# Look on my thesis, ye mighty:
## Gaze Interaction and Social Robotics

Vidya Somashekarappa

*Doctor of Philosophy*

APRIL 2024

UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

Doctoral Thesis in Computational Linguistics

Advisors: Asad Sayeed and Christine Howes

The Author is supported by grant 2014-39 from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

Cover design by © Jelle Tjeerd Fokkens

The title of the cover image: 'Robots and I'

Print: Stema Specialtryck AB, Borås, Sweden 2024

Publisher: University of Gothenburg (Dissertations)

Photographer: Monica Havström

Distribution:

Department of Philosophy, Linguistsics and Theory of Science

University of Gothenburg

Box 100, SE-405 30, Gothenburg

ISBN: 978-91-8069-659-3 (PRINT)

ISBN: 978-91-8069-660-9 (PDF)

The publication is also available in full text at:

http://hdl.handle.net/2077/80112

*To my family, friends & cats,*

*I met a traveller from an antique land*
*Who said: Two vast and trunkless legs of stone*
*Stand in the desert. Near them, on the sand,*
*Half sunk, a shattered visage lies, whose frown,*
*And wrinkled lip, and sneer of cold command,*
*Tell that its sculptor well those passions read*
*Which yet survive, stamped on these lifeless things,*
*The hand that mocked them and the heart that fed:*
*And on the pedestal these words appear:*
*"My name is Ozymandias, king of kings:*
***Look on my works, ye Mighty**, and despair!"*
*Nothing beside remains. Round the decay*
*Of that colossal wreck, boundless and bare*
*The lone and level sands stretch far away.*
*–Ozymandias by Shelley*

# Abstract

Gaze, a significant non-verbal social signal, conveys attentional cues and provides insight into others' intentions and future actions. The thesis examines the intricate aspects of gaze in human-human dyadic interaction, aiming to extract insights applicable to enhance multimodal human-agent dialogue. By annotating various types of gaze behavior alongside speech, the thesis explores the meaning of temporal patterns in gaze cues and their correlations. On the basis of leveraging a multimodal corpus of dyadic taste-testing interactions, the thesis further investigates the relationship between laughter, pragmatic functions, and accompanying gaze patterns. The findings reveal that laughter serves different pragmatic functions in association with distinct gaze patterns, underscoring the importance of laughter and gaze in multimodal meaning construction and coordination, relevant for designing human-like conversational agents. The thesis also proposes a novel approach to estimate gaze using a neural network architecture, considering dynamic patterns of real-world gaze behavior in natural interaction. The framework aims to facilitate responsive and intuitive interaction by enabling robots/avatars to communicate with humans using natural multimodal dialogue. This framework performs unified gaze detection, gaze-object prediction, and object-landmark heatmap generation. Evaluation on annotated datasets demonstrates superior performance compared to previous methods, with promising implications for implementing a contextualized gaze-tracking behavior in robotic interaction. Finally, the thesis investigates the impact of different gaze patterns from a robot on Human-Robot Interaction (HRI). The results suggest that manipulating robot gaze based on human-human interaction patterns positively influences user perceptions, enhancing anthropomorphism and engagement.

# Sammanfattning

Avhandlingen undersöker nyanserade aspekter av blick som kommunikativt fenomen i mänsklig tvåpartsinteraktion, med syftet att utvinna insikter som kan användas för att förbättra multimodal dialog mellan människa och system. Blick är en viktig ickeverbal social signal som förmedlar uppmärksamhet och ger inblick i andras intentioner och framtida handlingar. Genom att annotera diverse varianter av blick tillsammans med tal undersöks betydelsen hos temporala mönster i blicksignaler och deras korrelationer. På basis av en multimodal korpus av tvåpartsinteraktioner kretsande kring smakprovning undersöker avhandlingen relationen mellan skratt, pragmatiska funktioner och tillhörande blickmönster. Resultaten belyser att skratt fyller olika pragmatiska funktioner förbundna med distinkta blickmönster; därigenom understryks skrattets och blickens vikt i multimodal betydelsebildning och -koordinering, vilket är relevant vid utformning av människoliknande konverserande agenter. Avhandlingen föreslår också ett nytt tillvägagångssätt för blickestimering genom en neuronnätsartitektur som tar hänsyn till dynamiska mönster i verkliga blickbeteenden vid naturlig interaktion. Detta ramverk syftar till att underlätta responsiv och intuitiv interaktion genom att möjliggöra för robotar/avatarer att kommunicera med människor genom naturlig multimodal dialog. Ramverket sammanför blickförutsägelse, blickobjektsförutsägelse och färgdiagramsgenerering för landmärken. Utvärdering på basis av ett annoterat dataset som presterar överlägset jämfört med tidigare metoder, med lovande implikationer för utveckling av kontextualiserat blickspårningsbetande i robotinteraktion. Slutligen undersöker avhandlingen betydelsen av olika blickmönster från en robot för människa-robotinteraktion (HRI). Resultaten indikerar att manipulering av robotblick utifrån mönster i människa-människainterkation påverkar användares upplevelser positivt samt förstärker antropomorfism och engagemang.

# Contents

# Acknowledgements

CAUTION: This section should not be considered in any way representative of the seriousness of the thesis but much rather reflective of the whimsical side of the author. Not sure I'm gonna go and get another PhD, so I'll use this platform to thank all of the people that mean/meant something to me during this journey. You have shaped or reshaped me over the years and I'm very grateful for your existence. Some of you will be surprised, some not and some definitely saw it coming. Read at your own risk.

Even though many of you do not know me personally, I'm certain you picked up the book and went directly to the acknowledgements. I appreciate you being here. Please join the party later this evening.

Somashekarappa and Prema, my parents, thank you for being there through thick and thin and choosing to have me and only me. It is a lot of pressure to not f**k things up when you're an only child. Thank you for making education an utmost priority from a very young age without which it wouldn't have been possible for me to be here today. Your dedication and resilience serve as inspiration to strive for. Hope I've made you proud. Yes mom, I will send you the certificate to frame, as long as you hang it in the attic of your bedroom for no one to ever see!

Pooja and Priya, my sisters, for believing I could do anything, anytime, anyday and always. For being selfless and constantly by my side during crisis while laughing in tears, no matter what. I could probably write another thesis of all the craziness that we share together. I'll be there for you (when the rain starts to pour), I'll be there for you (like I've been there before), I'll be there for you ('cause you're there for me too). There is so much more to be said, but lets move on, who has the time, really..

Rutger, my backbone, for keeping me grounded, I love you!

For my in-laws, Bob, Jacqueline, Frederique, Benjamin, Max and Matthijs for sharing this wonderful journey with me. You are home away from home. Our cats, Particle and Wave, I thank you for being the most unpredictable marshmallows and breaking multiple pots. You have taught me patience and how to let go of any desire to own something nice. Your obsession with Harold shall always stay a mystery unless one of you speaks up. I'm forever grateful for all the comfort and happiness you bring to our lives. Your softest meow is worth a thousand hugs.

Pär and Maria, thank you for adopting me; the pandemic of '20 and '21 wouldn't have been the same without you and I'm grateful that it brought us closer. You inspire me everyday to do more and live a better life (lagom eh!). I am running the Göteborgsvarvet this year, am I welcome in the club? I'll make sure to have yogurt always in case you come around Pär. I'd be surprised if Maria hasn't popped open her laptop during the defense and finished most of her work. Thanks for being here! Virtually or physically. Elias, my ex-house mate, those years were filled with joy, discussions about politics, investments and four hour long debates about DNA in genetic inheritance. You did make/invent the best potato pancakes and what a blast it was hosting parties that lasted until day-break. Thanks for being my hype-man around-the-clock and trying out strange recipes that I cooked.

Anastasia, for comforting and being a central pillar of support. Your strength and agility has inspired me to look beyond my limitations. Thank you for introducing me to and sharing many of my first experiences, including ice bathing, bouldering, and more. For consistently unintentionally twinning on any given day and taking long walks by the docks.

Rashmi and Akshaya, my decade old friends from back home, for reminding me of the taste, feel and touch of Mysore thousands of miles away (not the Swedish miles, that would bring it down to a hundred). Shoutout to my pub quiz crew Sarah, Jane, Elise, Thomas, Bastian and many more for making dreary Sunday evenings much more entertaining (especially during 'em winters, that lasts for 12 years in Göteborg).

Asad and Chris, my academic parents, thank you for guiding and shaping me through these years. You made hard work seem fun and easy. I'm forever grateful to have had you as my supervisors and for being the only two people to have read the thesis probably more than I have. Thank you for burning the midnight candle during deadlines despite having a packed schedule. Asad for sharing the south-east asian roots and being there to laugh about cultural indifferences. Chris, for patiently listening to me rant various theories about Gen Z's. While writing the acknowledgements I found out that they can't go to restaurants with more than 3 items in the menu, apparently causes 'em distress and anxiety. I'm pretty sure you've already heard this one. Shoutout to gen Z's - we can now place an order under 30 seconds. Efficiency is key to loneliness. Asad, I will miss the board game nights during department conferences. Chris, I'll always look forward to dancing many more dances with you.

Sharid, for heart warming conversations and reminding me of the reality. Eva-Marie for chairing the defense and trusting me with your plants. They may or may not be dead, you will only find out once you open the door to my office which I wouldn't recommend doing (they are Schrodinger's plants now). Staffan, for being so very kind and considerate when I was in crisis and s**t hit the fan. Ellen, for inspiring elegance on the 5th floor.

Adam, for many pizza dates and introducing me to Swedish culture, wait scratch that, Adam culture. I could have an hour long mundane conversation with you and not be bored. Anything syntactic, you'll always pop in my mind. Mehdi, for being the big brother (not like the show), and a constant subject of my pranks. Bill, for hosting board game nights, for the chili, chili oil and sour bread, delicious. Vlad, for co-authoring a fun paper, literally. Hope you never stopped making kombucha. Nikolai, for being an amusing frenemy, eventhough he'll deny it. Hana, for always being so very kind, showing me how to bark and for being care free. Your colour palette is very soothing and I shall always find you. Cheers to the comp-ling PhD's for all the stimulating discussions, fun night outs, taking strolls in Delsjön, picking blueberries/mushrooms and canoeing till dark.

Robin, Shalom, Simon, Stergios, Jean-Phillip, Eleni, Rasmus, Nina, Robert, Aram, Amandine and Fahima, thank you for your guidance and encouragement. I'm forever grateful for how homey the group has made me feel over the past few years.

Maria, my work husband, apple of my eye, mischievous, mysterious, hand-standing, cart-wheeling, push-up queen of a co-worker, thanks for being my alter ego. Every minute we spend together is memorable and we definitely should not be left to our own devices.

Dominik, Orvar and Alex for being the best office mates and choosing the coolest room, what a view! Thank you for keeping silence during my afternoon naps and wondering how I manage to do it. Dominik, thank you for being a constant ear and providing TeX support, not tech, he's not Indian (only I can say that, comes with privileges). You have helped me navigate some difficult situations and never got bored of my constant life updates, I intend to continue to do so. Thanks for making stressful weeks/months more palatable. Orvar for being the first in office everyday so that I didn't have to struggle to find my keys from the bottom of the bag with a million things in hand. You always put a smile on me. You have been a constant support when looking for a job and completing the thesis was extremely overwhelming. Alex for figuring out the etikprövningen and paving way for future ethical approval applicants so that they struggle a lot less. Also, I really appreciate you taking the time and translating the abstract to Swedish.

Tjeerd, for designing the beautiful cover of this thesis. I have to say, no one can rock my 6 inch shiny black leather boots like you. I'm always in awe of your courage, let's please sing role-reversed "I'm a barbie girl" while you wear my hair extensions again. Embarrassing yourself in front of your colleagues takes a lot of guts. I want us to watch more of the same trash reality-tv and have scheduled serious fika discussions. Leave no stone unturned.

Markus, thank you for being a tea nerd and serving me precisely made luscious specialty tea. Monica for taking exquisite photos and deep meaningful dinner dialogues. Iines, thank you for those countless shit-chats and laughs. If a carved pumpkin shows up again on your desk around Halloween, I'll help you figure out the culprit. Niclas, everyone does need to tie up on Fridays! Thank you for everything from educating me about training dogs to wonderful conversations discussing wine and french vineyards. Doris for lending me your shoulder, wait, was it arms, when extremely needed. Thomas, for all the insightful unworthy colloquy and always keeping me delightfully entertained during lunch. I will never forget the way you can butcher a beet burger. Yes, beet! When I told you my defense date was the 25th of April, you enthusiastically delighted me with the fact that the date is exactly in the middle of when Hitler died and was born. Regular people would have said- "congratulations".

Bahareh, I'm captivated by your charisma and thank you for all the helpful advice you have given me about raising cats. My stupendous colleagues Giacomo, Gianluca, Leila, Jasmine, Dorna, Alexander, Jacob, Hadi, Annika, Davide, Filippa, and Mattias for lasting memories during my time here. Richard Endörfer for not mentioning me in his thesis, I'll do it here, just for revenge. Lena, Johan, Alexander, Fredrik, Christopher, Susanna, Emelie, Jennifer, Mikaela, Cecilia, Ulrik, Sandra and Hektor for supporting with anything and everything. You've been reliable and delivered when it mattered the most.

Are you still with me? Finally, cheers to all the cool PhD's, postdocs and researchers I have met abroad during conferences and workshops. You know who you are, only because you would received this book by mail. Hope you haven't changed addresses. Thank you for educating me on what not to do!

My deepest appreciation to the green-reader Timo Baumann for reading the thesis and providing insightful suggestions. I am grateful for the opponent Dominique Knutsen for accepting the challenge of not only reading but presenting the thesis. I really appreciate the committee members Danielle Matthews, Alexandrs Berdicevskis, Emilia Barakova and Ellen Breitholtz for taking on the role. You will forever be some of the very few people that would ever read the book.

I would like to thank the Adlerbert scholarships and the Donation Board scholarships for their generous grants. CLASP for making this research possible, and sending me to places I'd otherwise never visit. The fireplace I put a sticker on at Dicksonsgatan was never moved to Renströmsgatan unfortunately. I'm still very sore about that.

PS: You're reading the 26th version of the thesis and I'm yet to be convinced its the final version.

Thank you!
Vidya

# Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Gothenburg or any other institution.

**Vidya Somashekarappa**

# Declaration

The thesis contains research work that has been accepted/published in peer-reviewed conference proceedings. The contributions from "An annotation approach for social and referential gaze in dialogue" published in the *Proceedings of the Twelfth Language Resources and Evaluation Conference* (LREC, May 2020) and "A deep gaze into social and referential interaction" in the *Proceedings of the Annual Meeting of the Cognitive Science Society* (CogSci, June 2021) are discussed in Chapter 2. The contributions from "Looking for laughs: Gaze interaction with laughter pragmatics and coordination" in the *Proceedings of the International Conference on Multimodal Interaction* (ICMI, Oct 2021) co-authored with Dr. Chiara Mazzocconi and Dr. Vladislav Maraev are discussed in Chapter 3. The contributions from "Neural Network Implementation of Gaze-Target Prediction for Human-Robot Interaction" in the *IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN, Oct 2023) are discussed in Chapter 4. Finally, the contributions from "Good Looking: How Gaze Patterns affect Users' Perceptions of an Interactive Social Robot" to be published in the *Advanced Robotics and Its Social Impacts* (ARSO, May 2024) is discussed in Chapter 5. All the work if not explicitly mentioned was carried out in close collaborations with my PhD supervisors Associate Prof Asad Sayeed and Prof Christine Howes.

# Abbreviations

- AI - Artificial Intelligence
- ASR - Automatic Speech Recognition
- DCNN - Deep Convolutional Neural Network
- DL - Deep Learning
- EAC - ELAN Analysis Companion
- EAI - Event Attention Interface
- ECA - Embodied Conversational Agent
- EEG - Electroencephalography
- ELAN - EUDICO Linguistic Annotator
- ERP - Event-Related Potentials
- EUDICO - European Distributed Corpora Project
- FACE - Facial Automaton for Conveying Emotions
- GAT - Gesprächsanalytisches Transkriptionssystem
- GHI - Good-housekeeping Institute Corpus
- HHI - Human-Human Interaction
- NIR - Near-Infrared Cameras
- HRI - Human-Robot Interaction
- LSTM - Long Short-Term Memory
- NLU - Natural Language Understanding
- NMMC - Nottingham Multimodal Corpus
- NN - Neural Network
- RNN - Recurrent Neural Network
- VR - Virtual Reality
- TTS - Text-Speech Synthesis
- UES - User Engagement Scale

# Chapter 1

# Theoretical Foundations of Gaze in Interaction

A Latin proverb by Marcus Tullius Cicero from the second century BC states: 'The face is the portrait of the mind; the eyes, its informers' [1]. The act of looking at something or someone involves different cognitive, sensory, and motor systems in the human body (Vickers 2011). Gaze is a form of nonverbal communication that is used to convey a variety of social and emotional messages (Burgoon and Bacue 2003). The subtle informative cues of gaze facilitate fluent interactions among people (Hessels 2020). Incorporating these gaze predictions into a robot can yield a better level of engagement with humans and enable the robot to anticipate a human's intentions and goals (Saran et al. 2018).

Imagine you are having a conversation with a friend. When you are talking to them, you might naturally look at them to show that you are paying attention and engaged in the conversation. That is "gaze behavior." Now, let's say you are telling your friend a funny story, and both of you burst out laughing. While you're laughing, you might notice that you tend to look at your friend to share the moment. That is another aspect of gaze behavior – using your eyes to connect with someone emotionally. The thesis researches how people use their eyes to communicate during conversations

1. https://en.wikipedia.org/wiki/Cicero

and social interactions, and to teach robots to replicate the same behaviour. This is achieved by analyzing how people look at each other, how long they look, and when they look away. By understanding these patterns, it is possible to teach robots to "read" human gaze cues and respond in a more natural and human-like way.

For example, let's say you have a robot assistant at home. Instead of just responding to your voice commands, the robot could also understand where you're looking and react accordingly. If you are talking to the robot and then look at a bookshelf, the robot might understand that you are interested in the books and offer to help you find something to read. It is about making interactions with robots feel more like interactions with people.

Ultimately, the goal is to create robots that can understand and respond to human gaze behavior, making them more intuitive and helpful in various situations. Whether it is assisting with tasks at home or providing companionship, these socially intelligent robots could make a big difference.

The thesis proposes a novel approach to predict human gaze for human-robot interaction for different types of gaze in real-time. It is divided into three main parts. First, a human-centered approach provides an in-depth understanding of gaze in human-human interactions present via multimodal corpus studies (Chapter 2 & Chapter 3). Secondly, an automation approach which leverages machine learning to automatically detect gaze in human-human interaction (Chapter 4). Finally, an exploratory approach to integrate dialogue and gaze cues in human-robot interaction (Chapter 5).

## 1.1 Human-Human Gaze Interaction

The world inhabited by humans is complex and rich with information. How does everyone survive without becoming overwhelmed? We are limited in our sensory and cognitive abilities, even though there are hundreds or thousands of objects and other types of information visible. Thankfully, not everything that exists has a bearing on

our immediate objectives or long-term survival. Humans have progressively evolved techniques for choosing relevant information and learning where to attend while discarding irrelevant visual information. The same difficulty is faced by artificially intelligent (AI) agents as they transition from a simple digital environment to the complex real world: how do they select crucial information from a sea of information?

Gaze has a dual functionality, comprising "a signalling function" that relays information back into the environment, and "an encoding function" that gathers information from other individuals. It is a vital communication cue that can convey mental and emotional states (Mason et al. 2005), as well as the direction and object of attention (Kuhn et al. 2008; Gobel et al. 2015). The purpose of a human-centered research approach is to understand interactive human behaviour. While on the other hand, the design-focused approach tend to manipulate features of the robotic gaze, i.e., length of fixation, and includes lab-based or field-based evaluations. Technology-focused research aims to build the computational tools for robot eye-gaze in human interaction. The thesis address all three designs of human gaze behaviour during interaction.

### 1.1.1 Gaze in a Social Context

Social gaze can be interpreted as communicative by observers. Humans have a unique ability, beyond that of non-human primates, to interpret others' attention through eye-gaze (Emery 2000). Researchers reason that the depigmentation of the human sclera, unique among primates, has evolved for effective communication and social interaction based on eye contact (Kobayashi and Kohshima 1997). Linguists and psychologists have long been interested in non-verbal communication relating to speech and gesture, including eye-gaze (Kendon 1967a; Argyle and Cook 1976a; Goodwin 1980; Goodwin 1981), which is the focus of this work.

Chapter 2, outlines the research on social and referential functions of eye-gaze in dialogue and argues that these should not be treated independently, as eye-gaze information is multifunctional. Further, it gives a brief description of the existing multimodal corpora that include eye-gaze information (Chapter 2.3), discussing the need for a new annotation scheme for eye-gaze. Finally, it presents our annotation scheme and preliminary observations (refer Chapter 2.4).

Humans perceive robots differently from how they perceive other humans – even though robots cue higher-order responses to gaze, they do not trigger the face-specific neural pathways at very short timescales (Admoni and Scassellati 2012; Ragni and Stolzenburg 2015). To establish gaze to indicate next speaker or a desire to take the next turn it is important to interpret and produce interactional cues. For example, gaze can be used to signal the beginning and end of a speaking turn in social interaction (Ho et al. 2015a). To establish more natural dialogues with humans, a conversational agent's ability to direct attention to the most appropriate target in a multimodal interaction is important (Norris 2004). Bousmalis et al. (2009) presented cues such as head nodding and smiles, and Hunyadi (2019) used gestures but did not include gaze while investigating the temporal patterns of non verbal cues to study agreement and disagreement.

Previous studies tend to focus on either social functions of gaze (e.g., turn-taking or other interaction management (Jokinen et al. 2013)) or how gaze is used in reference resolution (Kontogiorgos et al. 2018), with few researchers combining these. Hence in Chapter 2, we explore the social, referential and pragmatic aspects of gaze in human-human interaction with an eye towards their future implementation in human-robot interactions.

Argyle and Cook (1976b) showed that listeners display longer sequences of uninterrupted gaze towards the speaker, while speakers tended to shift their gaze towards and away from the listener quite often. Later work has refined these observations, with, for instance, Rossano (2013), noting that these distributional patterns are dependent on the specific interactional activities of the participants; for example, a

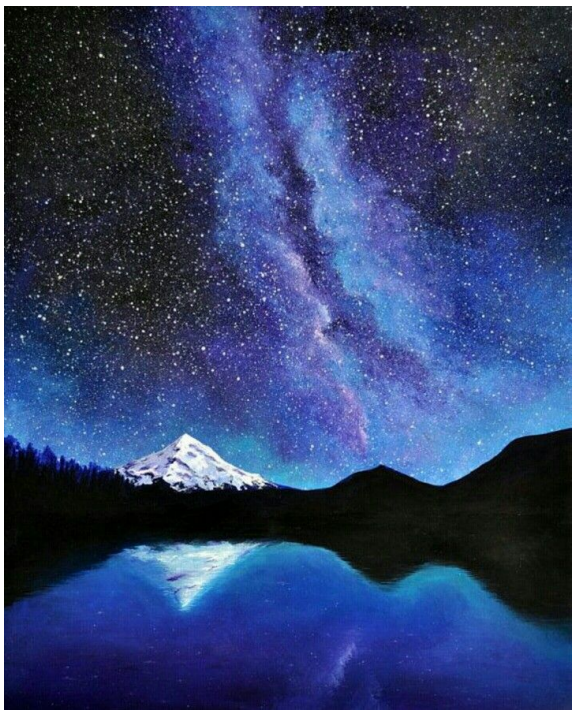


Figure 1.1: Illustration I

more sustained gaze is necessary in activities such as questions and stories, since gaze is viewed as a display of attention and engagement. Brône et al. (2017a) also found that different dialogue acts typically display specific gaze events, from both speakers' and listeners' perspectives.

