language Models (LLM) zoals GPT-3 van OpenAI om de emotionele uitdrukkingen op een sociale robot te automatiseren. LLM's hebben de laatste tijd veel aandacht gekregen en hebben laten zien dat ze een groot aantal problemen kunnen oplossen. Ik wilde deze mogelijkheden benutten om real-time emotionele uitdrukkingen op het gezicht van een robot te genereren wanneer deze interactie heeft met een mens. Ik heb een model geïmplementeerd om GPT-3.5 (de voorganger van ChatGPT) te gebruiken om de emotie te voorspellen die de robot waarschijnlijk zal hebben tijdens een interactie, op basis van de dialooggeschiedenis van het lopende gesprek als context. Er werd een gebruikersonderzoek met een interactieve taak voor het sorteren van afbeeldingen uitgevoerd om het model te evalueren en om te zien of de deelnemers de emoties van de robot konden waarnemen. De resultaten van het onderzoek toonden aan dat de deelnemers de robot significant menselijker, emotioneel gepaster en positiever vonden wanneer deze context-geschikte emotionele uitdrukkingen vertoonde met behulp van GPT-3.5. Daarnaast werd ook vastgesteld dat de deelnemers de robot als meer humaan, emotioneel gepast en positief beschouwden wanneer deze context-geschikte emotionele uitdrukkingen vertoonde. Daarnaast werd ook vastgesteld dat de deelnemers het hoogst scoorden in de taak wanneer de robot context-adequate emoties vertoonde, wat het belang aantoont van op emoties afgestemde reacties bij het bevorderen van effectieve samenwerking tussen mens en robot.

In **Hoofdstuk 5** onderzocht ik de vragen hoe het gezicht gemodelleerd moet worden op een robot zodat we zijn gezichtsuitdrukkingen beter kunnen herkennen: "*Moeten we het hele gezicht modelleren op een robot of zijn alleen de ogen genoeg?*" en "*Moet het gezicht eruit zien als een mens of zou een mechanisch uitziend gezicht ook goed zijn?*". Deze vragen zijn belangrijk omdat ze niet alleen licht werpen op het belang van verschillende gezichtsregio's in emotieherkenning, maar ook een mogelijkheid bieden om de complexiteit van het genereren van robotemoties te verminderen. Resultaten van een online gebruikersonderzoek gaven aan dat mensen de emoties van een robot beter herkenden wanneer deze er meer menselijk uitzag dan wanneer deze er meer mechanisch uitzag. Bovendien herkenden mensen de uitdrukkingen van een robot binnen een acceptabel

nauwkeurigheidsbereik wanneer alleen de ogen zichtbaar waren in plaats van het volledige gezicht.

Tot slot vat ik in **Hoofdstuk 6** de algemene bevindingen uit de hoofdstukken 2 tot en met 5 samen en bespreek ze als geheel. De gecombineerde bevindingen komen overeen met het overkoepelende doel om robots te ontwikkelen die empathisch en sociaal bewust zijn en in staat om emotionele verbanden te leggen door het vertonen van mensachtige blikken en affectief gedrag. Robots die dergelijk non-verbaal gedrag vertonen, vergelijkbaar met mensen, hebben het potentieel om de kwaliteit en diepgang van interacties sterk te verbeteren. Het is cruciaal om te erkennen dat non-verbaal gedrag, waaronder elementen als oogcontact, gezichtsuitdrukkingen en gebaren, diep geworteld zijn in menselijke communicatie. Deze gedragingen stellen ons in staat om emoties, bedoelingen en begrip, subtiel en complex over te brengen. Door robots in staat te stellen deze non-verbale signalen te herkennen en te beantwoorden, wordt een brug gevormd tussen de interactie tussen mens en robot en wordt een meer natuurlijke en naadloze integratie van robots in menselijke omgevingen bevorderd.

Curriculum Vitae

Chinmaya Mishra was born on November 18th, 1989 in Burla, Odisha. He earned a Bachelor's degree in Computer Science and Engineering from the Institute of Technical Education and Research, SOA University, Bhubaneswar in 2012. Following his graduation, he worked as a Systems Engineer at Tata Consultancy Pvt. Ltd focusing on the development and maintenance of critical client infrastructure. In 2015, he resigned from his position to pursue formal education in AI, ultimately obtaining his Master's in Intelligent Systems with a minor in Psychology from the Technical University of Kaiserslautern (now RPTU) in 2020.

During his Master's studies, his focus areas were Natural Language Processing (NLP) and Human-Robot Interactions (HRI). He conducted his Master's thesis research at the Robotics Research Lab, TU Kaiserslautern, under the supervision of Prof. Dr. Karsten Berns, developing a behavior control architecture to automate the affective behaviors of a humanoid robot named ROBIN. He also worked as a Research Assistant during this period, where he conducted exercise sessions for the lecture "Applications of Artificial Intelligence".