Unaddressed participants also display interesting gaze behaviour showing that they anticipate turn shifts between primary participants by looking towards the projected next speaker before the completion of the ongoing turn (Holler and Kendrick 2015a). This may be because gaze has a 'floor apportionment' function, where gaze aversion can be observed in a speaker briefly after taking their turn before returning gaze to their primary recipient closer to turn completion (Kendon 1967a; Brône et al. 2017b).

Figure 1.2: Illustration II

### 1.1.2 Gaze in Referential Context

In identifying an image on display by referring to "the painting of a night sky", our attention is drawn automatically to Illustration I (fig. 1.1)[2] but not to Illustration II (fig. 1.2)[3] even without any necessary pointing gesture. The process of identifying application-specific entities which are referred to by linguistic expressions is called reference resolution.

One area in which multi-modal reference resolution has been previously studied is in the context of sentence processing and workload. Sekicki and Staudte (2018) showed that referential gaze cues reduce linguistic cognitive load. Earlier work, Hanna and Brennan (2007), showed that gaze acts as an early disambiguator of referring expressions in language.

Campana et al. (2002) proposed to combine the reference resolution component of a simulated robot with eye tracking information; they intended to deploy it on the International Space Station. However, they did not address the integration of eye movements with speech. Also, eye-gaze information was used only in case of inability to identify unique referenced objects. Zhang et al. (2004) implemented

2. https://www.pinterest.com/pin/717268678124480359/
3. http://www.digitalpicturezone.com/digital-photography-tips-and-tricks/taking-photos-in-the-mid-day-sun/

reference resolution by integrating a probabilistic framework with speech and eye-gaze; results showed an increase in performance. They also found that reference resolution of eye-gaze could also compensate for lack of domain modelling. Visual input has a immediate effect on language interpretation like reference resolution.

In humans, gaze has evolved to play a central role in social communication and interaction. Humans have highly developed visual systems and the ability to direct their gaze in a controlled manner, which allows us to convey a wide range of social and emotional messages through gaze behavior (Bailey et al. 2009; Harwerth et al. 1986). This is thought to have been important for the development of language, as well as for the formation and maintenance of social relationships (Brooks and Meltzoff 2005; Abele 1986).

### 1.1.3 Cultural Differences in Gaze Behaviour

Gaze behavior can vary across cultures, reflecting differences in social norms and expectations for eye contact and other aspects of gaze (McCarthy et al. 2008; Haensel et al. 2022). In Western cultures, direct eye contact is often seen as a sign of attentiveness and sincerity, while avoiding eye contact can be interpreted as a sign of dishonesty or disinterest (Uono and Hietanen 2015; Akechi et al. 2013).

In some Asian cultures, direct eye contact is not as highly valued, and it may be seen as impolite or aggressive to maintain eye contact for too long (Yuki et al. 2007). Instead, these cultures may place more value on looking down or averting the gaze in certain social situations. In some indigenous cultures, gaze can be used as a form of nonverbal communication, and individuals may use their gaze to convey respect, challenge, or other social messages (Matsumoto and Hwang 2016; Adetunji and Sze 2012).

These cultural differences in gaze behavior can have important implications for social interaction and communication, and it is important for individuals to be aware of and respect the cultural norms and expectations for gaze in the context in which they are communicating. While the cultural differences in gaze behavior exist, they are not absolute and can vary within a culture based on individual differences and situational factors.

Robots need to be sensitive to these cultural cues and adapt their gaze behavior to align with local expectations. In conversational dynamics, the role of gaze in turn-taking can vary across cultures. Some cultures may expect consistent eye contact during conversations, while others may interpret constant eye contact as impolite or confrontational (Marchesi et al. 2023). Implementing machine learning algorithms can help robots learn and adapt their gaze behavior over time based on user interactions and cultural context. This adaptive capability enables robots to improve their cultural sensitivity and effectiveness in various settings.

### 1.1.4 Gaze in Decision Making

Eye-gaze plays a significant role in human decision making. Research has shown that people pay attention to the gaze direction of others and use it as a cue to determine demonstrate their own beliefs and intentions (Frischen et al. 2007). In social situations, people often rely on eye-gaze to infer the emotional state and mental state of others, and to understand social norms and expectations (McKay et al. 2021). For example, if someone is looking at you while they are speaking, it can signal that they are engaged and paying attention to you, while if they are looking away, it can indicate that they are distracted or not interested. eye-gaze can also influence our own behavior and decisions. For example, if someone is looking at us in a certain way, it can influence our confidence levels, our perception of the situation, and our decisions about what to do next (Smith and Krajbich 2019).

In some species of primates, such as chimpanzees and bonobos, eye-gaze can be used as a means of communication and to establish dominance or cooperation. In these species, making eye contact with another individual can signal aggression or submission, and the direction of gaze can be used to negotiate social relationships and determine which individual is in charge (Fröhlich et al. 2016). In other species, such as dogs, eye-gaze can play a role in communication and bonding. For example, dogs will often look at their owners to get their attention or to understand what is expected of them (Koyasu et al. 2020). In birds, eye-gaze can be used to assess the dominance of potential mates or to signal readiness to mate (Dawkins 2002). In many species, eye-gaze can also play a role in decision making by helping animals to determine where to focus their attention and what actions to take in response to potential threats or opportunities in their environment.

Artificial agents, such as robots and artificial intelligence systems, have been designed to simulate animal or human-like behaviors, including eye-gaze. However, compared to humans and non-human animals, the ability of artificial agents to use eye-gaze in decision making is limited and still in its early stages of development. Current artificial agents are capable of detecting and tracking the gaze direction of individuals, and can use this information to make simple decisions, such as deciding where to direct their attention or which objects to focus on. This is an area of ongoing research and development, as researchers aim to develop artificial agents that are more capable of mimicking human-like gaze behavior and decision making (Hortensius and Cross 2018).

### 1.1.5 The Interaction of Gaze and other Non-Verbal Signals

During laughter, gaze behavior can provide important social cues and signals to others, indicating a person's level of engagement and interest in a conversation or situation (Grammer 1990). For example, people tend to make more eye contact with others while laughing together, which can help to build rapport and strengthen social bonds. This can be especially true in situations where laughter is used to

signal agreement, intimacy, or shared enjoyment (Glenn 2003). In some cases, gaze behavior during laughter may also reflect a person's level of comfort or discomfort in a given social context (Gironzetti et al. 2016). People who are laughing nervously may avoid eye contact, which can signal that they are uncomfortable or unsure about the situation. On the other hand, people who are comfortable and confident may engage in direct eye contact while laughing, which can be perceived as a sign of dominance or assertiveness (Sporer and Schwandt 2007). In these cases, gaze behavior during laughter may also play a role in establishing and maintaining power dynamics within a group.

Laughter and gaze have an important role in managing and coordinating social interactions. In Chapter 3, using a multimodal corpus of dyadic taste-testing interactions, we explore whether laughs performing different pragmatic functions are accompanied by different gaze patterns towards the interlocutor, both from the point of view of the laughing participant and from the partner. Chapter 3.3, describes how we investigate the role of gaze in laughter coordination between interactants. Our results (as elaborated in chapter 3.4) show that laughs performing different pragmatic functions are related to different gaze patterns, both for the laugher and her partner, and that gaze is an important cue exploited by interactants when reciprocating laughter or laughing simultaneously. We discuss our data in relation to the literature about laughter and gaze functions in interaction, linking them to dialogic context (ref chapter 3.5). The results stress the importance of laughter and gaze for modeling of multimodal meaning construction and coordination in interaction, and are therefore relevant for researchers designing human-like embodied conversational agents.

## 1.2 Gaze Estimation

The second strand of work in this thesis concerns gaze automation. Gaze estimation evaluates human intent and interest by measuring the gaze of the human eye. The roots of human gaze estimation and eye-tracking trace back to the 18th century when researchers employed invasive eye-tracking techniques to observe eye movements (Kar and Corcoran 2017; Khan and Lee 2019). However, with the advancements in digital signal processing and computer vision, non-invasive gaze estimation approaches have become more prevalent, leveraging unique physical characteristics of the eye (Chennamma and Yuan, 2013). The photometric and motion characteristics of the human eye have proven crucial in providing the necessary features for this task (Akinyelu and Blignaut, 2020). Gaze estimation involves two key metrics: gaze direction and point of gaze. Gaze direction is determined by the visual axis, which deviates from the optical axis. Eye properties such as the pupil and corneal reflection, extracted from eye regions, are utilized at the application level to ascertain gaze direction. Subsequently, the gaze point is defined as the intersection of the gaze direction and the object's surface.

Automatic gaze analysis develop methods for estimating the position of the target objects, by tracking the movements of the eyes (Valenti et al. 2011). An accurate technique should be able to separate gaze while withstanding a wide range of difficulties, such as identity bias, occlusions, eye-head interplay, lighting fluctuations, and eye registration errors. Moreover, studies have demonstrated how human gaze follows an arbitrary trajectory when eye moves, which adds another level of difficulty to gaze estimation (Alnajar et al. 2013; Saran et al. 2018). The main focus here is to accurately identify the gaze direction on intended objects using state-of-the-art machine learning approaches. Gaze analysis has three components: registration, representation, and recognition.

The identification of the eyes, or eye-related critical areas, or maybe even simply the face, is the initial step, or registration. Eye-tracking technology is commonly used for registration (Chung et al. 2005), employing techniques such as infrared light to monitor the position and movement of the eyes (Ramdane-Cherif et al. 2004). The goal is to accurately register the gaze points and track the eye movements over time. Registration is crucial to establish a foundation for further analysis. Accurate data captured during this stage ensures that the subsequent representations and interpretations are based on reliable information. Representation involves the conversion of raw gaze data into a meaningful and understandable format. The raw gaze data, often in the form of coordinates or sequences of points, is processed and transformed into visualizations or representations (Duchowski 2018). These representations may include gaze plots, heatmaps, or gaze path diagrams, providing a visual summary of where the person looked and for how long. By representing gaze data visually, researchers and analysts can gain insights into patterns, trends, and areas of interest. Visualization aids in the interpretation of the data and helps communicate findings more effectively. Recognition is the final stage where the interpreted gaze data is used to infer or recognize specific behaviors, intentions, or cognitive processes. Advanced algorithms and models are applied to analyze the gaze patterns and make inferences about the person's cognitive or visual attention. This may involve identifying points of interest, understanding reading patterns, or recognizing emotional states based on gaze direction. Recognition enables a deeper understanding of the individual's cognitive processes and behaviors. This information can be valuable in various fields, including human-computer interaction, psychology, marketing, and user experience design.

In chapter 4, we develop an automated system for analyzing visual scenes, by extracting gaze direction and target information in the scene simultaneously (such as turn taking, joint attention, gaze following, gaze aversion and mutual attention) in a natural dyadic interaction. We propose a model that utilizes the manual annotation of gaze targets in a natural dialogue setting and generate simultaneous gaze

prediction for both parties in the video, along with attention heatmaps that provide exclusive information of the target object-of-interest in the scene, by also providing out-of scene gaze predictions. The novel tool presented can assist in automatic extraction of gaze data for both AI/HRI applications and clinical/psychological research. There is currently no single dataset that covers all of the different gaze and scene combinations that we address in this chapter.

Currently, robotic gaze systems are reactive in nature but the proposed Gaze-Estimation framework can perform unified gaze detection, gaze-object prediction and object-landmark heatmap in a single scene, which paves the way for a more proactive approach. We generated 2.4M gaze predictions of various types of gaze in a more natural setting (GHI-Gaze). The predicted and categorised gaze data can be used to automate contextualized robotic gaze-tracking behaviour in interaction. We evaluate the performance on a manually-annotated data set and a publicly available gaze-follow dataset. Compared to previously reported methods our model performs better with the closest angular error to that of a human annotator. Existing baseline methods and the data that was generated covers different categories of gaze. Furthermore, we discuss the ethical implications and considerations of the study and propose an implementable gaze architecture for a social robot.

The majority of classic gaze analysis models rely on customised low-level attributes (such as colour, shape, and appearance) and specific geometrical algorithms in order to get beyond their restrictions and handle typical unconstrained scenarios. Similar to other computer vision tasks, gaze analysis has seen a change in methodology since 2015, with a shift towards deep learning. Over the past several years, the difficulties related to variations in illumination, camera setup, eye-head interaction, etc., have significantly decreased due to deep learning-based models and the availability of massive training datasets.

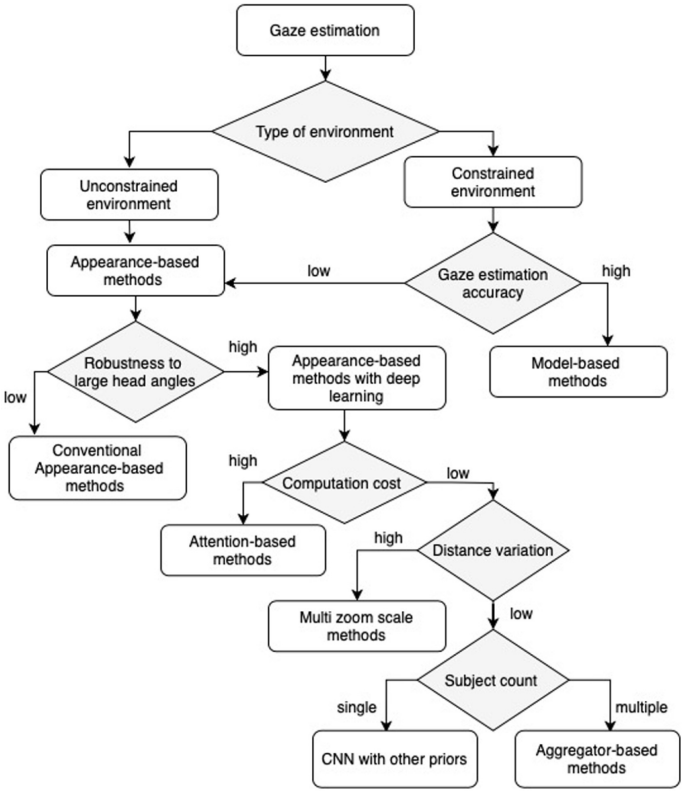



Figure 1.3: Gaze estimation approaches

### 1.2.1 Gaze Estimation Approaches

As discussed in the previous section, while there are many eye-gaze estimate systems available, they are costly, unreliable, requiring human intervention, and inaccurate in real-world applications. Additionally, certain conventional methods' performance is restricted by things like poor image quality and lighting. Deep Learning (DL) based eye-gaze estimation techniques are useful in these situations because of their high accuracy, flexibility, automation, learning from pre-existing data, and improved decision-making. Prevalent deep learning methods have demonstrated efficacy in enhancing performance in eye gazing applications (Sangeetha 2021). Model-based and appearance-based techniques are the two main categories into which human gaze estimation methodologies can be divided (Azad et al. 2006).

### 1.2.2 Model-Based approaches

In order to manually regress the eye features and create a geometric model, model-based methods essentially require specialised equipment, such as near-infrared (NIR) cameras (Kar and Corcoran 2017). This approach is limited to confined spaces and person-specific (Akinyelu and Blignaut 2020).

### 1.2.3 Appearance-Based Approaches

Due to outstanding resilience and applicability in unconstrained situations, gaze estimation approaches based on DL techniques have garnered significant attention in the last ten years in the eye-tracking field. Appearance-based methods are neither environment or device-specific and do not require special equipment. These techniques can be further classified into appearance-based methods using DL and standard appearance-based techniques.

The ability to extract high-level gaze cues from images and the capacity to learn a non-linear mapping function directly from the image to eye-gaze are only a couple of the advantages that DL-based methods have over traditional appearance-based approaches (Cheng et al., 2021; Kellnhofer et al., 2019). Due to their capacity to map image features directly, manage large-scale datasets, and learn complex non-linear mappings when challenged by significant head-pose variations, eye occlusions, and lighting conditions, deep convolutional neural networks (DCNN) have been used in nearly every DL-based gaze estimation approach.

The primary appearance-based approaches using DL, can be further subdivided into two groups according to the quantity of subjects: single-user gaze estimate and multi-user gaze estimation. The need for multi-user gaze estimation methodologies is growing, even if there has been a notable change in gaze estimation strategies towards unconstrained situations. By the end of 2021, only a small number of these techniques had been studied, including time and space-shifting single-user gaze estimation.

Multi-user gaze estimate is primarily needed in open environmental situations like retail stores, public meeting places, and public arenas, as opposed to traditional single-user gaze estimation. Therefore, it calls for reliable, high-speed, low-overhead gaze measurement techniques. Time-sharing approaches and space-sharing approaches comprise the two types of existing multi-user gaze estimation studies (Pathirana et al. 2022; Zhang et al. 2019). The number of users is dispersed across a certain amount of time using the time-sharing mechanism. However, the space sharing method processes more than one user concurrently. Owing to their limited scalability and lack of resilience, time-shifting techniques have not received much attention in the literature.

### 1.2.4 Calibration-Free Approaches

Calibration-free gaze estimation aims to determine where a person is looking without the need for an extensive calibration procedure (Alnajar et al. 2013). Traditional gaze estimation systems often require users to undergo a calibration process, where they follow a set of visual targets to establish a mapping between their eye movements and the corresponding gaze directions. Calibration-free approaches seek to eliminate or minimize this calibration step, making gaze estimation more user-friendly and applicable in various scenarios.

Pupil center methods estimate gaze direction based on the position of the pupil in the eye images (Zhang et al. 2011; Wan et al. 2021). These methods often assume a fixed relationship between the gaze direction and the position of the pupil center. Corneal reflection methods utilize the corneal reflections (purkinje images) in the eyes to estimate gaze direction. The relative positions of these reflections in the eye images can be used to infer the gaze point. Meanwhile, synthetic data generation involves training deep learning models on synthetic datasets that simulate a wide range of gaze directions and eye appearances (Trampert et al. 2021). By training on diverse synthetic data, models can potentially generalize well to real-world scenarios without the need for user-specific calibration.

### 1.2.5 Non-Intrusive Methods

Remote gaze estimation techniques use external sensors, such as cameras, to estimate gaze direction without any physical contact with the user. These methods are more user-friendly and applicable in various scenarios. Video-based eye tracking uses webcams and RGB cameras to capture images or videos of the user's face and eyes. Computer vision algorithms analyze the images to track eye movements and estimate gaze direction. This method is non-intrusive as it doesn't require any specialized hardware. Infrared-based eye tracking uses infrared light which is often invisible to the human eye but can be detected by specialized cameras. These systems use infrared light sources and cameras to capture eye movements.

Depth cameras (e.g., Kinect) provide information about the distance of objects in the scene, by combining depth information with RGB data, these cameras can capture the three-dimensional position of the eyes and face, enabling gaze estimation without physical contact. Integrated eye-tracking sensors are smart glasses and head-mounted displays that come with built-in eye-tracking sensors, often based on infrared technology. These sensors track eye movements and can estimate gaze direction without requiring additional external cameras or sensors.

Remote photoplethysmography (rPPG) measures variations in skin color caused by blood flow. By analyzing subtle color changes in the face, particularly around the eyes, it's possible to estimate heart rate and, to some extent, gaze direction. While not as precise as dedicated eye-tracking methods, rPPG is non-intrusive and can provide additional context. Electrooculography (EOG) measures the electrical potential generated by eye movements. While traditional EOG involves placing electrodes on the skin around the eyes, newer non-contact methods, such as using capacitive sensors, are being explored for gaze estimation.

## 1.3 Human-Robot Gaze Interaction

In human-robot interaction, the robot's gaze behavior can be used to signal its intentions and attention, making it easier for humans to understand what the robot is trying to communicate. For example, a robot that maintains eye contact while speaking can signal that it is paying attention to the person and actively engaged in the conversation. On the other hand, if the robot avoids eye contact or looks away frequently, it can signal that it is not fully engaged or is distracted. Gaze behavior can also be used to direct the attention of the human towards specific objects or areas. For example, a robot that points its gaze towards an object can signal to the human that the object is important or relevant. This can be particularly useful in tasks where the robot needs to guide the human's attention, such as in a museum tour or a training scenario.

Anthropomorphism is frequently used in the domains of human-robot interaction to enhance users comfort with machines. Assigning human characteristics to robots in an effort to reduce the complexity of technology provides an extra measure of comfort (Marakas, Johnson, Palmer, 2000; Moon Nass, 1996). "Face-to-face" interactions are still regarded as the gold standard of communication whether communicating with people or conversational agents, even though interactions between humans involve many subtle social cues (Adalgeirsson Breazeal, 2010). Therefore, in order to be regarded as socially intelligent partners in interactions, agents must use anthropomorphic designs and a diverse range of social behaviours.

These components are used by a large number of social robots, particularly those with human-like designs, and they enable the creation of non-verbal social behaviours during interactions with people (Fong, Nourbakhsh, Dautenhahn, 2003). Numerous behavioural components, such as subtle facial expressions and gaze movements, are crucial for contextualised human conversational contexts. The fact that a lot of known information is stored in the non-verbal signs that are being transmitted makes face-to-face interaction desirable. Nevertheless, creating and deciphering

these cues results in increased cognitive burden, which could lengthen interaction times. This shows that conversational agents that resemble humans and are able to convey predictable nonverbal behaviours may facilitate social interactions in users but may be less effective at completing tasks.

Chapter 5 investigates whether interactive behaviour can be altered by a human-like face (a social robot) that can convey nonverbal cues as opposed to conversational agent (a smart speaker) that does not make use of these multimodal aspects. Our contribution is a user research in which participants engaged in conversation with a social robot with and without gaze cues. In order to better understand the implications of the comparison, we compare the social robot's non-verbal conduct in three conditions and investigate whether various social eye-gaze characteristics contribute to the observed differences.

In addition to gaze behavior, robots can use other nonverbal cues, such as head and body movements, to enhance their communication and interaction with humans. These cues can provide additional information about the robot's emotional state and intentions, making it easier for humans to understand what the robot is trying to communicate. Gaze behavior can be artificially applied in a variety of ways to enhance the capabilities of artificial agents and make them more effective in their tasks.

### 1.3.1 Human-Robot Interaction

Gaze behavior can be used to make robots more effective in communicating and interacting with humans. For example, a robot that can mimic human gaze patterns can make it easier for people to understand what it is trying to communicate.

### 1.3.2 Object Tracking

Gaze behavior can be used to track and identify objects in an environment. For example, a robot equipped with gaze tracking can identify and track a moving object, such as a ball or a person, and use that information to make decisions about what actions to take. Multimodal perception and sensor fusion techniques have long been established, but in recent years, there has been a surge of interest primarily driven by the growth of smart environments and sensor networks, particularly those found in autonomous or semi-autonomous vehicles. Various approaches to multi-sensor data fusion can be categorized into high-level fusion (HLF), low-level fusion (LLF), and mid-level fusion (MLF). LLF and MLF strategies involve integrating data from different sensors to facilitate collaborative detection and tracking, ultimately leading to the creation of shared perceptual maps. For instance, one approach (Luiten et al. 2020) combines optical flow, scene flow, stereo-depth, and 2D object detections to track objects in 3D space, while another proposed method (Kim et al. 2021) focuses on 2D and 3D bounding box detection to develop a more scalable fusion system. The implementation of this system is to identify objects and individuals to facilitate interaction for social robots (Cruces et al. 2022).

### 1.3.3 Attention Allocation

Gaze behavior can be used to determine where an artificial agent should direct its attention. For example, an AI system equipped with gaze tracking can use information about where a person is looking to determine which parts of an image or video to focus on. It is an important aspect of perception and decision making, as it enables individuals to prioritize information and attend to the most important or relevant stimuli. Within the framework of CORTEX (Cognitive Robotics Architecture), exists a unified and dynamic working memory known as Deep State Representation (DSR). This mechanism captures information across various levels of abstraction,

ranging from raw perceptual data to high-level symbols and action plans. DSR serves as a short-term dynamic representation of both internal factors like inner state and proprioception, as well as external elements such as environmental data, objects, and individuals within the surroundings for attention allocation (Bustos et al. 2019).

### 1.3.4 Emotion Recognition

Gaze behavior can be used to recognize and understand human emotions. An AI system equipped with gaze tracking can use information about where a person is looking to determine their emotional state and make decisions about how to respond. An individual who is happy and engaged in a conversation might maintain eye contact, while someone who is feeling uncomfortable or sad might avoid eye contact or look down frequently. Similarly, changes in gaze direction and patterns can be indicative of other emotions, such as anger, fear, or surprise.

### 1.3.5 Importance of Gaze in Human-Robot Interactions

**1. Trust and credibility**: A natural and convincing gaze behavior can help to increase the trust and credibility of artificial agents in the eyes of users. By making eye contact, nodding, and gesturing, artificial agents can convey a sense of engagement and attentiveness that is similar to human behavior. **2. Empathy and emotional connection**: Gaze behavior can also help to create a sense of empathy and emotional connection between users and artificial agents. By using gaze to convey emotions, such as smiling or frowning, artificial agents can appear more human-like and foster a stronger emotional bond with users. **3. Attention and interaction**: Gaze can also be used to direct the attention of users and facilitate interaction. For example, an artificial agent may use gaze to signal that it is ready for a user's input or to indicate that it is paying attention to a user's conversation. **4. Social cues and**

**context**: Gaze behavior can also provide important social cues and context that help to guide interactions between users and artificial agents. For example, an artificial agent's gaze behavior may change depending on the type of interaction, such as providing information, asking a question, or providing feedback.

## 1.4 Experimental Evaluation of Human-Robot Gaze Perception

The Chapter 5 of the thesis examines the role of gaze automation within social robots, highlighting its significance in promoting more engaging, intuitive, and effective interactions between humans and robots. It emphasizes the importance of a robot's ability to control its gaze, encompassing aspects such as where it looks, how it looks, and when it looks, as essential for establishing natural and meaningful communication with users. Through a comprehensive review of existing literature, the chapter emphasizes the impact of human-like behavior in robots on people's perceptions, focusing particularly on the establishment of trust, rapport, and engagement through appropriate eye contact.

The objectives of the study encompass implementing different gaze patterns in social robots, experimentally evaluating the impact of these patterns on human-robot interaction, and analyzing how specific gaze patterns correlate with users' perceptions and experiences. Through rigorous experimental investigation and analysis, the chapter aims to shed light on the intricate dynamics of gaze behavior in human-robot interaction, ultimately paving the way for more effective and natural interactions in diverse contexts.

Chapter 5, investigates whether different gaze patterns from a Furhat robot can lead to more effective, natural and engaging interactions. The results indicate that gaze manipulations based on gaze patterns from human-human interaction positively impact user perceptions compared to the neutral and random conditions. Participants

rate the anthropomorphism and animacy of the robot in the experimental conditions and the findings contribute to understanding the impact of robot gaze on user perceptions and engagement, offering insights for the design and improvement of interactive social robots.

Moreover, it highlights how gaze behavior serves as a potent nonverbal communication tool, enabling robots to convey intentions, emotions, and social cues, thereby facilitating more seamless and intuitive interactions. The research aims to contribute significantly to the field by assessing various gaze patterns and their implications on interaction quality and engagement.

# Chapter 2

# An Annotation Approach for Social and Referential Gaze in Dialogue

This chapter of the thesis introduces an approach for annotating eye-gaze considering its social, referential and pragmatic functions in multi-modal human dialogue. As we discussed in the previous chapter, to assess gaze patterns it is essential to obtain quantitative temporal data in order to draw implications. This is achieved by providing novel observations that can be executed in a machine to improve multimodal human-agent dialogue. Gaze is an important non-verbal social signal that contains attentional cues about where to look and provides information about others' intentions and future actions. Detecting and interpreting the temporal patterns of gaze behaviour cues is a natural and mostly unconscious process for humans. However, these cues are difficult for conversational agents such as robots or avatars to process or generate. It is key to recognise these variants and carry out a successful conversation, as misinterpretation can lead to total failure of the given interaction. This chapter introduces an annotation scheme for eye-gaze in human-human dyadic interactions that is intended to facilitate the learning of eye-gaze patterns in multi-modal natural dialogue. In this work, various types of gaze behaviour are annotated in detail along with speech to explore the meaning of temporal patterns in

gaze cues and its co-relations. Considering that 80% of the total stimuli perceived by the brain is visual (Kamitani and Tong 2005), gaze behaviour is complex and challenging; hence, implementing human-human gaze cues in an avatar/robot could improve human-agent interaction and make it more natural.

The following section of the chapter provides a detailed understanding of gaze behaviour in different situations by proposing an annotation paradigm to annotate where people look and what they say during a conversation. Is it possible to predict where a person would look when they speak based only on their eye movements? Do we share a typical pattern of gaze when we speak and how does these patterns evolve to affect where the attention is drawn? After annotating the conversations, an analysis is conducted to see how eye movements relate to speech such as observing when people start or stop looking at each other while talking, and how this relates to what is being said. For example, looking away during a conversation could mean someone disagrees with what is being said.

## 2.1 Functions of Gaze

### 2.1.1 Interaction of Social and Referential Functions

One of the main reasons to look into someone's eyes is to determine their intended goal, since the eye direction of a person reliably signifies what they are going to act upon next. In an experiment (Phillips et al. 1992), eye contact was investigated in young normal infants who observed adults performing actions with ambiguous or unambiguous interpretations and found instant eye contact for ambiguous actions but rarely with unambiguous actions.

The phenomenon of 'eye contact effect', moderates certain facets of concurrent/immediately following cognitive processing (Senju and Johnson 2009). Developmental studies demonstrate proof of preferential orienting and processing of faces by means of direct gaze from early in life. The ability of 2- to 5 day old newborns to discriminate

direct and averted gaze was tested to measure the brain electrical activity to assess neural processing of faces when accompanied by direct (as opposed to averted) eye-gaze. The results illustrated that from birth human infants prefer to look at faces that engage them in mutual gaze and that, from an early age, healthy babies show enhanced neural processing of direct gaze (Farroni et al. 2002).

By the 4th week infants fixate and smile at eyes (Argyle and Cook 1976a). This visual interaction between the newborn and caregiver plays a major role in developing attachment (Nijenhuis and Bouchard Jr. 2007). The earliest gaze behaviours act as a foundation to build nonverbal and verbal communicative or social behaviours in later stages (Gillberg 1998). A visual attention cueing paradigm was used to study gaze in 2 year old children which consisted of eye movements and non-biological movements, the results suggested that the visual attention is cued by perceived eye movements (Chawarska et al. 2003). While on the other hand low-confidence conscious meta-cognitive knowledge and unconscious meta-cognitive knowledge through eye-gaze was measured in 3-5 year old children (Ruffman et al. 2001) which established that they were not aware of the knowledge conveyed through their eye-gaze. Children develop an increased understanding of social information and intentions carried through dynamic facial cues mainly changes in eye-gaze direction during middle childhood (Mosconi et al. 2005). Hence, it becomes exceedingly important to understand gaze behaviours for improving interactions involving robots/avatars.

### 2.1.2 Importance of Gaze Interaction in Dialogue

Many of the difficulties in interacting with robots/avatars have been attributed to the "uncanny valley" effect which hypothesizes that there is a relationship between the degree of an object's resemblance to a human being and the emotional response to it (Mori 2020). Several studies have been working on finding approaches that overcome the "uncanny valley" effect, but the focus on the appearance of a robot leaves a missing part that is the influence of non-verbal behaviour.

Thepsoonthorn et al. (2021) conducted an experiment to determine the relationship between the perceived human-likeness of a robot and participants' affinity towards it. Participants were asked to interact with a robot with different non-verbal behaviours ranging from no non-verbal behaviour (speaking only) to a model capable of gaze, head-nodding, and gestures. Results showed that for fixation duration there was a biphasic relationship between affinity and human likeness, with the longest fixation durations observed when the robot expressed non-verbal behaviours.

Terzioğlu et al. (2020) showed improvement in likeability and perceived sociability by using interactive gaze cues alone. Hence, in the design of social robots or understanding of human-robot interactions, the multifaceted robotic visual perception understanding of how to (and how not to) use social cues such as gaze becomes increasingly important.

Social cues can convey social information, such as attention, interest, agreement, dominance, and intimacy (Burgoon and Le Poire 1999). Eye contact, for example, is often seen as a sign of engagement and interest in a conversation, while avoiding eye contact can be seen as a sign of discomfort, dishonesty, or disinterest (Phutela 2015). Gaze can also be used as a way of signaling interest. Mutual gaze and lingering eye contact are often seen as signs of romantic interest (Moore 2010). Researchers have studied cognitive processing underlying gaze intend using eye tracking technology where individuals look at visual information in order to gain insight into their thought processes and decision-making strategies (Newell and Shanks 2014).

Social eye-gaze encompasses various intentional and expressive eye movements, including gaze aversions that may convey thoughtfulness. It also includes eye movements that serve a purpose during interactions, even if not explicitly communicative, such as when a child or a robot adjusts its view towards an object of interest, as long as these movements might be noticed by others. However, social eye-gaze doesn't involve eye movements that typically go unnoticed during social interactions, like isolated gaze actions, reflexive stabilization of one's viewpoint (vestibulo-ocular reflex), or visual processing routines that don't change the focus of the camera.

Figure 2.1: Session in progress (Original Scene)

In this context, several types of eye-gaze used in this thesis are defined as follows: **a) Mutual gaze**, often known as "eye contact" involves directing one's gaze towards another person's eyes or face, with reciprocation. Gazing at someone's face without receiving it in return does not constitute mutual gaze. **b) Referential gaze**, or deictic gaze, is when one's gaze is directed at an object or a specific location in space. This gaze can sometimes coincide with verbal references to the object but doesn't necessarily require speech. **c) Joint attention** involves sharing focus on a common object. It starts with mutual gaze to establish attention, moves to referential gaze to direct attention to the object of interest, and returns to mutual gaze to ensure a shared experience. **d) Gaze aversions** are shifts of gaze away from the primary point of focus, often a partner's face. These aversions can occur in various directions, influenced by the purpose behind the shift.

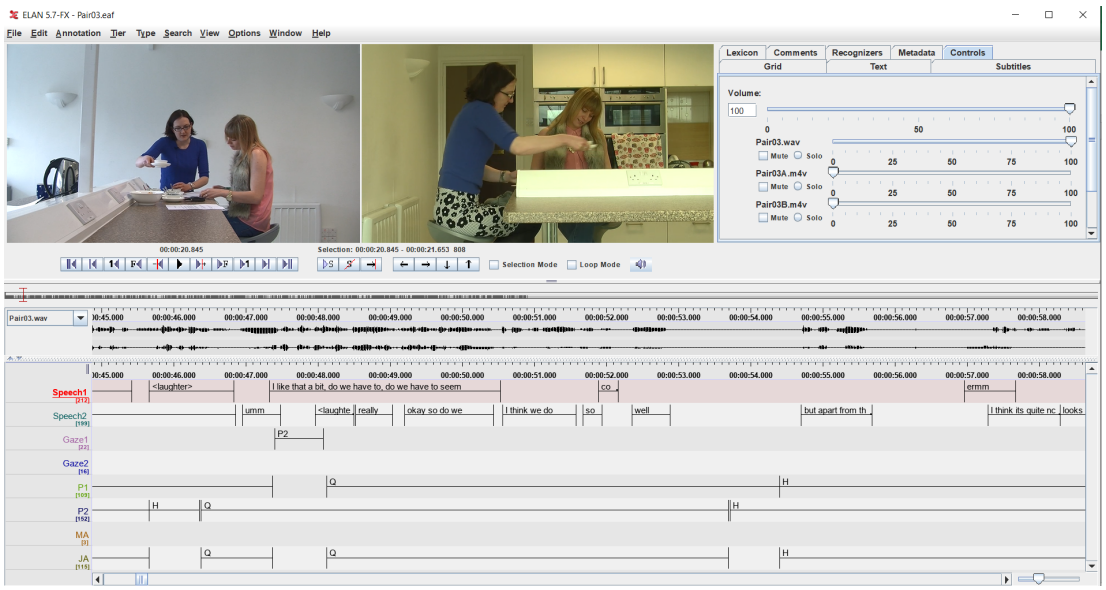



Figure 2.2: Speech and Gaze Annotation Labels

### 2.1.3 Human-Robot Gaze Interaction

Chevalier et al. (2019) showed human and robot faces as stimuli to participants and demonstrated higher gaze attention for human faces compared to robot faces. These were paralleled by attention processes obtained from (Wykowska et al. 2014) event-related potentials (ERP's) of Electroencephalography (EEG) as well as functional Magnetic Resonance Imaging (Ozdem et al. 2017). These prior findings indicated that the observed critical behavioural difference is mirrored by differential patterns of activation in the bilateral anterior temporo-parietal junction (TPJ) which is typically involved in attentional reorienting as well as mentalizing. It is important to note that the perceptional difference is because of lower gaze-cueing rather than the appearance of the robot itself.

Despite, the development of the new generation perception devices such as Kinect and gaze control systems implemented on a FACE humanoid social robot, which also included multimodal features like field of view, proxemics, verbal and nonverbal cues from the environment, the robot still does not direct its gaze appropriately and lacks the gaze-coordination required for smooth interaction (Zaraki et al. 2014).

Yun (2016) proposed a hybrid approach, a computational model for selecting a suitable interlocutor for robots containing gaze control and perceptual measures for social cues in a multiparty situation. Physical space and conversational intimacy were the two factors that were added to the model calculation for controlling for the social gaze control effect. Although some research has been done to understand the attention processes and their implementation in smart devices, the effects of temporal difference between gaze cues such as joint attention, mutual attention, referential attention, gaze aversion and gaze transition in social setting of spontaneous interactions has not been studied before (Khan et al. 2016).

## 2.2 Research Questions

Our motivation for designing a new type of gaze annotation is to make progress on answering the following research questions:

- Annotation: Is it feasible to annotate eye-gaze and elements of dialogue? Is this type of annotation useful for machine learning systems?
- Given some annotations, is it possible to predict, for example, dialogue acts, turn-taking or reference based on eye-gaze alone?
- Are there specific gaze patterns pertaining to different speech acts, and what influences them?
- What are the temporal patterns of gaze, and what do they look like?
- Can gaze inform or predict speech, and how does speech influence gaze?
- What are the qualitative and quantitative findings that could help build a better model of gaze in dialogue for a conversational robot or avatar to interpret human gaze behaviour and produce human-like gaze behaviour?

Answers to these research questions will contribute both to understanding the cognitive neuroscience of language and to the development of improved human-computer interaction.

## 2.3 Reviewing Gaze Annotation in Multi-modal Interaction

Research focusing on multi-modal interaction needs high quality annotation in order to obtain a detailed view of the interaction between visual, verbal and bodily features. A number of projects are interested in collecting and annotating video data due to increase in its demand. The current dyadic interaction experiment proposes a new form of annotation schema.

In the past five decades, the measurement of gaze points and eye movements with eye-tracking techniques during online behaviour has influenced multiple areas of research in psycholinguistics and psychology (Bhattacharyya 2018). This type of study mainly explores eye-gaze as a measure of cognitive processing with participants who are provided with a physical stimulus (e.g., picture or passage on a screen). There has been significantly less attention given to the role of eye-gaze in production, particularly in identifying the communication function of gaze and its ties to co-occurring utterances (Ho et al. 2015b).

The CID corpus (Bertrand et al. 2006) for interactional data in French is a corpus with single camera perspective profile view with a disadvantage of restricting access to gaze. Camera frontal views are an unnatural environment for the conversationalist, losing the advantages of the traditional face-to-face setting, and limiting the possibility of gaze-based interaction in dialogue. Corpora such as the Nottingham Multimodal Corpus (NMMC), the Swedish Spontaneous Dialogue Corpus (Spontal corpus), and the IFA dialogue video corpus (IFADV) in Dutch are examples of multiple-angle recorded data, but they focus either on social function or only on the referential functions (Brône and Oben 2014). In the current study, we combine these two functions where reference is part of the interaction. This is more common in a natural multi-modal dyadic dialogue and uses two cameras to gain access to multiple angle interactions that allow analyses of fine-grained, reliable behavioural features.

In multi-modal interaction, this creates a new area of research into gaze as a directive instrument or a disambiguation instrument and provides the opportunity to find potential correlations between gaze, facial expressions, and gesture. The results of a study (Jokinen et al. 2010) exploring non verbal signals for turn-taking and feedback in direct face to face interactions revealed that the gaze, head movement, or gesture primarily function as indexed signs linked to the whole context where they occur rather than symbols which carry meaning.

## 2.4 Methods and Materials

This section describes the multi-focal and multi-modal dialogue corpus used for annotation and presents the recording setup, task design, and annotation scheme used to code speech and gaze.

### 2.4.1 Data

By using video recordings, we are able to study the interactional dynamics specific to face-to-face dialogue along with understanding the collaborative processing and production of language. Hence, during the recording session, participants had to perform a collaborative task having a free range of conversations yielding natural multi-modal interaction. The data contains explicit information of social and referential gaze since the dialogue is open ended and task requires joint attention while performing the task.

Participants were twenty four dyads recruited from staff at the Good Housekeeping Institute (GHI, a consumer product testing organisation in the UK[1]). In each session a pair of participants taste-tested eight different types of hummus in the GHI test kitchen (see figure 2.1), and provided ratings on a single (shared) questionnaire.

1. https://www.goodhousekeeping.com/uk/the-institute/

They tasted and judged each product's appearance, aroma, flavour and texture, then provided a rating and described the product in three words by discussing with each other and coming to an agreement. Hummus was rated on a 9-point Hedonic scale ranging from dislike extremely to like extremely. Each session lasted for about 20-30 minutes and within this, participants organised their time themselves, e.g., deciding how long to take for tasting and rating each hummus, switching freely, choosing their strategies in performing the task and organising their interaction. The dialogues are task-directed rather than completely spontaneous. This type of dialogue is ideal for our purposes as it allows the internal dynamics of the conversation to be entirely free while the task creates an external trigger about which participants are communicating, meaning that both referential and interactive aspects of gaze ought to be present (which might not be the case in spontaneous dialogue as the topic under discussion may not include any shared referents available to visual attention).

#### 2.4.1.1 Recording Set-up

One important factor to be considered while annotating gaze is the duration of gaze fixation on a respective entity. Since people switch between objects extremely quickly the gaze behaviour may seem disorderly, so both directional and durational information needs to be recorded for a reliable categorisation of gaze episodes. Figure 2.1 depicts the configuration of the recording. Static external cameras were fixed with a profile and a frontal shot of each participant who sat at right angles to each other enabling us to record a clear gaze, tracking face, hands and body from multiple angles, resulting in a very rich representation of interaction providing access to extensive variability of multi-modal cues.

The core data of the multi-modal video recording comes from these fixed cameras that record the ongoing conversation, and the subjects are free to move and gesticulate from where they are seated. Consent is taken from the participants to record the complete session.

#### 2.4.1.2 Recording Devices

The set up required recording video from two perspectives (two cameras) and audio signals via microphones (unidirectional microphones that pickup audio from the respective speaker) which were worn by the participants.

- 2 fixed color cameras (JVC JY-HM360E Profesional HD camcorder, 1.56M pixel LCOS Color Viewfinder and 920K pixel LCD Display)
- 2 microphones (Schertler Cello Microphone, Output impedance: 4.7 kOhm at 1 KHz, Frequency range: 20 - 20,000 Hz)

The fixed camera recordings and the wave-forms of the microphones were synchronised prior to annotation.

## 2.5 Annotation

### 2.5.1 Annotation Tool

Data was annotated in ELAN (Berez 2007), a tool that provides a framework for annotation of audio and video recordings. This enables us to have precise time-alignments and hierarchically organise annotation tiers as outlined below. ELAN records data in a stand-off XML format.

- ELAN : The audio and video files of each session are added to the software separately, which contains profile and front shot of the participants showing the two videos side by side, as shown in figure 2.3. The software is compatible for using several video files along with the audio wave file. We used two video files since each session had dyadic interactions. The video files were time aligned to make sure the beginning of each videos were synchronized. These appear next to each other on the left-top corner of the window right below the menu. The audio file contains the waveform of the spoken speech and is convenient to annotate the speech data. This file once uploaded appeared right below

the videos. The software then has a blank space below to code the annotation data where we encoded different different tiers separately for speech and gaze along with the associated audio files. Because of several export options and comparability it can be used for further statistical processing.

The synchronized videos were played to listen to the speech, then the beginning the end of the utterance is marked accordingly and typed along the line of the duration of the speech. The wave form file assisted in marking the onset and offset of the utterances accurately. These were played again to double check for errors. Speech of the dyads in each session were annotated in two tiers as Speech1 and Speech2. Speech1 (S1) and Speech2 (S2) contains all the utterances of participant 1 and 2 respectively during the session.

- Transcription: For the video transcription general norms and principles of Gesprächsanalytisches Transkriptionssystem (GAT) were considered (Selting et al. 1998). The orthographical transcription for each participant was done in two separate tiers speech1 and speech2. These tiers contained metadata indicating the beginning and end of the excerpt encoded with respective spoken utterance per unit, in few occasions a short description of the interactional context in unicode (<>) such as laughter, cough, uhm, etc.