In 2021, he was awarded the Marie Curie PhD Fellowship under the COBRA ITN project and began his industrial PhD at Furhat Robotics in Stockholm under the supervision of Prof. Dr. Gabriel Skantze and Prof Dr. Peter Hagoort, with co-supervision from Dr. Susanne Fuchs and Dr. Rinus Verdonschot. His PhD research involved collaborations with the Neurobiology of Language department at the Max Planck Institute for Psycholinguistics and the Leibniz-Centre General Linguistics (ZAS), Berlin. He has also supervised students in India and is actively collaborating on a research project with IIT Mandi and SIT Pune. Additionally, he served as a member of the supervisory board of the COBRA project, representing the PhD students.

Currently, he is a postdoctoral researcher at the Max Planck Institute for Psycholinguistics' Multimodal Language Department, working with Prof. Dr. Aslı Özyürek and Dr. Judith Holler.

For more information scan the following QR code:

or visit www.chinmayamishra.com

Publications

Kejriwal, J., **Mishra, C.**, Offrede, T., Skantze, G., & Beňuš, Š. (2024). Does a Robot's Gaze Behavior Affect Entrainment in HRI?. (Pre-print)

Mishra, C., Skantze, G. , Hagoort, P. & Verdonschot, R., (2024). The Influence of Human-likeness and Facial Regions on the Perception of Social Robot Emotions. *12th International Conference on Affective Computing and Intelligent Interaction (ACII 2024)*, (under Review)

Mishra, C., Verdonschot, R., Hagoort, P. & Skantze, G., (2023). Real-time Emotion Generation in Human-Robot Dialogue Using Large Language Models. *Frontiers in Robotics and AI*, doi: 10.3389/frobt.2023.1271610

Mishra, C., Offrede, T., Fuchs, S., Mooshammer, C. & Skantze, G., (2023). Does a Robot's Gaze Aversion Affect Human Gaze Aversion?. *Frontiers in Robotics and AI*, 10:1127626, doi: 10.3389/frobt.2023.1127626

Offrede, T., **Mishra, C.**, Skantze, G., Fuchs, S., & Mooshammer, C. (2023). Do humans converge phonetically when talking to a robot?. *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, Prague, Czech Republic.

Mishra, C., & Skantze, G. (2022, August). Knowing Where to Look: A Planning-based Architecture to Automate the Gaze Behavior of Social Robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1201-1208). IEEE.

Acknowledgements

I have pushed writing this section till the very end (literally writing it the day before I am supposed to submit the thesis for printing) because, to me, it signals not only the end of a journey but also the beginning of writing probably the hardest part of this thesis! Hardest because I have to condense 3 years' worth of support and encouragement into a few pages. Considering that I have made it this far, I am fairly confident that I will manage to finish this as well. That said, let me get into writing this section in earnest. This has been a wonderful experience and I could not have done it without the support and help from all the amazing people mentioned in these acknowledgments.

I consider myself really lucky to have had the best team of supervisors that a candidate could ever ask for: **Gabriel Skantze**, **Peter Hagoort**, **Rinus Verdonschot**, and **Susanne Fuchs**!

Gabriel, thank you for always encouraging me to pursue my research ideas. From the very first day, you encouraged me to present my ideas and treated me as a collaborator rather than an inexperienced student. You gave me the space and support that I needed to become independent and confident in my research. You were always there to silently guide me and nudge me onto the right track when necessary. Perhaps the most memorable moment for me was when you presented me with the book "Gaze and Mutual Gaze" after I gave my first conference presentation in ROMAN. That was not only encouraging but also signified a form of acknowledgment of my research from you. You never imposed a deadline and it always kept me wondering if what I had done was enough. In hindsight, this was what allowed me to explore things on my own and be better prepared for the next steps in my career. I have learned so much from you that it is impossible to put into words — a heartfelt thanks for showing me the right path to becoming a researcher.

Peter, thank you so much for being my guardian in academia. You took me under your wing and have always shielded me from the bureaucratic pressures in academia. You allowed me to focus on my work without worrying about other distractions. Even though I was away from Nijmegen, you ensured that I felt welcomed and at home when I visited for the first time. During my interview when everyone was asking about my research ideas and experience, I remember how you suddenly jumped in and asked only 2 questions: "I would be a fool not to hire you because..." and "I will bring these two things to your group with me". Meetings with you were always to the point, but you also took time to ask me about my future plans and offer advice. Thank you for always having my back and the confidence that I will succeed.

I was introduced to **Rinus** as my supervisor for when I would visit Nijmegen for a secondment. Little did I know that I would not only find a supportive supervisor but also a good friend in you. You always made sure to drag me

on coffee walks after lunch. You might not know this, but these walks had a profound impact on my work: they either prevented me from dozing off at my desk or helped me relax when I was getting overwhelmed with work, both of which directly resulted in increased productivity. Perhaps that was always your plan? Thank you for always appreciating my work and being the cool person that you are.