### 2.5.2 Gaze Annotation

As shown in figure 2.3, gaze was annotated in six tiers.

1. **Referential gaze (P1, P2)** is gaze directed at an object or a location in space. Gaze information of each participant was annotated in separate tiers, as P1 and P2 (Participant1 and Participant2). The labels were the objects in their shared visual field such as bowl of hummus (H), Questionnaire (Q), breadstick (B), etc., or a location in space (Z).

2. **Joint attention (JA)** is sharing attention focus on a common object (Mundy 2017). The overlap between P1 and P2 indicates where both participants are fixating on the same thing in the shared visual field. This tier was generated by temporal and object overlap of the P1 and P2 tiers. The excerpt was encoded when both the participants attend to the same stimulus/fixate on a particular object. During statistical analysis it becomes easier to understand on which objects were the most or least fixations on.
3. **Gaze1 and Gaze2 (G1, G2):** For each participant, these encoded times they were looking at their partner. Gaze1 is annotated as P2 when P1 is looking at P2. Gaze2 is annotated as P1 when P2 is looking at P1.
4. **Mutual attention (MA)** is nothing but eye contact, it occurs when the gaze of both conversationalists is drawn to each other's face or eyes (Argyle et al. 1994). Mutual Attention is obtained by temporal overlap in G1 and G2.
5. **Gaze aversions** are the shifts in the the main direction of gaze that is away from the partner's face (Korkiakangas 2018). This was obtained by *lack* of overlap in G1 and G2. This denoted P1 looking at P2, while P2 was looking at something else, and vice versa.

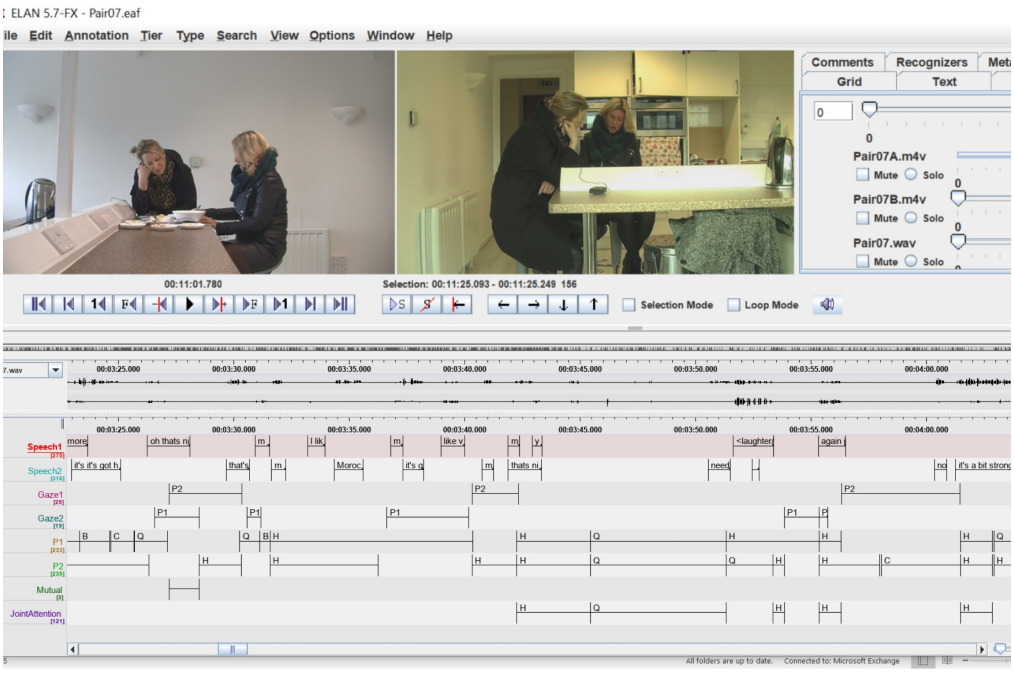



Figure 2.3: ELAN Annotation

Table 2.1: Annotation summary by duration (D). Tiers represent the various gaze acts annotated for 11 minutes and 40 seconds. The Number of annotations, minimum, maximum, average, median, total Annotation duration are in seconds. Annotation duration is a percentage of time spent on a specific gaze act during the session.

| Tiers | # of Ann. | Min D | Max D | Avg D | Median D | Total Ann. D | Ann. D % |
|---|---|---|---|---|---|---|---|
| Speech1 | 157 | 0.29 | 4.32 | 1.13 | 0.87 | 177.68 | 25.37 |
| Speech2 | 173 | 0.23 | 4.33 | 1.04 | 0.82 | 180.08 | 25.71 |
| Gaze1 | 29 | 0.31 | 8.10 | 2.18 | 1.66 | 63.29 | 9.04 |
| Gaze2 | 19 | 0.363 | 5.22 | 1.92 | 1.67 | 36.47 | 5.21 |
| P1 | 232 | 0.23 | 25.91 | 2.69 | 1.56 | 624.15 | 89.11 |
| P2 | 235 | 0.17 | 25.91 | 2.77 | 1.69 | 649.66 | 92.75 |
| MA | 3 | 1.00 | 1.69 | 1.33 | 1.31 | 4.00 | 0.57 |
| JA | 121 | 0.25 | 25.91 | 3.58 | 2.13 | 433.35 | 61.87 |

## 2.6 Preliminary Results

As shown in table 2.1, in the approximately 11 minutes 40 seconds of one dialogue that has thus far been annotated, the participants had equal amounts of speech (157 versus 173 utterance events equal to 25/26% of the time each).

Interestingly, P1 spent somewhat more time looking at P2 than P2 did at P1 (9% to 5%) and these looking events overlapped (such that the participants were looking directly at each other) only on 3 occasions (see the MA row in table 2.1), and for less than 1% of the total duration of the annotated interaction.

In line with Argyle and Cook (1976a) and Rossano (2012), the listener looked at the speaker more frequently than the other way round (4.9% of the time compared to 2.8% of the time). Further investigation of these eye-gaze events is needed to see if they co-occur with particular dialogue acts or points where a floor change may occur, as suggested by Brône et al. (2017b), but if so, this is potentially useful information to a dialogue system.

However, participants' visual attention was far more often on one of the objects in the shared visual field, such as the hummus or the questionnaire (P1 89%; P2 92%) with these annotations overlapping for 61% of the duration of the annotated dialogues (JA) indicating that participants were looking at the same thing more often than not. Interestingly, while both participants spent a lot of time looking at both the hummus and the questionnaire (P1 H: 34%; P2 H: 36%; P1 Q: 40%; P2 Q: 44%) they had joint attention on the questionnaire nearly twice as often as the hummus (JA H: 20%; JA Q: 35%) showing how gaze behaviour is affected by the particular constraints of the sharedness of the sub tasks even within a dialogue (here, the rating is specified to be a joint action, while the tastings can be carried out in parallel).

In terms of reference resolution, based on the intuition that to use gaze behaviour to aid reference resolution it is necessary to look at the other participant whilst their visual attention is on the referent, we compared the overlaps where P1 looked at P2 while P2 looked at something else, and vice versa. Interestingly, P1 looked at P2 while P2 looked at something else for 8% of the time, but the inverse was true only 4% of the time. Further investigation into how this maps to linguistic information, for example, whether P2 was more ambiguous in their speech, is pending.

## 2.7 Analysis

We conducted a post-annotation qualitative and quantitative analysis. For qualitative measurement, we considered the onset and offset of gaze in relation to speech, and the numbers for quantitative analysis were exported directly from ELAN.

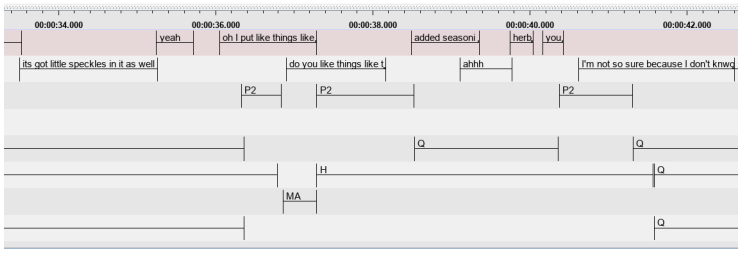



Figure 2.4: Number of Annotations

## 2.7.1 Qualitative Analysis

The qualitative data contains excerpts (see figure 2.4) of speech segments from ELAN. As shown in figure 2.2 the speech information of each participant is followed by gaze tiers.

### 2.7.1.1 Pragmatic gaze assessment during disagreement

We assessed how the shifting of gaze on the partner helped to check if the statement made was being accepted or not. *Excerpt 1.1:* referring to the appearance of the hummus

***P2: it's got little speckles in it as well***

P1: ————————- H ————————-

P2: ————————- H ————————-

***P1: yeah, oh I put like things like that***

P1: — H — P2 —- MA ————————

P2: —— H ———MA————————

***P2 : do you like things like that***

P1: —- MA ———- P2————

P2: ——MA —— H————————-

***P1 : added seasoning, herbs, you?***

P1: ———— P2————————

P2: ————— H————————

***P2 : I'm not so sure cause I don't know***

P1: ———P2 ——— Q———————

P2: ——— H——— Q———————-

When P2 says "it's got little speckles in it as well", the gaze attention of both P1 and P2 is on the hummus, as they are evaluating its appearance. In the next utterance, P1 responds in favour of the appearance, during which their gaze transitions to their partner. Their partner shows understanding by reciprocating the gaze, leading to the establishment of Mutual Attention by the end of P2's utterance. This is followed by P2 asking a question to re-assess P1's opinion, and here we can see that P1 continues to look at P1, but P2's gaze shifts back to the hummus, in an example of active gaze avoidance during disagreement. P1 pursues a verbal answer by asking "you?", while still maintaining their gaze on their partner, which changes to looking at the questionnaire after obtaining a negative response from P2, whose gaze attention has either been on the hummus or the questionnaire, avoiding looking at P1.

This is one of many examples that suggests that gaze can reveal information pertaining to negation or disagreement well before declaring it verbally.

#### 2.7.1.2 Pragmatic gaze assessment during agreement

*Excerpt 1.2:* referring to the texture of the hummus

***P1: plain***

P1: —- Q —-

P2: —- H —-

***P2: yeah, its pretty normal isn't it***

P1: — Q ———- MA ——— Q ——

P2: — H ———- MA ———- Q —-

***P2 : do you like things like that***

P1: —- MA ——- P2———

P2: —MA —— H———-

***P1 : yeah***

P1: — Q —-

P2: — Q —-

While discussing about the texture of the hummus, P1 says "plain", where the gaze attention of P1 is on the questionnaire and P2 on the hummus. Right before P2 utters "yeah, its pretty normal isn't it", gaze attention is on the partner and P2 looks at P1. It is followed by an agreeing utterance from P1 "yeah". This excerpt shows that mutual gaze was established which reveals information about agreement well before declaring it verbally.

#### 2.7.1.3 Gaze Check

*Excerpt 2:* tasting the hummus

***P1 : feel a bit of pepper in there***

P1 : ——-H———P2———-

P2 : ——-H———————-

***P2 : Um-m***

P1 : –P2–Q–

P2 : —-H—-

In excerpt 2, the gaze from the joint attention of P1 shifts to P2 before completing the sentence, representing a "Gaze Check" phenomenon. Here, P2 still continues to look at the hummus while responding. But P1's attention shifts from P2 to a different entity after receiving agreement in response.

#### 2.7.1.4 Gaze Prediction

*Excerpt 3:* rating the hummus

***P1 : like it moderately***

P1 : —-Q——-P2——-

P2 : ———H————

***P2 : yeah I would say like it moderately***

P1 : —-P2————H————-

P2 : ————Q————-

This is another example of 'Gaze Check' from P1 to understand if P2 agrees to rate the hummus, "like moderately". Another important observation is that the gaze shift of P1 after looking at P2, is toward the object that P2 has fixated on for the entire duration when P1 was looking at P2. This type of gaze attention could further help in predicting the next movement of gaze towards the intended object of interest without any assistance from speech.

#### 2.7.1.5 Overlapping Gaze Transition

The P1 and P2 tiers give us most of the gaze information of each participant except Mutual Attention and attention on the partner. It is interesting to think about what factors might influence the subsequent gaze shift. Does it depend on speech? If there is no speech then does that mean that there is no shift in gaze at all? Here is an example of one such phenomenon:

*Excerpt 4:* gaze shift in the absence of speech

P1: —-Q————-H———B—————H——

P2: —Q————H———Q——B——H————

P1's gaze attention follows P2's just 2 ms after P2 looks at the Questionnaire. Once P2's attention is on Hummus, approximately 3 ms later, P1 joins again before looking at the breadstick. P2 briefly focuses on the questionnaire before joining P1, and when P2 shifts gaze, P1 continues to join. This is a very interesting pattern seen consistently throughout the experiment, where the overlap occurs just a few milliseconds before joint attention.

#### 2.7.1.6 Referential gaze

Excerpt 5: reaching for a breadstick

***P1 : okay breadstick***

P1 :———-B—————-

P2 :————Q———-

***P2 : sure here it is***

P1 :—-B—————-

P2 :—-B—————-

***P2 : do you mind my hands on it***

P1 :————B———————

P2 :————B———————

The example above explains how speech can influence gaze transition. Particular speech utterances especially ones mentioning the objects in the shared space drives gaze attention to the particular object away from the initial point of interest. For example, the phrase "okay breadstick", shifted the gaze of the partner to look away towards the location of the bowl containing breadsticks. These would enable gaze prediction based on referential speech and gaze attention. It is also important to note that P1 looked at the breadstick before verbally uttering the word.

### 2.7.2 Quantitative Analysis

The number of times the variables were annotated for different conditions is noted here. The total annotated data (Table 4.2) contains approximately 45 minutes (2700 seconds) of 8 participants in pairs, and the annotation comprises of 1,700 spontaneous speech utterances and 2,300 annotations for various types of gaze. A total of five dependant variables were measured across various gaze behaviours.

Table 2.2: Annotations summary. The tiers are Speech (S), Gaze at partner (G), Participants referential gaze (P), Mutual Attention (MA), Joint Attention (JA). The minimum, maximum, mean and total duration are in seconds.

| | Duration of annotations | | | | | Number of |
|---|---|---|---|---|---|---|
| Tier | Min | Max | Mean | Total | % | annotations |
| S | 0.15 | 3.98 | 0.97 | 271.61 | 24.97 | 1706 |
| G | 0.27 | 4.90 | 1.44 | 48.55 | 4.29 | 269 |
| P | 0.17 | 20.83 | 2.98 | 527.62 | 41.84 | 1529 |
| MA | 0.41 | 2.73 | 1.09 | 17.66 | 1.99 | 51 |
| JA | 0.15 | 20.43 | 2.90 | 379.56 | 32.08 | 519 |

As shown in Table 2.2, the fewest gaze annotations (51) were coded in mutual attention (MA), with most in referential gaze (P: 1529). Interestingly, when we look at the total annotation duration for each individual compared to their partner, although there is an extremely high correlation between the amount of speech of each participant in a dyad ($r = 0.97, p < 0.001$) and the amount of gaze at reference

objects between participants ($r = 0.96, p < 0.001$), showing the symmetry of the task, there is no such correlation between participants gaze patterns to each other ($r = -0.02$). There were also no correlations between amount of speaking and gaze at partner (self; $r = 0.009$ or other; $r = 0.002$).

#### 2.7.2.1 Fixation Duration

The duration of fixation is the total amount of time participants looked at different entities. Data from the separate tiers of Speech and Gaze were combined. Hence, all eight participants devoid of their interaction partner were taken into account to measure the co-relation between different entities.

*Joint Attention* For Joint attention, minimum and maximum fixation durations were under 2 miliseconds to 20 seconds with a mean of 2.9 seconds. Another aspect here is that the joint attention on one particular object did not last long (avg 2.9 sec); instead, there was a constant gaze transition and interaction with the surroundings. Overall the total annotation duration was approximately 380 seconds (out of 2700). This showed that the participants spent nearly as much time (JA) looking at the same object together, as they did looking at objects their partner was not looking at (P, 527 seconds) with equivalent average durations, showing how coordination of gaze is critical in a task requiring coordination with a partner in other respects, such as coordinating on which aspect of the task was being undertaken on a moment to moment basis.

*Mutual Attention* The duration of mutual attention accounted for as little as 0.4 seconds to 2.7 seconds. On average participants looked at each other for approximately one second at a time which is extremely short compared to Joint Attention (JA) or Gaze at the partner (G). In total, the annotation duration on Mutual attention was only 17.6 seconds (of 2700 seconds). Looking at the partner eye-to-eye for an extended amount of time can lead to an uncomfortable situation. This could

result in the participants spending the least amount of time making eye contact and avoiding uncomfortable eye contact, which could be defined as "eerie mutual attention". This is a novel observation that needs to be taken into consideration while improving gaze interaction in robots.

*Individual Attention* The attention on the partner while the partner looked somewhere else accounted for 4.29% (see row G in table 4.2), for which the annotation duration was 48.5 seconds (of 2700 seconds). Mutual attention where the participants looked at each other was notably rare in the data.

Our annotated data, a detailed coding of gaze in conversation, adds another dimension in understanding multimodal dyadic interaction, showing that gaze is a complex non-verbal mechanism that still follows a very coordinated pattern in interaction upon analysis. These patterns are especially evoked in social and referential scenarios (Somashekarappa et al. 2020). Gaze behaviours shift in split seconds; researchers have been able to study these behaviours through psychological techniques, but without large amounts of observational data.

This multi-modal corpus of transcribed speech is annotated and synchronized with eye-gaze details as described above and is produced in ELAN (.eaf) format. It contains high resolution video and audio files, and the video data is in M4V format while the audio data in WAV-format. The corpus presented in this chapter is a contribution to the growing quality of multi-modal research.

# Chapter 3

# Gaze Interaction with Laughter Pragmatics and Coordination

Our conversations are highly coordinated, with synchronisation occurring even across modalities (Fusaroli and Tylén 2012; Dale et al. 2013). Both laughter and gaze have been the object of in depth independent analyses and their crucial role in managing and coordinating interaction is not in doubt. Both gaze and laughter are perceivable actions (termed *visible/audible acts of meaning* in Bavelas et al. 2002) which affect the unfolding of the upcoming dialogue (Mazzocconi et al. 2020). While there is some work on the interaction of smiles, laughter and gaze in relation to humour (Gironzetti 2017; Brône 2020), less is known about the relation of laughter and gaze when this is not related to humour, but rather to what we call *social incongruity*. The only exception we are aware of is Romaniuk (2009), who take a micro-analytic approach on the use of gaze to decline a laughter.

Figure 3.1: Data collection setting from Chapter 2

An example of the fine coordination between laughter and gaze is presented in (1), where we see the onset of gaze at the partner from A shortly before the onset of A's laughter. The onset of A's laughter is then shortly followed by B gazing at A, just before joining B's laugh with her own.

(1) GHI Corpus (**Somashekarappa.etal20**), Pair03 (00:02:17)[1]

**A:** `It's "like slightly"?..`

**B:** `yeah ((shrugs))`

**A:** `I like hummus|||<laughter>`

**B:** `yeah<laughter>`

It is clear that both gaze and laughter are crucial elements to be taken into account when implementing algorithms for Embodied Conversational Agents (ECA) (Ochs and Pelachaud 2013; Becker-Asano and Ishiguro 2009), both for what concerns the interpretation of the users' dialogue acts and for what concerns their own behaviour, in order to have ECAs more competent from a pragmatic perspective and also more human-like in terms of emotional displays, where this is desirable.

In the current work, we aim to fill this gap by investigating the following, to our knowledge, as yet unexplored questions. The answers are to provide insights into how meaning is constructed in interaction across modalities, as well as provide empirical data for the implementation of human-like ECA:

Q1 Does the laughing participant gaze at their partner, differ in terms of probability and timing, depending on the pragmatic function performed by the laughter?

Q2 Is the interlocutor's gaze at the partner influenced, in terms of probability and timing, by the type of laughter produced by the partner?

Q3 Does gaze play a significant role in laughter coordination and alignment between participants?

---

1. Speech that overlaps with gaze at partner is shown in bold, with continuation of gaze marked by |||.

The chapter is structured as follows: in Section 3.1 we briefly present a literature review about laughter and gaze studies that constitute the background motivation of our questions, stressing how the study of laughter and gaze is increasingly important for ECA design. In Section 3.2 we explicitly state our hypotheses in relation to our motivating questions presented above, while in Section 3.3 we outline the method chosen to test them. In Section 3.4 we present our results, discussing them in Section 3.5 in the light of literature on gaze, speech turn-taking and interactional studies. We conclude in Section 6 highlighting the importance of the insights gained from our exploration for the implementation of human-like ECA.

## 3.1 Background

### 3.1.1 Laughter in interaction

Laughter production in conversation is not exclusively related to humour or to the appreciation of a *pleasant incongruity*. Many studies, particularly in conversation analysis, have shown its crucial role in managing conversations at several levels: dynamics (turn-taking and topic-change), lexical (signalling problems of lexical retrieval or imprecision in the lexical choice), pragmatic (marking irony, disambiguating meaning, managing self-correction) and social (smoothing and softening difficult situations or showing (dis)affiliation and marking group boundaries) (Glenn 2003; Jefferson 1984; Mazzocconi et al. 2020).

In friendly conversation, interactants typically aim at an optimal level of cooperation and equilibrium avoiding direct disaffiliation as much as possible (Pomerantz and Heritage 2012). Nevertheless, social interactions often require the production of speech acts that can make this equilibrium unstable or at risk (Raclaw and Ford 2017). Following Mazzocconi et al. (2020), we refer to any situation in which a clash is perceived between the current situation and a social norm and/or comfort as a *social incongruity*. Laughter, which can be used for bonding and showing friendliness, often comes in handy to cope with these situations. For example, in the case of

embarrassment or awkward silence, laughter can smooth the situation; when putting forward a criticism, a laugh can soften the statement; or when asking a favour or advancing a proposal, a laugh can induce benevolence from the listener (Jefferson 1984; Glenn 2003; Petitjean and González-Martínez 2015; Holt 2012).

Moreover laughter has been identified also as an attract-attention device, both in children (Stevenson et al. 1986; Reddy et al. 2002) and adults (Pinheiro et al. 2017), especially for its emotional salience, making therefore extremely relevant to explore its effect on interactants' gaze behaviour in natural conversation.

### 3.1.2 Gaze in interaction

The role of gaze in maintaining the conversational flow and coordinating dialogue acts is not in doubt. While many works have argued for the importance of individual gaze for fine regulation of turn-taking (Duncan 1972; Goodwin et al. 1980), some scholars actually highlighted a lack of systematic relation between gaze and turn-taking (Beattie 1978; Torres et al. 1997; De Ruiter 2005), proposing rather that gaze might function to solicit a response (Harness Goodwin and Goodwin 1986; Bavelas et al. 2002), which is not necessarily a speech turn (Rossano 2013). More specifically, it has been argued that turn-taking is only a partial explanation for gaze behaviour in conversation, and that our study of gaze has to take into account both turn-taking and informational structure (Torres et al. 1997; Bonin et al. 2012).

Despite turn-taking not being the only function performed by gaze, and the fact that not all turn shifts are accompanied by gaze towards the listener, it has been consistently observed that there is a tendency for listeners to display more gaze at the speaker during the course of dyadic interaction, while the speaker tends to direct their gaze at the listener mainly towards the end of their speaking turn (Kendon 1967b; Duncan and Fiske 1979). In this way, when a speaker gazes at the listener mutual gaze is attained (Goodwin et al. 1980), a brief mutual *gaze-window* (Bavelas et al. 2002) is established, and a change of floor may occur, having the previous

listener looking away as they begin their speaking turn (**Somashekarappa.etal20**; Kendon 1967b; Rossano 2013). Gaze patterns to the interlocutor have also been found to differ depending on the speech act they accompany and on their pragmatic function (Sandgren et al. 2012; Mirenda et al. 1984; Rossano et al. 2009)

Of interest in the study of gaze in interaction is not only gaze directed at one's partner, but also its absence or avoidance (Rossano 2013). For example, using a microanalytic analysis, Romaniuk (2009) observed how gaze aversion can be used to decline laughter and terminate its relevance; while Kendrick and Holler (2017) report that most preferred responses are produced with gaze toward the questioner, whilst most dispreferred responses are produced with gaze aversion. Moreover, it has been proposed that gaze aversion could also be explained (or influenced) by social stress (Stanley and Martin 1968), with evidence from patients with social disorders (Schneier et al. 2011). Conversely, results from other studies (Doherty-Sneddon and Phelps 2005) suggest that cognitive load has the most impact on gaze aversion (Glenberg et al. 1998). The latter hypothesis is based on the fact that visual cues are an important source of information and facilitate conversation, but cause higher cognitive load. This explanation seems to be supported by results observing more gaze aversion in the initial phase of request formulations (Sandgren et al. 2012; Kendon 1967b), and by speakers showing less fluency when forced to constantly look at their listener (Beattie 1981), even though these result could also be explained by the social stress factor.

### 3.1.3 Gaze and Laughter in ECA

Recently, there has been a growing research interest both on gaze and other non-verbal expression, especially in Affective Computing community, for the implementation of ECAs which are more competent from a pragmatic perspective and able to process and produce appropriate emotional responses (Stevens et al. 2016; Bailly et al. 2010; Lee and Marsella 2006; Niewiadomski et al. 2009). Virtual agents benefit from a detailed analysis of multimodal input and output patterns observed

during human-human interactions and from the interplay with their cognitive interpretation. Bailly et al. (2010) established a basis for a context-aware eye-gaze generator for an ECA. In order to develop an improved gaze generator we should isolate the significant events detected in the multi-modal scene that impact the closed-loop control of gaze. Lee and Marsella (2006) discuss the interpersonal role of gaze in interaction to signal feedback and direct conversation flow which current ECAs still lack. Simultaneously, in a dynamic environment, even the state-of-the-art ECAs struggle to direct gaze attention to peripheral movements. An embodied conversational agent should therefore employ social gaze not only for interpersonal interaction but also to possess human attention attributes so that its eyes and facial expression portray and convey appropriate distraction and engagement behaviours. ECA simulations for face-face conversation are mainly dyadic and turn allocation using gaze signals Gu and Badler 2006. Non-verbal behaviours also can help create a stronger relationship between the ECA and user as well as allow applications to have richer, more expressive characters. Overall, appropriate nonverbal behaviours should provide users with a more immersive experience while interacting with ECAs, whether they are characters in video games, intelligent tutoring systems, or customer service applications.

Becker-Asano and Ishiguro (2009) evaluated the role of laughter in perception of social robots and indicated that the situational context, determined by linguistic and non-verbal cues (such as gaze) played an important role. In particular, in their experiments, the Geminoid robot's direct gaze at the participant while laughing led to the perception of the robot's laughter as "laughing at someone" rather than "laughing with someone". Nijholt (2002) discusses the challenges of integrating humour into ECAs, and existing integration of smiling and laughter in ECA is typically triggered by a joke told by a user or an agent (Ding et al. 2014; Ochs and Pelachaud

2013). El Haddad et al. (2019) looked at the mimicry of smiles and laughs between interlocutors, which also might be used as the basis for an ECA's behaviour. Urbain et al. (2010) take a similar perspective, equipping ECAs with the capability to join in with a conversational partner's laugh.

Our work will provide empirical data useful for the implementation of systems able to engage in multimodal interaction, profiting of the availability of cross-modal cues (i.e. gaze and laughter).

## 3.2 Hypotheses

Based on the literature reviewed above, our predictions in relation to the three main questions motivating our work are the following:

H1 Based on the social stress hypothesis of gaze aversion (Stanley and Martin 1968; Schneier et al. 2011), and on research showing that gaze aversion is more likely when subjects are offering a dispreferred answer (Kendrick and Holler 2017), we expect laugher gaze towards the partner to be less likely if the laugh produced is related to *social incongruity*/discomfort (both around the onset and offset of the laugh) rather than to *pleasant incongruity*.

H2 On the basis of studies indicating that laughter can function as an attention getting device (Stevenson et al. 1986; Reddy et al. 2002; Pinheiro et al. 2017), we hypothesise that interlocutors will direct their gaze at the laugher after laughter production.

H3 Given the role of gaze in soliciting a response from one's partner (Rossano 2013; Bavelas et al. 2002), we expect laughs where one participant joins in with another's laugh (*joining in* laughs) to be preceded by an "inviting" gaze from their partner (as in Extract (1)). Similarly we expect the interactant joining the laugh to gaze herself at the partner, in order to instantiate the "gaze window" which may enable a turn shift (Bavelas et al. 2002).

| | Laughable Type | | | | Laughter positioning | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Pair | Pleasant | Social | Pragmatic | Friendly | Isolated | Antiphonal | Coactive | Total | Minutes |
| 03 | 27 | 19 | 1 | 5 | 28 | 20 | 4 | 52 | 10 |
| 07 | 11 | 10 | 3 | 1 | 15 | 4 | 6 | 25 | 10 |
| 15 | 2 | 5 | 0 | 0 | 5 | 2 | 0 | 7 | 3 |
| Total | 40 | 34 | 4 | 6 | 48 | 26 | 10 | 84 | 23 |

Table 3.1: Distribution of different laughter annotations across dyads and minutes of interaction analysed.

## 3.3 Materials and methods

### 3.3.1 Corpus data

Our data consist of 23 minutes taken from three female-female dyadic interactions from the Good Housekeeping Institute (GHI) Corpus (**Somashekarappa.etal20**). The GHI corpus contains video and audio of pairs of participants discussing and rating different kinds of hummus on a chapter questionnaire (see Figure 3.1). We annotated the interactions for laughter and gaze as described in the following sections.

### 3.3.2 Laughter Annotation

Our annotations have been conducted using the software ELAN (Brugman and Russel 2004). Coding was carried out by the first author watching and listening to a video until a laugh occurred. The coder then marked the onset and offset of the laugh, and, following the annotation scheme proposed in Mazzocconi et al. (2020), annotated the laughter's form, temporal sequence in relation to speech and others' laughs, context of occurrence, laughable it was related to (i.e. the argument of the laughter), and pragmatic function. In the current chapter we focus on two of these features: (1) the type of incongruity present in the laughable, (2) the positioning of laughter in relation to the partner's laughter (laughter coordination).

We assessed the agreement on laughter identification and segmentation (start-time and end-time boundaries) using the Staccato algorithm implemented in ELAN (Lücking et al. 2011), having two annotators marking 70% of the data. We run the analysis with 1000 Monte Carlo Simulations, a granularity for annotation length of 10, and $\alpha = 0.05$. The degree of organisation is 0.8386.

#### 3.3.2.1 Laughable classification

Following Mazzocconi et al. (2020) we consider laughter as an event predicate, the meaning of which is constituted by two dimensions: the laughable and arousal, which we do not consider in the current chapter. By laughable we mean the argument the laughter predicates about, an event or state referred to by an utterance or exophorically (Glenn 2003). Different kinds of laughable can be distinguished based on whether they contain an incongruity or not, and if so, which kind of incongruity (see (Ginzburg et al. 2020) for a formal definition of incongruity). The annotation categories are as follows:

1. **Pleasant incongruity** is a clash between the laughable and certain background information perceived as witty, rewarding and/or somehow pleasant. Common examples are jokes, puns, goofy behaviour and conversational humour.
2. **Social incongruity** is a clash between social norms and/or comfort and the laughable. Examples include social discomfort (e.g. embarrassment or awkwardness), violation of social norms (e.g., invasion of another's space, asking a favour), or an utterance that clashes with the interlocutor's expectations concerning one's behaviour (e.g., criticism).
3. **Pragmatic incongruity** arises when there is a clash between what is said and what is intended. This kind of incongruity can be identified, for example, in the case of irony, scare-quoting, hyperbole etc. Typically in such cases laughter is used by the speaker in order to signal changes of meaning within their own utterance.

4. **Friendliness** refers to cases where no incongruity can be identified. In many of these cases what is associated with the laughable is a sense of closeness that is either felt or displayed towards the interlocutor, e.g., while thanking or receiving a pat on the shoulder.

In the current work we focus on the observation of gaze patterns accompanying laughs related to *pleasant incongruity* compared to *social incongruity*, as these are the most frequent kinds of laughable across contexts of interaction and languages (Mazzocconi et al. 2020) (see also Table 3.1). These categories are also the furthest apart in terms of pragmatic function, since *pleasant incongruities* are related to something pleasant and rewarding, whilst *social incongruities* are related to potential discomfort and unpleasantness.

Our dataset is therefore constituted of 74 laughs: 40 related to *pleasant* and 34 to *social incongruity*. 60% of the data (50 laughs) were annotated by two of the authors. The inter-annotator percentage agreement was 82%, with Krippendorff's $\alpha = 0.69$.

#### 3.3.2.2 Laughter coordination

In our annotation we distinguish 3 classes pertinent to the sequential distribution of the laughter in relation to laughs produced by the partner:

1. **Isolated laughter:** a laugh not preceded by or co-occurring with another laughter;
2. **Antiphonal laughter:** a laugh shortly following a laugh from the partner, starting during the partner's laugh, or within one second after its offset;
3. **Coactive laughter:** a laugh with the same onset time as a laughter from the interlocutor. We did not give an exact time definition for shared laughter onset, rather we relied on annotators' intuitions. We tested whether this intuitive notion was appropriate by calculating inter annotator agreement, which was high. Laughs which were considered to be coactive had a relative onset time of less than 100ms.

Inter-annotator agreement for this variable conducted over 60% of the data (50 laughs) reached 85.7%; Krippendorff's $\alpha = 0.76$.

### 3.3.3 Gaze Annotation

Following **Somashekarappa.etal20empty citation** as discussed in 2, the gaze annotation was coded for four aspects:

1. **Participant1 and Participant2 (P1, P2):** The gaze of each participant to an object, for example, Hummus (H), Questionnaire (Q), Breadstick (B) etc
2. **Joint attention (JA):** Looking at the same object, obtained by temporal and object overlap in P1 and P2.
3. **Gaze1 and Gaze2 (G1, G2):** For each participant, these encoded whether they were looking at their partner.
4. **Mutual Attention (MA):** Looking at each other, obtained by temporal overlap in G1 and G2.

In the current work we explore only gaze at each other (G1, G2), leaving a more fine grained analysis of whether the gaze reciprocity (MA), and questions about gaze to objects including joint attention (JA), for future work.

### 3.3.4 Data extraction

In order to perform our analysis we made use of the ELAN Analysis Companion (EAC) software (Andersson and Sandgren 2016) to conduct event-related analysis. Our dependent variables is *Gaze at partner* (G1 and G2), both from the laugher and to the laugher from the other participant. In order to address questions (1) and (2) we used *Laughable Type* (whether a laughter was related to a *pleasant* or *social incongruity*) as predictor (sec. 3.3.2.1; while in order to address question (3) the predictive variable is *Laughter coordination* (isolated, antiphonal, or coactive) (Sec. 3.3.2.2). Each analysis is centered either on the onset or the offset of the

laughter. Following Andersson and Sandgren (2016) we considered a time window of 3000 milliseconds (i.e. 1500 seconds before and after the laughter onset/offset). We selected 10ms resolution, using a "first come first served" overlap handling and binned the data at intervals of 100ms, rounding up any fractions to 1.

Given the type of gaze $G$ and the type of laughable $L$ (or the laughter coordination classes for question (3)), the probability of gaze before or after the event for a given time window "bin" $b$ is calculated as follows:

$$P_b(G|L) = \frac{\sum_{i=1}^{N} P(g_b|l_i)}{N} \tag{3.1}$$

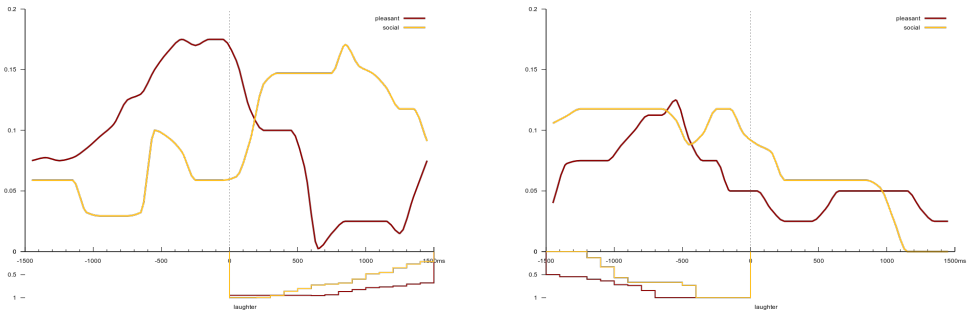
where $N$ is the total number of laugh events, and $P(g_b|l_j) \in \{0,1\}$ is the probability of gaze for a single bin $b$ for a given event $l_i$.

For each of our models, reported below, we ran a mixed-effect logistic regression in R, using the *glmer* function from the *lme4* package (Bates et al. 2015), with subjects as a random factor.[2] The dependent variable, *Gaze* (either at the partner or the laugher) was treated as a dichotomous dependent variable (present / not present) for each 100ms bin of the time window of interest (3000ms centered around the onset of the laugh)[3]

Together with *Laughable Type* (Q1 and Q2), and *Laughter Coordination* (Q3), we considered the binary variable *Time* as a predictor, contrasting the time-window preceding the laugh onset/offset (1500ms, *before*) to the time-window following it (1500ms, *after*).

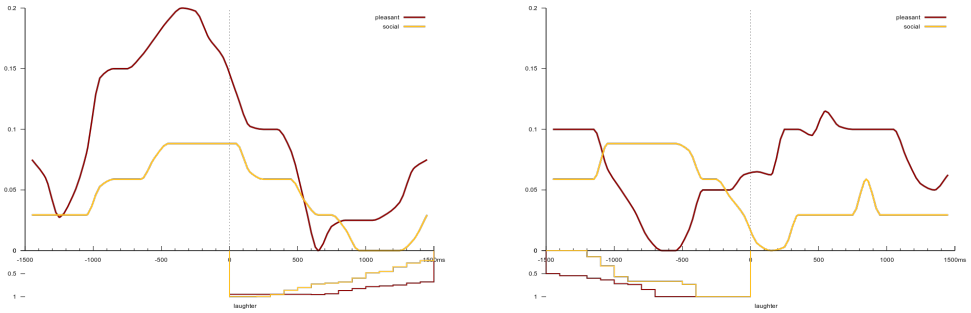
---

2. Including dyads as a random effect did not improve the models.
3. For any bin when gaze shift was occurring (the raw value therefore being between 0 and 1) we rounded up the value to one.




(a) Laugher - Laughter Onset (b) Laugher - Laughter Offset

Figure 3.2: Probability of Laugher's gaze at Partner according to Laughable Type. The probability of laughter duration is shown at the bottom of the figures.




(a) Partner - Laughter Onset (b) Partner - Laughter Offset

Figure 3.3: Probability of Partner's gaze at Laugher according to Laughable Type. The probability of laughter duration is shown at the bottom of the figures.

## 3.4 Results

### 3.4.1 Laugher's gaze × Laughable Type

#### 3.4.1.1 Onset laughter

Figure 3.2a shows the probability of the laugher gazing at her partner around the onset of their own laugh depending on whether the laughter is related to a *pleasant* or a *social incongruity*. We observe a contrasting pattern of *Gaze*, especially *after* the onset of the laughter. The laugher is more likely to be looking at the partner *during/after* a laughter related to a *social incongruity* than to a *pleasant incongruity*.

We observe main effects of *Laughable Type* ($CE = -0.81, SE = 0.22, z = -3.56, p < .001$.) and *Time* in relation to the onset of the laughter ($CE = -0.70, SE = 0.20, z = -3.51, p < .001$), and a significant interaction between the two factors ($CE = 1.60, SE = 0.30, z = 5.31, p < .001$). While the laugher is more likely to gaze at her partner *before* the onset of the laughter when it is related to a *pleasant incongruity* (as in (2)), the opposite is true *after* the onset of the laughter, when they are more likely to gaze at her partner when the laughter is related to a *social incongruity* (as in (3)).

(2) Pleasant incongruity [Pair15 (00:01:25)][4]

**A:** I quite like it 'cause I like it when
there's the little chickpeas on **top|||**

**B:** ((gazes at hummus)) **yeah**

**A:** **'cause it's quite posh** <laughs>

**B:** ((gazes at hummus)) yeah <laughs>

(3) Social incongruity [Pair03 (00:02:59)]

**A:** Shall we say... No.
Ta- **tasty <laughs>||||||**

**B:** <laughs>

**A:** ((returns gaze at hummus))

### 3.4.1.2 Offset laughter

Fig. 3.2b shows the probability of the laugher gazing at the partner around the offset of her own laugh depending on whether the laughter is related to a *pleasant* or a *social incongruity*. We observe a significant main effect of *Time* ($CE = -0.45, SE = 0.09, z = -5.09, p < .001$), but no significant main effect of *Laughable Type* ($CE = 0.21, SE = 0.17, z = 1.18, p = 0.23$) , nor a significant interaction ($CE = -0.12, SE = 0.17, z = -0.72, p = 0.47$). This means that regardless of the *Laughable Type*, the laugher is more likely to look at her partner *before* the offset of her own laughter rather than *after* the offset.

---

4. Speech that overlaps with gaze at partner is shown in bold, with continuation of gaze marked by |||.

### 3.4.2 Partner's gaze × Laughable Type

#### 3.4.2.1 Onset Laughter

Figure 3.3a shows the probability of the partner gazing at the laugher at the onset of the laugh depending on whether the laughter was produced in relation to a social or a pleasant incongruity. We observe a significant main effect of *Time* ($CE = -0.83, SE = 0.2, z = -4.13, p < .001$) and *Laughable type* ($CE = -0.99, SE = 0.22, z = -4.43, p < .001$), and no significant interaction ($CE = 0.31, SE = 0.35, z = 0.87, p < 0.38$); meaning that the partner is more likely to look at the laugher *before* the onset of the laughter, and in general more likely to look at the partner if the laughter was related to a pleasant incongruity.

#### 3.4.2.2 Offset Laughter

Fig. 3.3b shows the probability of the partner gazing at the laugher at the offset of the laugh depending on whether the laughter was produced in relation to a *social* or a *pleasant incongruity*. We observe the opposite pattern to Fig. 3.2a, with the partner more likely to gaze at the laugher if the laugh was related to a *pleasant incongruity* rather than a *social* one.

We observe a significant effect of *Time* ($CE = 0.51, SE = 0.22, z = 2.27, p = 0.02$), while the main effect of *Laughable Type* is not significant ($CE = 0.27, SE = 0.25, z = 1.07, p = 0.28$). Of particular interest is the significant interaction ($CE = -1.43, SE = 0.38, z = -3.71, p < .001$). This shows that *after* the offset of the laughter the partner is much less likely to be looking at the laugher if the laughter was related to a *social incongruity* rather than a *pleasant* one, while the opposite pattern is observed *before* the offset.

### 3.4.3 Laugher's gaze × Laughter coordination

#### 3.4.3.1 Onset Laughter

Fig. 3.4a shows the probability of the partner looking at the laugher at the onset of the laugh depending on whether the laugh produced was an isolated one (chosen as reference level), an antiphonal or a coactive one. We do not observe any significant difference in the probability of the laugher gazing at her partner *before* or *after* the onset of the laugh ($CE = -0.37, SE = 0.21, z = -1.78, p = 0.07$), while we do observe a main effect of *Laughter Coordination*, having isolated laughter as a reference level (Antiphonal-Isolated: $CE = -0.76, SE = 0.31, z = -2.46, p = 0.01$; Coactive-Isolated: $CE = 1.24, SE = 0.25, z = 4.79, p < .001$). No interaction was significant (Time × Antiphonal-Isolated: $CE = -0.24, SE = 0.45, z = -0.54, p = 0.58$; Time × Coactive-Isolated: $CE = 0.57, SE = 0.35, z = 1.63, p = 0.10$).

#### 3.4.3.2 Offset Laughter

Fig. 3.4b shows the probability of the partner looking at the laugher at the onset of the laugh depending on whether the laugh produced was an isolated one (chosen as reference level), an antiphonal or a coactive one. The laugher is more likely to be gazing at her partner *before* rather than *after* the offset of her own laughter regardless of the variable *Laughter Coordination* (*Time*: $CE = 1.62, SE = 0.32, z = 5.04, p < 001$). We observed gaze at the partner to be significantly more likely when the laugher is producing an antiphonal laughter in comparison to an isolated one (Antiphonal-Isolated: $CE = 1.6, SE = 0.39, z = 4.08, p < .001$), and even more likely when producing a coactive laughter (Coactive-Isolated: $CE = 2.49, SE = 0.4, z = 6.19, p < .001$). We observe a significant interactions of *Time* and *Laughter Coordination* (Time × Antiphonal-Isolated $CE = -1.62, SE = 0.46, z = -3.51, p < .001$; Time × Coactive-Isolated $CE = -0.94, SE = 0.47, z = -2.0, p = 0.04$).

### 3.4.4 Partner's gaze × Laughter Coordination

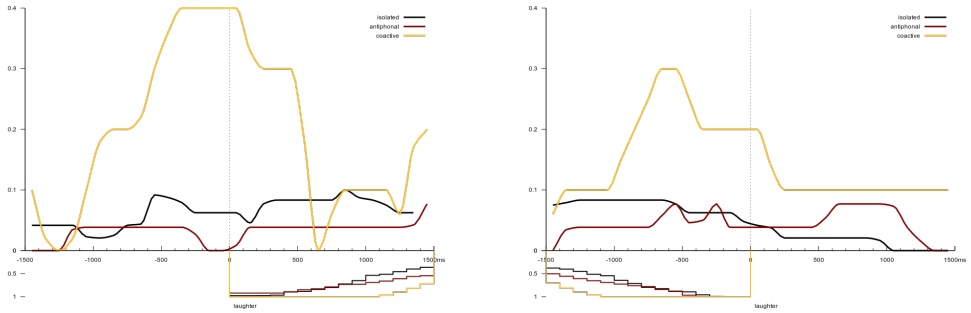
#### 3.4.4.1 Onset Laughter

Fig. 3.5a shows the probability of the partner looking at the laugher at the onset of the laugh depending on whether the laugh produced was an isolated one (chosen as reference level), an antiphonal or a coactive one.