Susanne you have been the kindest and most caring throughout my PhD journey. You always kept my well-being in mind and tried your best to guide me in the right direction. Anytime there is an event or article that you feel is relevant to me, you immediately share it with me which goes on to show how you are actively thinking about helping me. I cannot describe how grateful I am for all the pointers you have provided me throughout my PhD, be it on designing studies or preparing for my defense, thank you!

These short paragraphs are insufficient to capture the profound impact that all of you have had on my life in these past years and in shaping me into the researcher I have become. I will do my best to incorporate your teachings into my work and to support any students whom I end up supervising down the line, in the same way that I have been supported by you. You have made my PhD journey the wonderful experience that it was.

Besides my supervisors, I would also like to thank my manuscript committee for taking the time to go through this manuscript and provide their valuable comments. Thank you **Harold Bekkering**, **Judith Holler**, and **Gerard Bailly**. I look forward to exchanging ideas and discussing interesting topics with you all.

A big thank you to my paranymphs **Sho Akamine** and **Sachit Misra**. We got to know each other rather recently after I moved to Nijmegen this year, but both of you have come forward to support me on this last leg of my PhD. Thank you both for taking care of the social and admin part of my defence letting me focus on the research part! Sho, I am looking forward to having a lot more discussions on anime and maybe even binge watch a few together. We need to convert more people into anime lovers dattebayo!

Being part of a big consortium like the **COBRA** gave me an edge as an early-stage researcher with immediate access to a group of peers and a big network of senior researchers to learn from. A big thank you to my COBRA family for providing me the opportunity to learn and conduct inter-institute collaborations and interdisciplinary research. I was lucky to have so many wonderful ESRs as my peers who have all supported each other throughout our PhD journeys. Shoutout to **Jay, Tom, Carol, Lena, Adaeze, Greta, Dorina, Joanna, Johannah, Emilia, Byron, Junfei, Mercedes, Salome, Lavinia**, and **Xinyi**! Thank you **Jay** for being my brainstorming partner and a close friend. Thanks **Tom** for being a wonderful host in Berlin and taking me to the best ramen place on my birthday! **Lena**, thank you for being the LOTR expert that you are and for the awesome real-time tapestries. Thanks **Carol** for being my comrade in arms at Furhat and saving me from being the only person staring at the ceiling (read - PhD candidate) in the office.

This PhD would not have been possible without a conducive environment. This being an industrial PhD, made the work environment even more crucial.

Furhat was perhaps the best place that I could ask for. I am grateful for the wonderful culture that Furhat has and am privileged to have gotten the opportunity to work at a place where social robots are made, both of which have directly helped me realize this PhD journey. I remember **Samer** saying "Furhat is better with Chinmaya" when offering me the position, and I can say with surety that "my PhD was better with Furhat". Thank you **Susanne** for being such an emotional and caring person who was always willing to help me out. I could never thank **Johan** enough for being a great friend and my office brainstorming buddy! **Mathieu**, your music playlists were a lifesaver! I am looking forward to teaching you more about how to be better at ping-pong. **Nils**, thank you for the insightful discussions we always end up having. **Kaspar** and **Morgan**, thanks for being the coolest managers that I have ever worked with. A big shoutout to everyone at Furhat for always making me feel at home and supporting me in every way they can.

I was welcomed with open arms to the **TMH** department in KTH even though I wasn't a PhD candidate there. This allowed me to interact with fellow PhD candidates in the university which became a place to gain both inspiration and support from, not to mention, to have fun as well. The contributions of **#Room509** can never go unsaid. Perhaps I ended up becoming an external/ honorary 509 member and always ended up being there when at KTH. Thanks **Shivam** for inviting me to 509 and treating me to awesome Indian snacks every time I was there. **Alireza**, we started our PhD journey together and navigated the Swedish immigration process on the same boat. Thank you for being a dear friend and supporting me. I am still eagerly waiting for when we finally get to collaborate. Shoutout to **Charlotte, Jim, Birger, Harm, Ambika**, and **Katya**. Thanks **Bahar** for proving that quality papers can still be written in an insanely short amount of time and encouraging me to pursue different research avenues.

Being an external PhD candidate it is always difficult to coordinate and integrate with the host institution. But my colleagues at **MPI** and the **NBL** department made sure that this was not the case. Thank you **Kevin** for always being there to help me navigate the processes at MPI and ensuring that you advised the best course of action beneficial to me. You have been a constant source of support and encouragement to me. Thanks a lot **Angela** for making the HR process seem so easy and well organized. **Michaela**, thank you for always being there to help me and offering me chocolates. It is difficult to find good friends later in life but thank you **Talat** for being a great friend. Our discussions on interesting topics have always inspired me and I am eagerly waiting to start the work that we have agreed on as soon as I get a position in India. Thanks a lot to my colleagues at the NBL department (especially **Ellie, Hatice, Anne, Birgit, Danbi**) for accepting me into their fold and making me feel a part of it. I am also grateful to my new colleagues at that **MLD** department for their constant encouragement and support as I transition into the new role while wrapping up my PhD formalities.