We observed significant main effects of all the predictors included in the model, but no significant interactions: gaze at the laugher is more likely before the onset of the laugh (*Time*: $CE = -0.69, SE = 0.31, z = -2.28, p = 0.02$), significantly more likely at the onset of an antiphonal laughter than an isolated one ($CE = 1.06, SE = 0.26, z = 3.81, p < .001$), and even more likely before the onset of a coactive laughter ($CE = 2.22, SE = 0.29, z = 7.53, p < .001$).

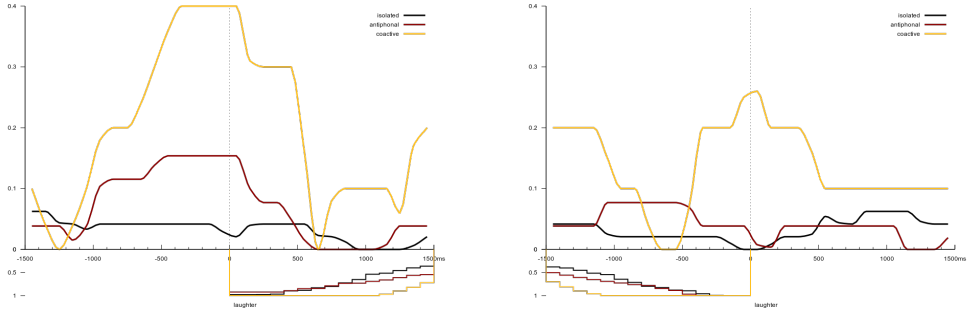
#### 3.4.4.2 Offset Laughter

Figure 3.5b shows the probability of the partner gazing at the laugher at the offset of the laugh depending on whether the laughter was produced in relation to a *social* or a *pleasant incongruity*. We observed a significant main effect of *Time* ( $CE = -0.72, SE = 0.28, z = -2.55, p = .01$), no significant difference between Antiphonal and Isolated laughter ($CE = 0.14, SE = 0.27, z = 0.51, p = 0.6$), but a significant difference between Coactive and Isolated laughter ($CE = 1.19, SE = 0.31, z = 3.81, p < .001$). We also observe a significant interaction between *Time* and *Laughter Coordination* (Time × Antiphonal-Isolated: $CE = 1.59, SE = 0.38, z = 4.14, p < .001$; Time × Coactive-Isolated: $CE = 0.89, SE = 0.43, z = 2.04, p = 0.04$).




(a) Laugher - Laughter Onset (b) Laugher - Laughter Offset

Figure 3.4: Probability of Laugher's gaze at Partner according to Laughter Coordination. The probability of laughter duration is shown at the bottom of the figures.




(a) Partner - Laughter Onset (b) Partner - Laughter Offset

Figure 3.5: Probability of Partner's gaze at Laugher according to Laughter Coordination. The probability of laughter duration is shown at the bottom of the figures.

## 3.5 Laughter and Gaze Pragmatics

Our data show that laughter related to different types of laughables, performing different pragmatic functions in interaction, is characterised by different accompanying gaze patterns both from the laugher and her partner. These observations confirm that both laughter and gaze play a crucial pragmatic role in the unfolding of dialogue; providing further evidence for the stance that to model gaze behaviour one needs to consider not only turns, but also propositional content and dialogue acts performed (Torres et al. 1997). Furthermore our results validate the laughter taxonomy proposed in Mazzocconi et al. (2020) showing that laughs belonging to different classes are produced and perceived as performing different pragmatic functions, eliciting different multimodal behaviours from the interactants. Below we discuss our results concerning laugher's and partner's gaze in relation to the type of laughable, and the results for gaze in relation to laughter coordination between interlocutors.

### 3.5.1 Laugher's gaze

We observe that the laugher is more likely to gaze at the participant *before* the onset of a laughter related to *pleasant incongruity* rather than *social incongruity*, while the opposite is true *after* the onset. This result clashes with our hypothesis 1, according to which we expected an absence of gaze to the partner both *before* and *after* the onset of a laughter related to *social incongruity*. Our hypothesis was based on Stanley and Martin (1968) and Schneier et al. (2011) proposing that social stress makes gaze aversion more likely. Kendrick and Holler (2017) also suggest that gaze aversion is more likely while producing a dispreferred answer, which is a dialogue act that belongs to the *social incongruity* laughable class in Mazzocconi et al. (2020)'s taxonomy.

However, we can explain this data considering that we are looking at gaze during the *laughter* rather than during the *laughable* production (e.g. a dispreferred answer). Most of laughs follows the production of the laughable (e.g. (Tian et al. 2016)). Our data might therefore still be consistent with (Kendrick and Holler 2017). Indeed, we observe a lower probability of gaze at the partner *before* the onset of laughter – a time that often coincides with laughable production. We might therefore imagine a scenario where the expression of a dispreferred answer is not accompanied by gaze at the partner, and only while laughing the laugher looks at the partner, to monitor that the laughter has smoothed the disagreement or the unmet expectation, having the desired positive effect. Our data cannot therefore neither confirm nor disconfirm the social stress hypothesis of gaze aversion. It is possible that the "social stress" component is what influences the laugher to not look at her partner *before* the onset of a laughter related to *social incongruity*, while at the same time being the motivation to check her partner's appraisal of the laughter (produced to ease the situation) during the laughter production.

The lower probability of gaze *after* the onset of laughter related to *pleasant incongruity* mirrors results reported by Gironzetti (2017), who observed a lower probability of attention directed towards smiling facial expressions (including laughter) when it occurred in the context of humorous exchanges in comparison to non-humorous ones. These observations can be interpreted in the light of Becker-Asano and Ishiguro (2009): they observed that when their robot was directing its gaze at the partner while laughing, the laughter was interpreted negatively as being directed *at* the participant, rather than being produced cooperatively. We can therefore speculate that the tendency to avoid looking at the partner while producing a laughter related to *pleasant incongruity* is a way to disambiguate the laugher's intention and social attitude towards the partner.

The opposite pattern observed *before* the laughter onset (i.e. higher probability of laugher gazing at the partner when about to produce a laughter related to *pleasant incongruity*), might be a result of the fact that the laugher is "careful" to assess whether laughter is an appropriate contribution, before producing it. There are indeed judgemental, moral, and cognitive aspects related to laughter production (e.g., not everything can be a subject for laughter, it is silly to laugh at some things, some laughter can be offensive for someone).

The consistency of our results with Gironzetti (2017) is interesting considering the differences between corpora in terms of physical arrangement, data considered, and task. Where participants were opposite each other, conversing freely (without the goal-directed task of our study), and they also considered smiling, suggesting that similar dynamics are at play in the multimodal integration of smiling and laughter pragmatic processing. In our corpus, in contrast, participants are engaged in a task which requires them to pay (and share) attention on objects on the table in front of them, and they are seated at a 45° angle (a setting similar to the referential communication task in (Sandgren et al. 2012)). This means that gaze at the partner and mutual gaze are rarer than in other corpora (engagement in competing activities allows interactants to look away from their partner more frequently (Rossano 2013)), which leads us to speculate that when gaze at the partner instead does occur it is specifically motivated by pragmatic functions.

### 3.5.2 Partner's gaze

Partner's gaze at the laugher is significantly more likely to occur *before* the laughter onset rather than *after*. This is compatible with the idea of gaze being a cue for soliciting a response (Bavelas et al. 2002; Harness Goodwin and Goodwin 1986), and that such a response does not have to be a verbal speech turn (Rossano 2013). The main effect of *laughable type* (i.e. that gaze at the laugher is more likely before the onset of laughter related to *pleasant incongruity*) has to be considered together with the data represented in Fig. 3.5a about gaze and laughter coordination. In Sec.

3.4.4 we report that antiphonal and coactive laughs are significantly more likely to be preceded by gaze from the partner, meaning that gaze can be interpreted as an invitation to join in. The distribution of antiphonal laughs is though skewed towards *pleasant incongruity* (pattern replicated in several corpora e.g. (Mazzocconi et al. 2020)), which therefore constitutes a confounding variable. Due to the small data set we cannot consider such factor in our statistical modelling. We leave this exploration to further work when a larger dataset will be available.

We also observe that *after* the offset of the laughter, partner is less likely to look at the laugher if the laugh was related to *social incongruity*. We interpret this as a "choice" from the partner to avoid direct gaze in order to not put extra pressure on the laugher (who has appraised some situation or dialogue act as potentially discomforting) and maybe choose to give feedback (reassuring the laugher) in another modality, signalling that the issue should be declined as not important (Romaniuk 2009).

The higher probability of gaze at the laugher *after* the offset of laughter related to *pleasant incongruity*, may be explained as a partner's strategy to check whether it would be appropriate to join the laughter. This is consistent with the results reported in Fig. 3.2a: i.e. laughers being more likely to gaze at the partner *before* the production of laughter related to *pleasant incongruity*. Laughing can indeed be a "no laughing matter": not all laughs should be reciprocated (Jefferson 1984).

Our hypothesis 2 is therefore partially invalidated in the setting of our corpus. We do not observe a significant higher probability to look at the laughing partner *after* the onset of the laughter (Fig. 3.3a), but rather *before* the onset. This data can be explained considering that the participants are sitting very close, engaged in a task, and mutual attention is already granted without the need to be signalled through gaze. Our data, on the other hand, highlight the role of laughter to elicit a (laughter) response from the partner (Bavelas et al. 2002).

### 3.5.3 Laughter Coordination

H3 was partly confirmed: we observed higher probability of gaze from the partner at the onset of an antiphonal or coactive laughter, but we did not observe a higher probability of looking at the partner from the laugher preceding the production of an antiphonal laughter. We therefore do not observe the need for a "gaze window" in order to respond to a laughter with a laughter. It would be nonetheless interesting in future work to control whether the laughter was produced by the participants in turn-initial position or not, in order to see if the observations mirror the patterns observed for speech (e.g. (Bavelas et al. 2002)). Nevertheless, our data show the important role played by gaze in the coordination of laughter production between participants and its role for eliciting responses from the partner, not just in terms of speech turns (Rossano 2013). The role of gaze for laughter coordination is particularly striking in the case of coactive laughter (i.e. both interlocutors start laughing at the same onset time), where participants seem to look at each other not only to synchronise on the simultaneous onset but also to terminate the laughter. This kind of gaze may not only serve the purpose of syncing the response, but also monitoring other non-verbal cues about the partner's disposition towards the laughable. This is an open question for future work.

# Chapter 4

# Neural Network Gaze-Target Prediction for Human-Robot Interaction

Gaze cues, which initiate an action or behaviour, are necessary for a responsive and intuitive interaction. Using gaze to signal intentions or request an action during conversation is conventional as discussed in 2.3. Gaze also plays a critical role in HRI tasks such as object recognition and manipulation, as a robot can use gaze to direct its attention to specific objects or areas of interest in its environment. We propose a new approach to estimate gaze using a neural network architecture, while considering the dynamic patterns of real world gaze behaviour in natural interaction.

Humans acquire the ability to decode complicated visual information in infancy (Zohary et al. 2022). Observed gaze is a common deictic ("pointing") signal used by many animals, including members of our own species, to direct action. These reactions are reflexive and pervasive in humans; they occur in a split second, act in a solitary state to task relevance, and seem to be the foundation for the early growth of language and theory of mind. Basic gaze-following behaviours appear to be shared by nonhuman animals and humans (Shepherd 2010b). Current models of vision, which rely instead on heavy supervision, are unable to mimic such learning. A prime example is gaze understanding, a form of early learning that is beneficial for social interaction and collaborative attention.

## 4. Neural Network Gaze-Target Prediction for Human-Robot Interaction71

A crucial social characteristic that facilitates human-robot cooperation is gaze-following (HRI). Robots that can track a person's gaze are better able to comprehend that person's attention, interest, and intentions. Gaze following enables us to make eye contact which can enhance our social presence and naturalness. To infer human visual attention, it is essential to carefully consider head posture, gaze direction, scene organization, and saliency (Parks et al. 2015).

Gaze estimation refers to the process of determining where a person is looking using techniques from computer vision and machine learning (Akinyelu and Blignaut 2022). In the context of dialogue, gaze estimation could be used to infer information about a person's attention, interest, or intentions. For example, if a person is looking at a particular object or person while speaking, it may indicate that they are referring to that specific object or directly addressing that particular individual (Cazzato et al. 2020). Gaze estimation is a challenging task, which can be affected by factors such as head pose, lighting, and occlusions. As a result, the accuracy of gaze estimation models can vary depending on the conditions under which they are used (Kar and Corcoran 2017).

In this chapter, we present a gaze-estimation model that estimates the gaze of two individuals in a scene, detecting different patterns of gaze behaviours in a single model.

Gaze following is a natural and important social behavior that allows people to communicate and coordinate their attention, and it plays a key role in social interaction and communication (Flom et al. 2017). In the context of gaze estimation, gaze following could be implemented using computer vision and machine learning techniques to enable a robot or virtual assistant to follow a person's gaze in a more natural and human-like way Saran et al. (2018). This is widely believed to have potential benefits such as improving the social interaction and communication capabilities of such systems, but remains an open challenge.

In addition to potential applications to Human Robot Interaction (HRI) (Drewes 2010), medical diagnostics (Bedford et al. 2012), virtual/augmented reality Blattgerste et al. 2018, and surveillance systems (Marois et al. 2020), our work is motivated in part by the potential for automated measures of gaze behaviour to contribute to clinical research and understanding of developmental disorders like autism. The burden of manual gaze coding, which is time consuming and costly (Somashekarappa et al. 2020), can be lifted in this situation by automated analysis. This, we believe, offers a useful tool for analysing gaze behaviours in clinical populations at a finer level of granularity, although the accuracy of the labels may be lower than with manual annotation.

Various systems have been created to estimate the direction of gaze. However, it is still difficult to quickly and accurately calculate gaze direction in a variety of environments, particularly when there are two or more parties involved in a conversation. The gaze-following challenge is designed to establish a connection between the scene and human visual attention, in order to evaluate how effectively it will perform during human-robot interaction.

The main objectives and contributions of this chapter are:

1. Automating robot gaze behaviour using machine learning.
2. Classifying elements of the dialogue based only on gaze behaviours (such as dialogue acts, intimacy regulation and referencing objects)
3. Presenting a dataset containing annotations of attention targets with complex patterns of gaze behaviour and out-of-scene target predictions
4. Developing an implementable-model of gaze in dialogue for a conversational robot/avatar to interpret and produce human-like gaze behaviour

Eye gaze supports and augments other social behaviours such as speech/gesture, and the mental states or cognitive effort can substantially influence gaze. Since speech is a dominant mode of communication in human interactions, it is non-viable to separate gaze from speech in human-robot dialogue. Researchers have shown that gaze improves speech-based interactions such as disambiguating object references,

maintaining engagement, conversation and narration, guiding attention, managing partners, and influencing turn-taking (McMahan and Evans 2018; Laube et al. 2011). We suggest a non-wearable eye-gaze detection technique that uses videos to study gaze in interaction by making use of the new developments in deep learning.

The majority of the research on gaze following in HRI points to its potential as a useful tool for enhancing robots capacity for social interaction and communication. However, there are still a lot of difficulties and unanswered problems, such as how to design and apply gaze in a reliable way, as well as how to gauge its efficiency in various HRI scenarios. Our method utilizes the manually annotated gaze predictions denoting various types of gaze to train the network that results in a more efficient and precise detection of gaze targets to its corresponding object in space at a given time. The rapid advancements in the field of robotic technologies presses the importance of social robots' prominence in the future, such as robots that are built for interacting with people and are designed for various applications such as therapy and education alongside industry (further discussed in 6.

## 4.1 Corresponding Work

Recent research has focused on developing algorithms for automatic gaze estimation using various approaches: **Deep Learning-based approaches**: Deep learning algorithms, such as convolutional neural networks have been used to predict gaze locations in images or video data. These approaches have shown promising results in terms of accuracy and efficiency (Lian et al. 2018; Koochaki and Najafizadeh 2018). There are different methods to estimate gaze using machine learning, some of which are based on deep learning algorithms such as convolutional neural networks (CNNs) (Lemley et al. 2019) or recurrent neural networks (RNNs) (Palmero et al. 2018) and also by analyzing head pose/gaze direction using 3D sensors such as depth cameras (Zhang et al. 2020).

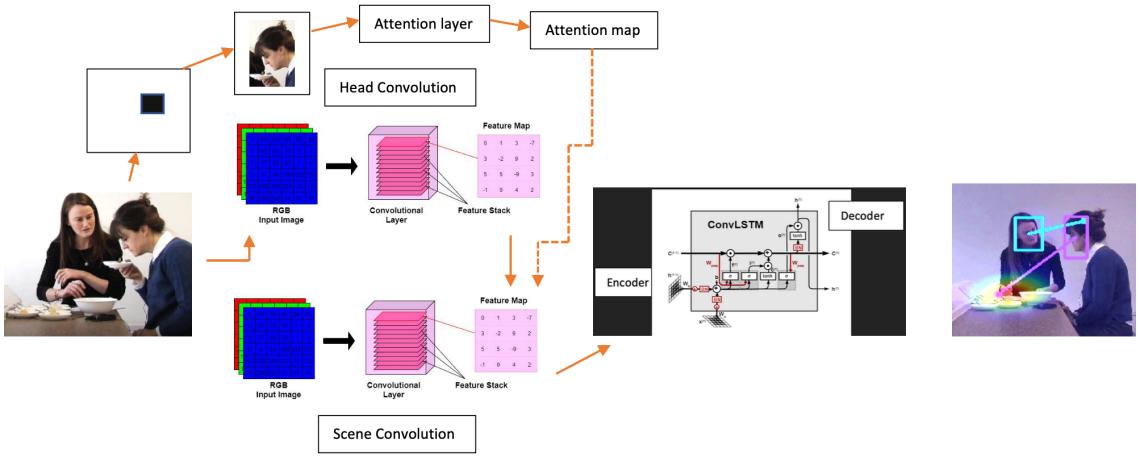



Figure 4.1: Network Architecture

**Appearance-based models**: Focuses on developing models that take into account the appearance of a person's eyes and face to predict gaze locations. These models have been shown to outperform traditional gaze estimation algorithms that rely solely on image data (Wood et al. 2016; Murthy and Biswas 2021).

**Multi-modal fusion**: Emphasizes on fusing multiple sources of information, such as eye movements, head movements, and facial expressions, to make more accurate gaze predictions. The multi-modal fusion approaches have been shown to improve the accuracy of gaze prediction compared to single-modality methods (Ishii et al. 2015; Aftab 2019).

**Gaze correction in virtual reality**: Virtual reality (VR) environments introduce new challenges for gaze estimation, as the participant's gaze is not directly visible. The main focus is on developing gaze correction methods that can correct for misalignment between the gaze location and the virtual scene in VR environments (Shi et al. 2020; Sidenmark and Gellersen 2019).

The basic pipeline of gaze estimation using CNNs involves:

1. Head pose detection: The first step is to detect the head in the image or video. This can be done using a variety of methods, such as Haar cascades, HOG+SVM, or deep learning-based approaches.

2. Eye detection: Once the head pose is estimated, eye region is segmented and is further cropped from the image or video and resized to a fixed size.
3. Feature Extraction: The CNN model is then used to extract features from the eye images. This typically involves passing the eye images through several layers of convolutional and pooling layers.
4. Gaze Regression: After feature extraction, the model uses a fully connected layer to estimate the gaze direction, in some cases, the model can use 3D gaze direction.
5. Gaze vector: The output of the model is a gaze vector that indicates the gaze direction.

Automatic gaze annotation is typically faster and less expensive than manual gaze annotation, but the accuracy of the labels may be lower than with manual annotation.

#### 4.1.0.1 Single-stage regression

It is commonly used in computer vision applications such as object detection, semantic segmentation, and gaze estimation. This approach has the advantage of being computationally efficient, as it requires a single forward pass through the network to make a prediction. Additionally, single-stage regression models can be trained end-to-end, allowing the network to learn complex relationships between the input features and the target variable. The network typically consists of several layers of neurons, where each layer performs a specific function such as feature extraction, non-linear transformation, or prediction (Sun et al. 2020).

However, single-stage regression models may be less interpretable than multi-stage approaches, as they can be more difficult to understand the internal workings of the model. Additionally, single-stage regression models may be more susceptible to overfitting to the training data if the model is too complex. In these cases, regularization techniques, such as dropout or early stopping, may be used to prevent overfitting.

#### 4.1.0.2 Auxiliary face landmark detection

It refers to the use of additional information related to facial landmarks, such as the positions of the eyes, nose, and mouth, in the process of gaze estimation. Facial landmarks provide important information regarding the structure of a face and can be used for face normalization, making it easier to detect the gaze direction. For example, the position of the eyes can be used to align the face so that the gaze direction is always in a consistent direction, regardless of the position of the head (Kendrick et al. 2018).

In gaze estimation, the auxiliary face landmark can be used to improve the accuracy of the gaze direction prediction by providing additional context about the face. For example, the position of the eyes can be used to estimate the position of the gaze direction relative to the face. Additionally, the position of the mouth can be used to distinguish between a smiling face and a neutral face, as smiling can affect the gaze direction. These techniques can be used in combination with a single-stage regression model for gaze estimation or as a pre-processing step before feeding the data into the gaze estimation model. The use of auxiliary face landmark information can be useful in improving the accuracy and robustness of gaze estimation, especially in scenarios where the gaze direction is occluded or the face is in an unusual position.

#### 4.1.0.3 Gaze following

In an initial attempt to solve the gaze following problem, Recasens et al. (2015) propose the GazeFollow dataset, which consists of 2D images, location of faces, and labels for the gaze following phenomenon. A video gaze-following dataset called Video-AttentionTarget, described in Chong et al. (2020), contains images, face locations, and labels from videos. The gaze target predicted by the above mentioned papers use a straightforward network topology that incorporates saliency and gaze

Figure 4.2: Various types of gaze: a) Joint attention (black box), gaze on hummus and questionnaire, b) Gaze aversion (green box), gaze on partner, c) Mutual attention (yellow box), looking at each other d) Individual gaze (orange box), attention on different objects in space

channels. To increase the accuracy of gaze-following, additional information such as human stance, gaze direction, or object identification are included. Gaze-following becomes more challenging as a result of the acquisition of the extra information, which requires additional dataset for training or extra processing steps for inference.

### 4.1.1 Predicting the Target

A significant difference between earlier studies on gaze target prediction is whether the attention target is situated in a 2D image or 3D space. Cheng et al. (2021) were among the first to show how a deep model can learn to find the gaze target in the image; our work focuses on the 2D case. By simultaneously learning gaze angle and saliency it is possible to address out-of-frame gaze targets. It can be done by taking into account various scales, body pose, and sight lines and within-frame gaze targets which improves the estimation even further.

We explicitly model the temporal gaze behavior and report results for gaze target prediction in video while taking out-of-frame targets into consideration. As per the architecture, head features are used to regulate the spatial pooling of the scene, and head location map in place of a one-hot position vector to produce a fine-grained heatmap using deconvolutions rather than a grid output.

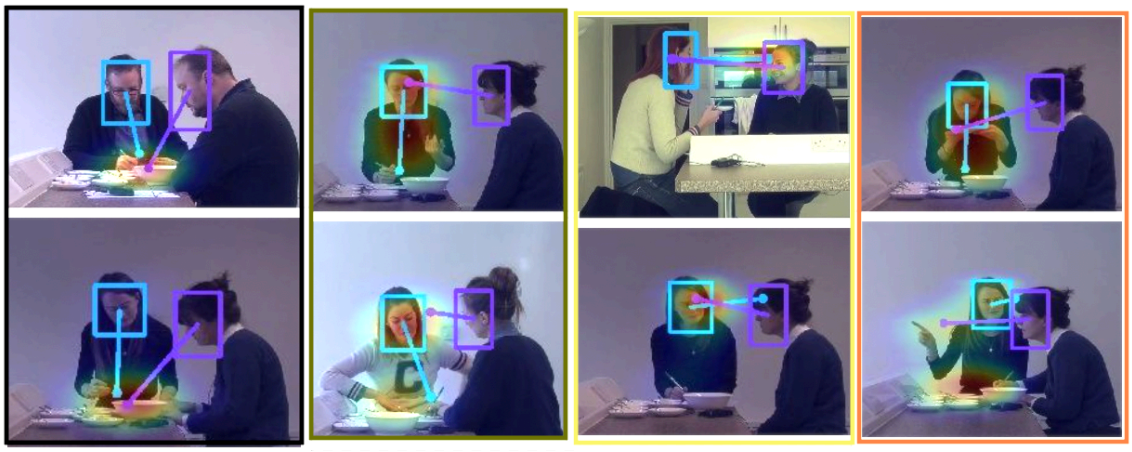


Figure 4.3: Simultaneous Gaze Generation with attention maps of both parties in a single scene

## 4.2 Method

### 4.2.1 Multiface processing pipeline

The technique uses a two-stage strategy. First, by considering the scene and head data as a separate network input that share the same scene properties. The scene image serves as a discrete input to the scene-channel for a one-shot feature extraction regardless of the presence of the individuals.

We implement active learning or semi-supervised learning approaches (Taha 2023) that make use of both manual and automatic annotation to improve the performance of gaze estimation models. The idea is to use the automatic annotation to pre-label a large dataset and then use a smaller set of manual annotations to correct and improve the automatic annotation.

The feature extraction backbone network is Resnet50 (Koonce and Koonce 2021) for network verification, where the head channel considers the number of persons present in an input image into account (dyadic in this particular situation). To forecast the gaze target, the two participants' head images and head locations are used as binary masks. Each head image location mask serves as the head channel

input.

$$f_h = C_h(I_h), f_s = C_s(I_s) \tag{4.1}$$

Second, in order to predict the gaze, same head ($f_h$) and scene ($f_s$) features are concatenated to a fusion layer which goes through several up sampling and convolution layers to predict the position heatmap of the gaze targets. To avoid exorbitant consumption of computational resources for scene and face feature extraction we opted for lightweight ghostnet module which uses inexpensive operations to generate feature maps similar to convolutional layers, although it does not transcend the convolution operation.

### 4.2.2 Face detection and heatmap generation

Single-stage methods for multi-stage face recognition are preferred for real-time applications due to their light weight and high accuracy. For example, face recognition methods apply a single-layer architecture to design more efficient modules for facial features.

We created an extensive face dataset using a large number of manual annotations and use a one-step gaze estimation method. The input to the model is a complete image with two faces and the output is the gaze directions of the people in the scene. Instead of processing each face individually, we propose a model that estimates the gaze of multiple people simultaneously with assistance from attention heatmaps within the image.

Using an image, we can infer a person's gaze direction and consider whether there are any salient objects along the estimated line of sight. The heatmap is learned by connecting two independent convolution pathways in accordance with this hypothesis. A multi-task learning framework is used to explicitly train for gaze angle

with a convolutional pathway connected to the face image. Auxiliary tasks for gaze angle estimation offer several advantages, including the possibility of devising a supervisory signal based on the relationship between gaze heatmaps and angles, which effectively enhances heatmap estimation performance (figure 4.3).

Heatmap generation for gaze predictions using machine learning involves creating visual representations of gaze data. Heatmaps are generated by plotting the gaze data onto a 2D image and color coding the points based on the density of gaze data in that area (Park et al. 2018). Gaze heatmaps can be used to evaluate the performance of gaze-based algorithms by comparing the generated heatmaps with ground-truth data.

We adopt a cross domain learning approach where the model learns partial information relevant to each task from different datasets (for example, looking outside the frame, looking at each other, fixation on an in-frame object, looking away in the frame, etc.).

### 4.2.3 Neural Network Architecture

The videos that were manually annotated were later used to predict robust gaze target location.

#### 4.2.3.1 Convolutional layers

The architecture consist of head convolutional layers for head feature extraction (ResNet-50). It is followed by an additional residual layer and an average pooling layer to reduce the spatial dimension of the feature maps. The black pixels shown in figure 2, denote the head bounding box of each person and the white pixels denote rest of the image. They are reduced using three max pooling operations. The

scene feature extraction network computes scene feature map similar to the head convolution module. Head feature map and head position are then concatenated while the scene convolution is the concatenation of the head position and the scene image.

#### 4.2.3.2 Dense layers

The final layers of the architecture consist of a fully connected layer where the attention layer computes attention maps by passing the two concatenated layers. Lastly, two convolution layers encode the features in the encoder module.

$$y = W_d * h + b_d \tag{4.2}$$

where $y$ is the gaze prediction, $W_d$ is the weight matrix of the final dense layer, $h$ is the output of the previous layer, and $b_d$ is the bias term of the final dense layer.

$$h_i = f(W_i * x + b_i) \tag{4.3}$$

where $h_i$ is the output of the $i^{\text{th}}$ convolutional layer, $f$ is the activation function, $W_i$ is the weight matrix of the $i^{\text{th}}$ layer, $x$ is the input image, and $b_i$ is the bias term of the $i^{\text{th}}$ layer.

#### 4.2.3.3 Recurrent layers

The scene information providing head position as spacial referencing enables the model to learn faster. Subsequently, the architecture includes convolutional Long Short-Term Memory (Conv-LSTM) to capture temporal dependencies in the eye movement data from the sequence of frames. Four deconvolution layers makeup the deconv module to up-sample the features computed by the convLSTM into a full sized feature map.

$$h_t = f_r(W_r * h_{t-1} + U_r * x_t + b_r) \tag{4.4}$$

where $h_t$ is the hidden state at time step $t$, $f_r$ is the activation function of the recurrent layer, $W_r$ and $U_r$ are the weight matrices of the recurrent layer, $x_t$ is the input image at time step $t$, and $b_r$ is the bias term of the recurrent layer.

#### 4.2.3.4 Object detection

The feature map is then modulated by a scalar that defines whether the gaze attention of the person is within the bounds of the scene or out-of frame, higher the value of the scalar the focus is within the frame. It consists of two convolution layers and a fully connected layer while the element-wise subtraction from the feature map normalization is performed. Following, heatmap with minimum values greater than or equal to zero are cropped resulting in the final heatmap that can be visualized with intensity maps of the object prediction.

### 4.2.4 Simultaneous Multiple Gaze Detection

The model was trained on NVIDIA RTX TITAN GPU and implemented in Pytorch. The network is optimized by the Adam algorithm for 100 epochs and the batch size was set as 32, and the initial learning rate was $10^{-}4$. The scene and the masked header image, acts as input to the model, scaled to 224*224. The output was a heat map of size 64*64 and the ground truth heat map was generated with 2D Gaussian weights around the ground truth of the tracked target. During training, to ensure a repeatable balance of the two channels, a single randomized head image was selected from the scene during each iteration.

In addition, weight loss varied depending on the duration of the training. In the early stages, we expect the output heatmaps to reflect the tracked target points, like previous gaze tracking methods. In the final stage, we focused on refining heat maps and minimizing errors through regression. Thus, in our implementation, the

value of (alpha) increases as the number of training epochs increases from 0 to 0.5. During inference, to track multiple people's gazes, the scene-channel extracts features from the scene only once, while the main channel runs multiple times for different people.

To capture the natural gaze behavior, VideosAttentionTarget dataset [1] videos include live interviews, sitcoms, reality shows, and movie clips, all available on YouTube containing 50 different shows represented by the clips ranging in length, from 1-80 seconds. From the extracted head image, we use the head conditioning branch to compute a head feature map. An additional residual layer and an average pooling layer are added to the head convolution of the network. Using three successive max pooling operations, a binary image of the head position is reduced and flattened. Previous work has used position encoding to encode the position of the head in the scene, but the binary image encoded it more effectively. Concatenating the face with its head positioning gives the head feature map. These two features are then passed through a fully-connected attention layer to create an attention map.

Following, Wang et al. (2020), we employ a single-stage regression task that maps image pixels to multiple gaze directions directly, estimates gaze directions and further predicts auxiliary face information including bounding box and facial landmarks. The scene convolution component of the network, which is equivalent to the head convolution module previously discussed, is used to compute a scene feature map. The model quickly learns when the head position is provided as a spatial reference in addition to the scene. The attention map generated by the head conditioning branch is then multiplied by the scene feature map. As a result, the model can learn to focus more on scene elements that are more likely to be noticed based on the characteristics of the head. Along with the weighted scene feature map, the head feature map is then concatenated. In the encoder module, the combined features are encoded using

1. https://github.com/ejcgt/attention-target-detection

two convolutional layers after which the model uses a convolutional Long Short-Term Memory network (ConvLSTM). The deconvolutional network, made up of four deconvolution layers, up-samples the features calculated by the convolutional LSTM to a full-sized feature map.

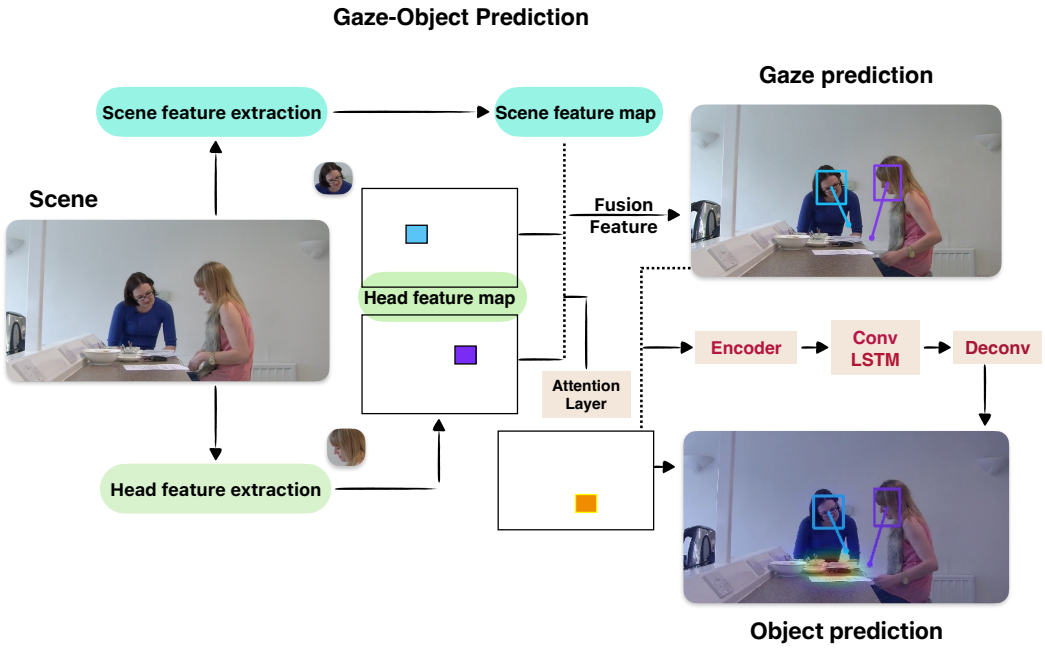



Figure 4.4: In head feature extraction, location mask of the head image performs multiple feature extraction (in this case two people) and the scene image acts an an independent output for a one time feature extraction. Following, head features are concatenated with the shared scene features. The fusion feature predicts the the final gaze output of the respective person. The object detection maps the corresponding gaze target point generating heatmaps

### 4.2.5 Heatmap generation for the intended object

Using the Model, we estimate the subject-dependent saliency of a scene in terms of a heatmap (the “what” component of visual attention) and how likely it is that the subject will fixate on the estimated gaze target in the scene (the overall “strength” of visual attention).

A fully-convolutional model has two pathways, one for the whole image and another for the face image (figure 4.1). In order to separate the scene information from the face information, we optimize the relationship between the scene and the face information in the input. Consequently, retrieved scene information can be shared when gaze following processes multiple individuals in the same scene.

Data inaccuracy in imaging systems due to amplitude quantization can be perceived as a random noise, that effects the accuracy of decisions based on the image data. Hence, to tackle the inconsistent size causing the quantization error, the numerical coordinate regression method is introduced on the heatmap to create a straightforward structure that allows multiple people to use the same scene feature for gaze-following prediction. In the following, we compute the gaze target point and remove the quantization error.

The likelihood of fixation is a single-valued measure that indicates how likely it is that the subject is looking at the estimated target region. The model is based on a fully connected layer. The model can estimate a person's visual attention much more accurately using this last output. In case the person is looking outside the image, the heatmap is close to zero. The value can be added to the heatmap with an operator, which, depending on the application, may be a weighting operator or a gating operator. This last layer is trained to create a higher value when it is more likely that the heatmap region is attended to and a lower value otherwise.

### 4.2.6 Implementation and Generation of Gaze

The GHI corpus contains annotation for different types of gaze in a dyadic conversation and is time stamped. It also has speech transcribed for each participant in two separate tiers. Each of these videos last about 20-30 minutes and have a high resolution recording from two vantage points (discussed in detail in 2).

The complete image, the subject's cropped face, and the location of the subject's face that requires attention estimation act as the input to the model. The two stills from the video are enlarged to 224x224 so that the network can see the face with improved resolution. The full image coordinates for the face position are supplied and these coordinates are flattened to a 168-dimensional 1-hot vector after being quantized onto a 13x13 grid.

Two convolutional (conv) pathways make up the model: a scene pathway and a face pathway (see figure 4.1). The conv pathways employ ResNet 50 (He et al. 2016) as their backbone network and for each of the conv paths, all conv layers of ResNet50 are specifically used. Later three convolutional layers (1x1, 3x3, and 1x1) with ReLu and batch normalization with stride 1 and no padding after each ResNet50 block are added. The filter depths of the convolutional layers are 512, 128, and 1, respectively which results in extracted features' dimensionality being reduced (Jin et al. 2022).

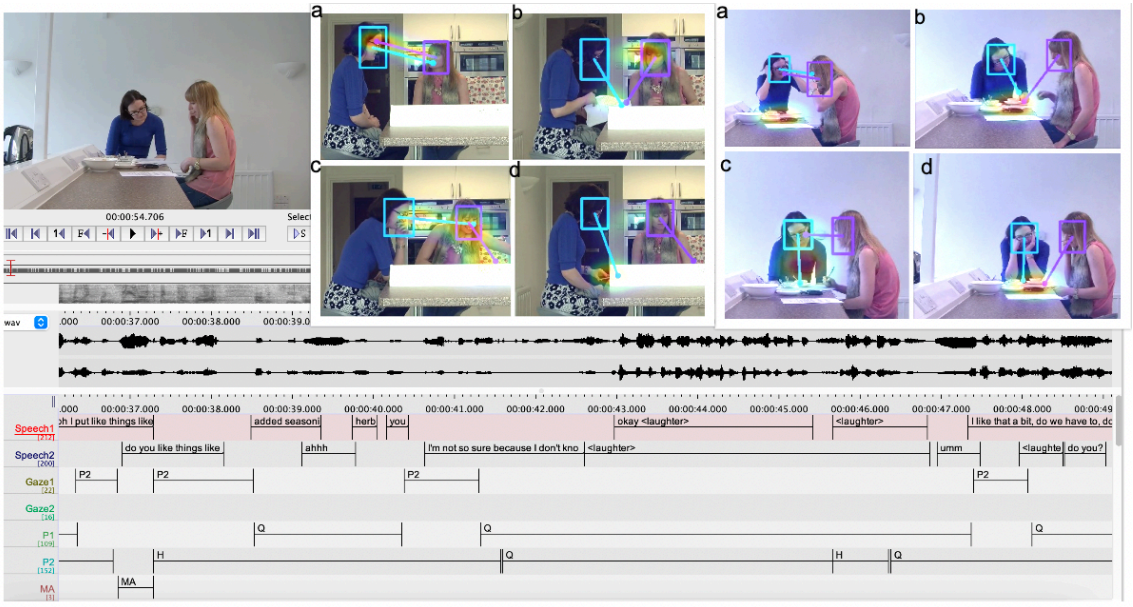



Figure 4.5: Various gaze estimates captured in two different angles in a particular scene. a. Mutual attention (looking at each other), b. Joint attention (gaze on hummus and questionnaire), c. Gaze aversion (gaze on partner while the partner looks away), d. Individual gaze (attention on different objects in space)

In the face pathway, the feature vector computed with the face input image goes through a fully connected layer to predict the gaze angle represented using yaw and pitch intrinsic Euler angles. In the scene pathway, the feature vectors extracted from the whole image as well as from the face image are concatenated with the face position input vector to learn the person-centric heatmap. Similarly to face position, the ground truth used for learning the heatmap is available as a gaze target position in (x,y) coordinates which is quantized into 10 grids in each dimension.

In order to determine the ”intensity” of the fixation that is, how likely it is that the individual is genuinely fixating at a gaze target within the viewable scene, the input vectors to the last layer of each pathway are concatenated and fed into the last fully connected layer. When the subject is gazing inside the image, the training label for this value is 1, and when they are looking outside the scene, it is 0.

In total for the 24 videos, 48 individualistic gaze predictions were generated accounting for close to 80k predictions per video. The heatmaps give us a clear understanding of the region of interest on which the gaze attention occurs (see figure 4.3).

## 4.3 Evaluation

The manual annotation has been done on 4 videos for the GHI corpus and we automate gaze for all 24 videos ranging between 24-30 minutes, and the generated images for each video is between 40k to 60k. Therefore, the resulting gaze prediction dataset (GHI-Gaze) approximately consists of 2.4 million images with facial landmark, gaze information and heatmaps were generated with two different angles for the same session (figure 1). The various types of gaze behaviour can be extracted by collecting the specific coded temporal information.

We conducted two types of evaluation to measure the perform of the model. Firstly, we use the spatiotemoral model on the existing dataset such as VideoAttentionTarget and Gaze-follow. Secondly, on the manually annotated Goodhousekeeping Institute (GHI) dataset containing 60k images and 90k annotations for specific types

of gaze (Somashekarappa et al. 2021). GazeFollow dataset is publicly available and is built on ImageNet, PASCAL and MSCOCO containing 122k images of different scenes and 130k annotations while the VideoAttentionTarget dataset contains youtube videos, with 164k extracted frames of labelled gaze behaviour and the targets.

### 4.3.1 Spatiotemporal Evaluation

In the context of gaze estimation, it involves evaluating the accuracy of the model's predictions in terms of both the location of the gaze target and the timing of the gaze. We evaluated the single image target prediction and the dataset contained annotations of the persons head and gaze locations in various types of images. The model was trained by using annotation labels from the baseline models mentioned in Table 4.2. Later these labels were extended to detect out-of frame gaze targets while the GHI dataset was not yet introduced for unbiased comparison.

We use four performance measures that are key indicators based on previous gaze-following methods to evaluate the model. **AUC** (Area Under the Curve) is the metric used to evaluate the performance of a binary classifier i.e the similarity between the predicted versus the ground truth heatmap. AUC ranges from 0 to 1, with a value of 1 indicating perfect classifier performance, and 0.5 indicating random performance. The averaged difference between the coordinates of the predicted gaze target point and the coordinates of the ground truth point that some annotators have assigned a label is the average distance, **Avg Dist**. The shortest distance between the anticipated point and the closest labeled point is the Minimum Distance, **Min Dist**. The angular error between the predicted and ground truth gaze direction from the head position to the attention target in the image is referred to as **Ang**.

Table 4.1: Evaluating on GazeFollow dataset

| Method | AUC ↑ | Avg. Dist ↓ | Min. Dist ↓ | Ang ↓ |
|---|---|---|---|---|
| Recasens et al.(2015) | 0.878 | 0.190 | 0.113 | 24.0° |
| Lian et al.(2018) | 0.906 | 0.145 | 0.082 | 17.6° |
| Chong et al.(2020) | 0.921 | 0.137 | 0.077 | - |
| Dai et al.(2021) | 0.922 | 0.133 | - | 16.1° |
| Jin et al.(2021) | 0.919 | 0.126 | 0.076 | - |
| Tu et al.(2022) | 0.917 | 0.133 | 0.069 | - |
| GHI-Gaze (ours) | 0.920 | 0.112 | 0.059 | 13.7° |
| Human | 0.924 | 0.096 | 0.040 | 11.0° |

Similar to the recent gaze prediction approaches for feature extraction we adopted resnet50 network. The resulting comparison with previous methods is shown in Table 1. The analysis shows that the GHI-Gaze, predictions fine tuned with human annotations and attention maps perform better with AUC of 0.924 and the angular error of 13.7° compared to the previous results. The human metrics have the best performance measure with 0.924 AUC and 11° of angular error.

In Figure 1, "*a*" represents mutual gaze where the individuals are looking at one another. Due to a clear view of the face it is easier to visualize the gaze in the first image. While in the second generated image of "*a*" due to face occlusion it is impossible for human annotators to recognize the gaze, but the model accurately detects the gaze taking into consideration other factors such as head pose estimates and attention heatmaps.

Images represented as "*b*" denote gaze on the same object in the scene, and "*c*" averted gaze where an individual looks at the partner and the partner looks away. Finally, "*d*" represents gaze on different objects with in the scene.

### 4.3.2 Evaluation on GHI Corpus

We evaluate our complete model on the VideoAttentionTarget (VAT) dataset and use AUC, Distance and Out-of-Frame measures. The analysis shows that the GHI gaze predictions fine tuned with human annotations and attention maps perform better with AUC of 0.889 to reported baselines although doesn't exceed human evaluations. Distance refers to the measurement of the separation between two points, objects, or entities. In the context of gaze estimation, it could refer to the distance between the gaze target and the eye gaze position. Out-of-Frame refers to an object or entity that is not within the bounds of an image or video frame. In the context of gaze estimation, it means that the gaze target is not visible in the image.