Before I move on to thank my friends and family, I must thank **Twitter** for keeping me up to date with the literature, recent happenings in the field and also for being my source of entertainment. It gave me a platform to connect with

researchers all over the world during COVID when it was difficult to network due to restrictions on in-person events. Thanks **Elon Musk** for destroying (almost) that.

I could not have made it this far without the support and encouragement of my friends and family. **Roshan**, we coincidentally ended up reconnecting overseas after so many years and even became neighbors. It was a blessing to have an old friend in a foreign place when starting a new chapter in my life. Thanks for all your help. **Manav, Nidhi** and **Praveen**, you guys were my constant support and source of entertainment (I guess we can all agree that it was mostly Nidhi on the stage, with all of us playing the role of the audience). Thanks **Daya** and **Titli** for always expressing how proud you felt in what I was doing. This inspired me to do better. I can't say how happy I get whenever I recall that both of you are also pursuing your PhDs now. Thanks **Sachin** for being my first comrade in pursuing a PhD together from TUK and for the endless PhD memes that kept the humor alive. A huge thanks to my **F.R.I.E.N.D.S** from college who have always appreciated what I was doing and encouraged me to push forward. Special thanks to **Amarjeet** for making sure that I realize that my younger brother is more of a friend to you than I am. Perhaps it is your way of getting back at me for leaving India for all these years. But I know how proud you are of what I am doing and I thank you for that.

Thanks **Jr. Mishras** (my cousins and sisters-in-law) for being the most amazing bunch of people out there. Coincidentally, in a group of 11, 5 of us ended up pursuing our PhDs in a wide array of fields. Every time we had a group call, it always ended up feeling like participating in a conference with animated discussions on so many interesting state-of-the-art topics. The number of eureka moments that we have had and the constant support that I have received from you all is precious. Thank you for that.

A skyscraper cannot be built without a solid foundation. For me, that foundation is my family. I could not have pursued my dream of obtaining a PhD had it not been for the selfless sacrifices and unwavering support I have received from my **Mom** and **Dad**. Thank you **Dad** for being a role model for me, not just with your academic achievements, but also for exemplifying what hard work and integrity mean. Thank you **Mom** for constantly being there to support every decision that I have taken so far, and for giving me the confidence to chase after my dreams. Not many families would support their son to leave an established job and jump into academia with all its uncertainties just to pursue his dreams, but I am lucky to have such a family. The blessings of my **grandparents** and the pride that I see in their eyes when they talk about me are all the encouragement I could ask for. A huge thanks to my grandfather for enthusiastically learning about AI and social robotics even at his age, and inspiring me further by engaging in deep philosophical discussions with me.

Thanks **Tanmaya** (my younger brother) for being my therapist, Grammarly, and reviewer. I could not have made it without our regular Fortnite therapy sessions with **Bimal** and **Omstavan**. Every time I write something with a deadline approaching, you are my go-to peer reviewer and grammar/spelling checker. Not to mention that you ended up designing my PhD thesis cover (and the GIF

at the end) as well. More importantly, thank you for always having my back, I couldn't have made it without you!

I would like to thank my **in-laws** for having faith that I will definitely finish this PhD and for giving me Arpita's hand! Shoutout to my **sister-in-law** for being a perennial source of entertainment.

Less than a month into my PhD journey, we started our journey together as well **Arpita**. I remember both of us planning to be together soon in Stockholm right after you finished your PhD. Little did we know how difficult it is to realize any plans during a PhD. Even though we ended up living away from each other, you never made me feel like we were apart. You took care of our family back home whenever needed so that I could have peace of mind and focus on finishing my work. I couldn't have asked for a more supportive and caring life partner who always has my back through all the ups and downs that we have experienced. I am really grateful for that. Of course, on another note, I would like to thank you for staying away from me for all these years, which helped me focus on research and complete my PhD in time! :)

Lastly, I would like to express my gratitude to God for showing me the right path and giving me the strength to push through difficult times.

Jai Jagannath!

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

VISITING ADDRESS
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

POSTAL ADDRESS
P.O. Box 310
6500 AH Nijmegen
The Netherlands

CONTACT
T +31(0)24 3521 911
F +31(0)24 3521 213
E info@mpi.nl
Twitter @MPI_NL
www.mpi.nl

GA n° 859588 - H2020

Furhat Robotics

Conversational Brains