Table 4.2: Quantitative Spatiotemporal Evalutaion

| Method | AUC ↑ | Distance↓ | AP↑ |
|---|---|---|---|
| Random | 0.505 | 0.458 | 0.621 |
| Fixed bias | 0.728 | 0.326 | 0.264 |
| Chong+LSTM | 0.833 | 0.171 | 0.712 |
| VAT | 0.860 | 0.134 | 0.853 |
| GHI | 0.889 | 0.117 | 0.869 |
| Human | 0.921 | 0.051 | 0.925 |

The first column in Table 4.2 lists various methods that were considered as baseline performance measures. The prediction made at a 50% chance is denoted as 'Random' and the bias term or constant that is added to a model's output to adjust its predictions is represented as 'fixed bias', where its value is not updated during training. It can be added to the output of a layer to improve the accuracy of predictions. The fixed bias is usually initialized with a small random value and remains unchanged during training. It helps to shift the predictions of the network towards the target values and can be used to account for any systematic error in the data.

We report the gaze target non temporal estimator with additional LSTM layer proposed by Chong et al. (2018) with an AUC of 0.833 for comparison. Our model was evaluated on the VAT dataset that previously only reported for individual gaze estimation in a scene and was not fine tuned with human feedback resulting in an AUC of 0.860. Human evaluation measures the manual annotation of gaze predictions of objects within the screen and has the highest AUC of 0.92. GHI performs better compared to other models with an AUC of 0.889 and excludes out-of scene targets but predicts that the gaze targets for both the participants in a single frame providing information about the specific types of gaze.

## 4.4 Modeling gaze behaviour in a Robot

Modeling gaze behavior in a robot often requires a combination of computer vision, and robotic control system, as well as a thorough understanding of human gaze behavior. It can be challenging, particularly when the gaze behavior is complex or subtle.

Some of the most difficult types of gaze to model include: **Mutual gaze/ Direct gaze**, where the robot directly looks at the human, simulating eye contact, because it requires the robot to respond in real-time to the human gaze, while also adjusting its own gaze in a natural and engaging way (Lombardi et al. 2022). Longer mutual attention can be considered eerie and induce uncanny valley effects. **Averted gaze**, where the robot looks away from the human, requires the robot to simulate a lack of attention or interest, which can be difficult to do in a way that is convincing to humans (Müller et al. 2020). **Gaze cues**, where the robot uses gaze to initiate an action or behavior, often requires the robot to respond to human gaze in a sophisticated and context-sensitive way (Mutlu et al. 2012; Massé et al. 2016). **Scanning**

**gaze**, where the robot moves its gaze around its environment, simulating exploration or attention to multiple objects or people. **Follow gaze**, where the robot tracks the movement of a human's gaze, implying attention or interest in what the human is looking at.

In a human-robot interaction, the prediction of a gaze target can be used as input for the determination of the best robot action in response to a human action, given the circumstances. The challenges arise from the need to simulate human-like gaze behavior in a way that is realistic, engaging, and responsive to human actions. Follow gaze, where the robot tracks the movement of a human gaze, can be easier to model compared to other types of gaze behavior in a robot. This is because gaze follow often requires less sophisticated modeling of human gaze and more straightforward control of the robot's gaze mechanism. These different types of gaze can be used in combination to create more sophisticated and nuanced interactions between the robot and the human.

In gaze follow, the robot uses computer vision algorithms to detect the position and movement of the human gaze, and then adjusts its own gaze accordingly. It does not need to respond in real-time or make sophisticated judgments about the context of the interaction. Instead, the robot simply follows the human gaze as it moves. However, it is important to note that gaze follow is still a challenging task, particularly when the gaze is fast-moving or unpredictable and goal/context dependent. The results of the study have a direct application on improving the contextualized gaze on a social robot and in this paper, we discuss the gaze architecture for implementation on a robot.

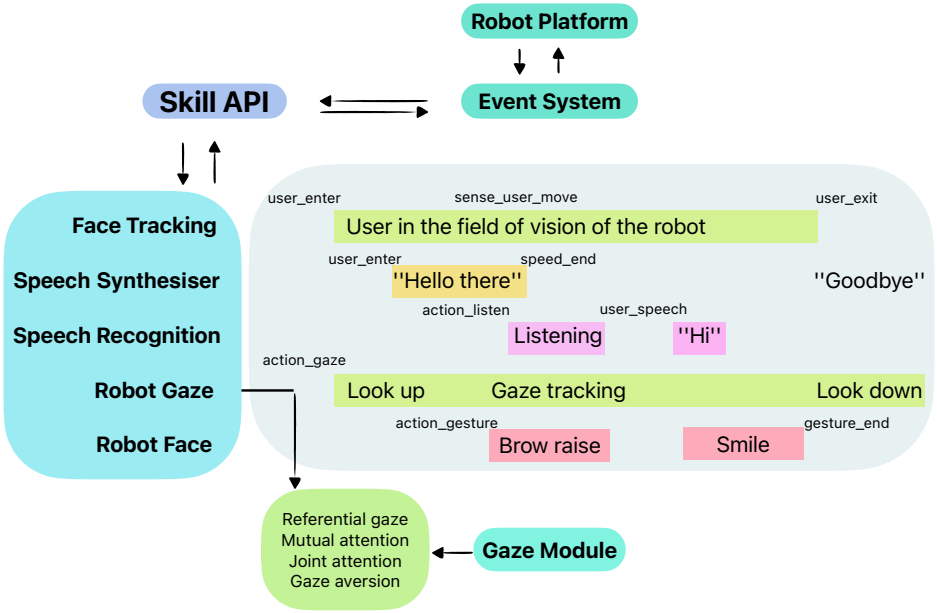



Figure 4.6: Real-time Human-Robot Interaction architecture for a Social Robot

## 4.5 Gaze Interaction Architecture for a Robot

It is anticipated that social robots, which are becoming more prevalent in society, would be able to communicate like humans do by using both verbal and non-verbal cues (Mishra and Skantze 2022). Anthropomorphic robots are favored to zoomorphic or machinelike robots because they are more likely to trigger mind attribution. Gaze predictability manipulation can help study gaze following strategies based on the task characteristic and following reflexive gaze. By adding invalid cues it is possible to determine the trustworthiness of the system (Morillo-Mendez et al. 2022).

The Furhat Robot platform consists of input and output interfaces (projector, neck servo motors, touchscreen, etc.) and software modules for automatic speech recognition (ASR), text-to-speech synthesis (TTS), face tracking, etc. The Event System mediates all of the sensory inputs, modules, and actuators in the Robot Platform. To create a gaze plan, the Gaze Planner advocates high-level events such as the user's

position, speech input, and the positioning of objects on the touchscreen. The Gaze Controller then takes advantage of this strategy to generate actions that causes the robot's head to turn and eyes to move. Using the Skill API, which defines all of the interaction-specific details, is where the interactions can be implemented.

The interaction is modelled using state charts where the dialog contexts are defined as hierarchical states and the generic behaviors are defined on higher levels in the hierarchy (figure 3). While the more specific behaviors are defined further down in the hierarchy, and may override generic behaviors. The intent classification (NLU) dynamically takes the current hierarchical contexts enabling multiple intent in each utterance. Complex behaviors may be defined in their own state charts, and reused across applications. Computer vision platform tracks real time multi-user face and estimates head pose from the video stream. The architecture provided in figure 3, describes the dialogue with gaze module implementation from the current work. A specific type of gaze, acts as an input for the robotic gaze based on the speech intent and face tracking by assessing the temporal predictions.

# Chapter 5

# Experimental Evaluation of Gaze Interaction with Social Robot

Gaze automation in a social robot is a crucial aspect of Human-Robot Interaction (HRI) that aims to make the robot more engaging, intuitive, and effective in its communication with humans. The ability of a robot to control its gaze —-where it looks, how it looks, and when it looks- is essential for establishing a natural and meaningful interaction with users (Ruhland et al. 2015; Breazeal 2004; Onyeulo and Gandhi 2020). The main purpose of the research in this chapter is to assess different patterns of gaze in robots in order to contribute to a more effective, natural, and engaging interaction. **How does human-like behavior of a robot influence people's perception of it?**

Maintaining appropriate eye contact helps the robot establish a connection with users, fostering a sense of engagement, trust, and rapport by directing attention towards users or relevant objects, signaling engagement and facilitating more focused interactions. Gaze behavior serves as a powerful nonverbal communication tool, allowing robots to convey intentions, emotions, and social cues (Aliasghari et al. 2020; Liu et al. 2012), and using these gaze cues to signal turn-taking in conversations, facilitating more natural and intuitive interactions. Gaze automation enables robots to provide visual feedback, such as nodding or looking towards objects, to

confirm understanding or indicate agreement and enables better task performance by focusing attention on critical elements, reducing distractions, and improving task efficiency (Somashekarappa et al. 2023).

The main objectives and contributions of the chapter are:

1. Implementing different gaze patterns in a social robot in an interactive setting
2. Experimental evaluation of robot gaze patterns in human-robot interaction
3. Analysis of how different gaze patterns may be predictive of people's engagement in an interaction with a social robot

In this experimental study, we investigate gaze patterns as a potential continuous metric to gauge people's perceptions of robots. Our specific research questions are i) Is there any correlation between robot gaze patterns participants' perception of its friendliness, cooperativeness and sociability? ii) Does the robot's gaze pattern significantly influence users assessments of the robot's usability and overall interaction experience? iii) If so, how do specific gaze patterns indicate the emotional engagement of the user during the interaction?

## 5.1 Corresponding Work

Research has shown that a robot consistently maintaining the user's mutual attention is viewed as more genuine (Hoffman and Breazeal 2004). Similarly, leveraging the speaker's visual attention through gaze-tracking positively impacts understanding speech, aiding in the prediction, clarification, and resolution of spoken references (Prasov 2011; Degutyte and Astell 2021). Mutlu et al., delved into the significance of mutual attention within collaborative scenarios, where both humans and robots engage in coordinated tasks within a shared environment (Mutlu et al. 2016). Additionally, they explored gaze strategies employed to structure dialogs and define distinct roles of speakers and listeners in human-robot interactions.

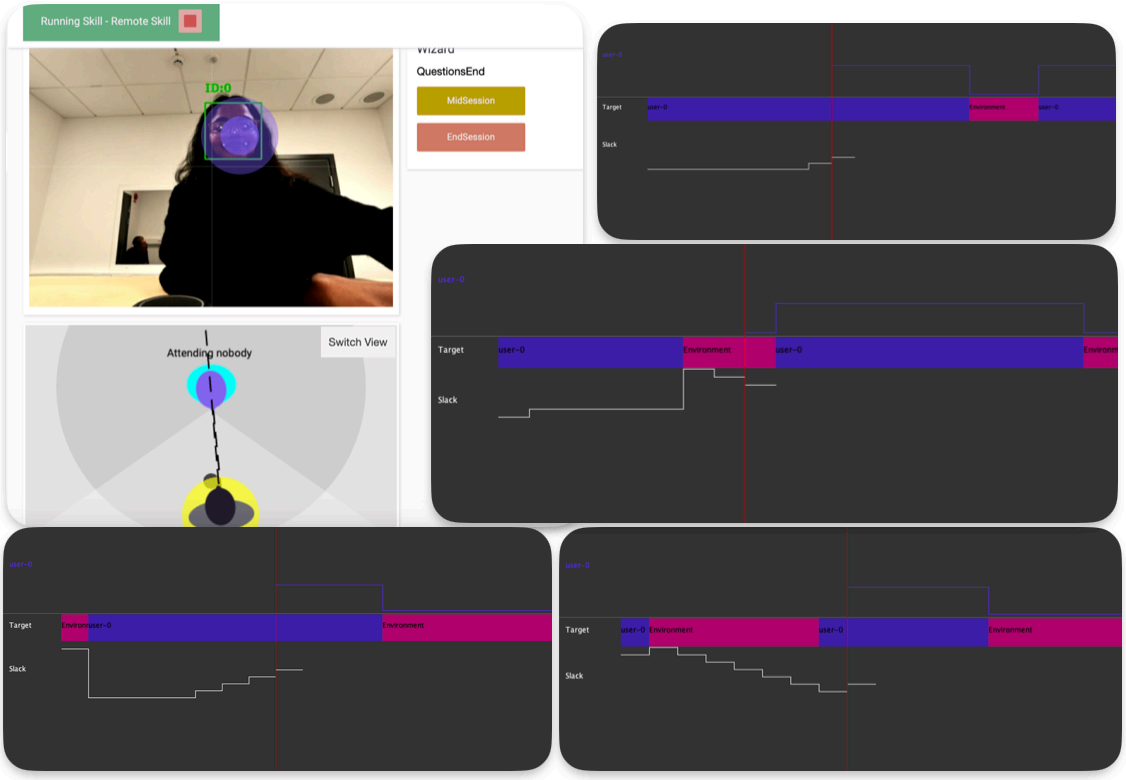



Figure 5.1: Event Attention Interface (EAI) showing the gaze model state transitions of the robot in the interaction space. Blue block indicates gaze on the user and pink block in the surrounding environment. The blue horizontal line depicts the user's gaze on the robot and the white horizontal line is the beginning and the end of speech turn of the robot.

The computational frameworks developed for recognizing and facilitating engagement, focus on "the method by which multiple participants establish, sustain, and conclude their perceived bond during shared activities" (Rich et al. 2010). This research highlights the influence of targeted and mutual gaze on the relational dynamics between interacting parties but doesn't address the role of gaze in clarifying speech or using gaze to prompt responses.

While certain investigations have focused on general modelling approaches to integrate various modes and dialog reasoning for multi-modal interfaces, embodied conversational agents, and human-robot interactions, they often overlook key concepts such as joint attention, mutual attention, or engagement (Wang et al. 2023). The current paper builds upon insights from existing literature but advances the field by

incorporating different gaze aspects into a unified and innovative modelling framework. The model for our experimental condition is based on detailed observations of gaze behaviors in human-human interactions, made from the manual annotation and analysis on the Good Housekeeping Institute (GHI) corpus in Chapter 2 (Somashekarappa et al. 2021).

Humans employ diverse communication modalities to convey information, while selecting channels based on their efficiency and communicative potential (Hessels 2020). To disambiguate, individuals rely on their counterparts' ability to integrate information from these channels for clarity. Solely relying on verbal expressions can introduce uncertainties, potentially disrupting mutual understanding (Cassell et al. 2000). Consequently, listeners often integrate a speaker's verbal content with accompanying gestures and eye movements to derive a clearer interpretation before seeking clarification (Gullberg and Holmqvist 1999). Notably, when using gaze during speech, the visual cue typically precedes speech by approximately 800-1000 milliseconds, while individuals tend to focus on the referenced object about 200 milliseconds after auditory input (Heyselaar et al. 2021).

In collaborative efforts, individuals strive to direct their partners' attention either to specific objects in their environment or to themselves. In addition to verbal and pointing cues, they commonly use gaze and a combination of these communication methods to achieve this goal. Furthermore, individuals track their partner's gaze to establish a shared frame of reference, leading to synchronized focus and mutual attention on a particular object. This interaction signifies active engagement in the joint activity and the ability to discern indicated items. When both parties maintain this focused interaction, they engage in mutual gaze (Mitterer and Reinisch 2017). Both gaze mechanisms, aimed at fostering joint attention, play a crucial role in preserving shared understanding. In our robotic application, the system draws attention to the speaker based on the demands of the conversational act such as maintaining attention while the speaker is looking at the robot or looking away during a long utterance to avoid discomfort. Effective communication relies on skillful mechanisms

that regulate the roles of speakers and listeners (Brown 1991). In this context, the orientation of gaze becomes a pivotal cue in either facilitating or constraining these role transitions. Generally, speakers divert their gaze from the audience, signaling their desire to maintain control of the conversation, whereas they redirect their focus toward another participant to relinquish the floor (Mason 2012). If a statement does not conclude with a gaze directed towards another individual, the transition between speakers may be prolonged.

Figure 5.2: Session in progress. Experiment setting showing cameras placed behind the users and Furhat during the interaction. Images to the left show various positioning of gaze movements of the robot in accordance with the conversation flow (consent has been provided my the participants to use the images from the experiment for publishing purposes with faces being swapped/anonymized)

Despite achieving good results in replicating human-like gaze behaviors in robots, a prevalent constraint is their predominantly reactive nature. Although certain systems devise gaze behavior plans for forthcoming utterances at the onset of speech (e.g., (Ishii et al. 2016)), these plans lack incremental updates and do not significantly influence the ongoing gaze behavior. Another common limitation is the static nature of many systems, utilizing fixed duration for gaze shifts. For instance, in (Das et al. 2015), the robot's gaze remained fixed on the relevant target for 1-5 seconds during interactions before transitioning to the target with the lowest priority. In contrast, Human-Human Interaction (HHI) entails extensive planning. Studies indicate that gaze behavior is intricately coordinated with the underlying speech plan (Holler and Kendrick 2015b). The duration of planning determines whether a swift glance suffices or necessitates head movement for a more extended gaze.

In summary, prior research has explored gaze models for robots, but it has often overlooked the consideration of the human partner's eye movements. The predominant gaze behavior of existing robots is designed in response to users' speech, or gaze behavior is accomplished through head movements. However, head movements are limited to approximating coarse gaze direction and cannot effectively convey nuanced eye movements. Given the identified gaps in existing research, the present study addresses this gap by directing attention toward the intricate interplay between human and robot gaze dynamics. Our focus is to discern how incorporating a more comprehensive understanding of human eye movements into the design of robotic gaze behaviors could enrich the overall human-robot interaction experience. This research is aimed to contribute valuable insights into refining robotic gaze models, moving beyond traditional constraints, and fostering a more natural and intuitive communication between humans and robots.

## 5.2 Human to Robot interaction setup

This section describes the implementation of our system that enables real-time gaze interaction with a social robot head. Furhat[1], is an anthropomorphic robot equipped with a *Software Development Kit (SDK)*, which provides tools tailored conception, deployment, and analysis of applications. It features a biomimetic neck design facilitating lifelike head movements, comprehensive control over facial expressions, gestures, and ambient lighting. The platform supports customization, enabling adjustments to facial characteristics, ethnicity, gender, multilingual modality, and even species, with adaptable faces securely attached through magnetic mechanisms.

The primary components of Furhat's programming infrastructure encompasses development of skills using *Kotlin API*, with integration into python for object detection. The skill framework constitutes an advanced layer building upon rudimentary I/O capabilities, allowing *Natural Language Understanding* (NLU), dialog management, multimodal utterances, interaction logging, and the incorporation of *Graphical User Interfaces (GUI)*. The facial behavior is controlled by a 3D face model which is similar to that of virtual agents.

The robot can execute rapid gaze shifts through digital animation and incorporates physical servos in its neck to mimic head movements. Convincing neck and eye gaze behavior are crucial for our specific task, where users are expected to assess the robot's visual behavior. The robot achieves precision in looking at various parts of the lab by standing in the fixed position, and is calibrated to ensure accurate gaze shifts towards specific locations.

---

1. https://furhatrobotics.com/

### 5.2.1 Interaction session

The experiment is a within-subject design where the participants interacted with the robot in 3 consecutive sessions. The social robot displayed three variations of gaze behavior: neutral, experimental and random (see section 5.3.2 for details). The participants had zero exposure to the robot prior to the experiment. The social interaction sessions lasted about 30-40 minutes. Each session was programmed to take up to 10 minutes where the robot asked questions and the user was requested to briefly answer at their own pace while the robot maintained expressive gaze movements throughout the experiment. Finally, at the end of each session two questionnaires were provided to measure user engagement and perception of the robot.

### 5.2.2 Participants

21 participants between the age of 25 to 48 with the average age of 36.5 years were recruited (M=13; F=7; Non-Binary=1). They were either first or second language English speakers, with a minimum of undergraduate education. Prior to the main session we conducted a pilot study on 5 individuals. At the beginning of the session, participants were presented with information about the study and provided their informed consent. The study was approved by the Swedish Ethical Review Authority[2].

## 5.3 Experiment and evaluation

The participants were seated approximately 150 cm away from the robot. The user and Furhat, were centrally aligned. We adjusted the participants' sitting height to guarantee that their eyes are approximately the same level as the robot's eyes. Our baseline condition is using continuous eye contact as similar to gaze behavior always

2. `https://etikprovningsmyndigheten.se/`, 2023-03044-01.

used in existing robotic systems: the robot attends to users' face when they are facing towards the robot. The gaze shift is generated via eye movement. The robot's neck moved to track the user during the experiment along with their gaze. Furhat's blink was consistent across all experiments.

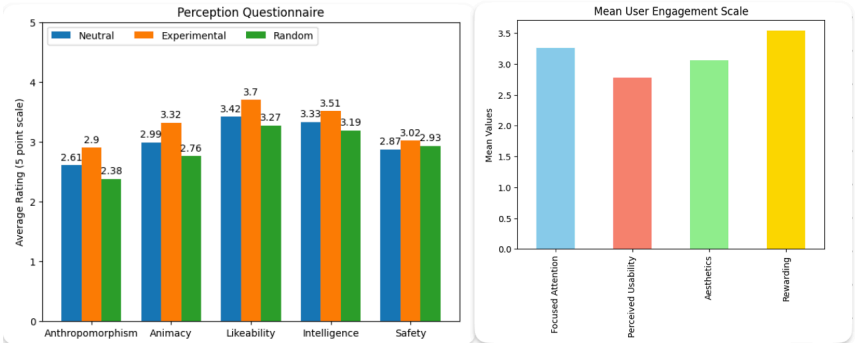



Figure 5.3: Questionnaire Analysis. Plot 1: Perception Questionnaire, Plot 2: User Engagement Scale

Prior to the study, the experimenter explained the purpose and procedure of the study to the participants. The entire experiment lasted for around 30 minutes per participant. During each session, the participants had three interactions with the robot (one in each of the gaze behavior conditions; neutral, experimental and random, with the order randomized between participants). In each interaction, the participants were asked to answer six unique questions. After each interaction, each participant filled in a questionnaire to assess their perceptions of the robot's behavior. The Perception Questionnaire consists of questions to assess users' perceptions of the robot's anthropomorphism, animacy, likeability, intelligence and safety, with each of these factors based on 3 to 5 sub-questions rated on a five point Likert scale (e.g. the subscales related to anthropomorphism included items such as rating the robot on a scale that ranged from 'fake' (1) to 'natural' (5) and 'artificial' (1) to 'lifelike' (5).

After all three interactions, the subject filled in a user engagement scale questionnaire about their overall experience and feedback. This questionnaire consists of 12 statements to be assessed in a 5 point Likert scale from "strongly disagree" (1) to "strongly agree" (5), with each statement relating to one of four factors. The four

factors are 'Focused Attention (FA)', 'Perceived Usability (PU)', 'Aesthetic Appeal (AE)', and 'Reward Factor (RF)' (O'Brien et al. 2018). Examples of items relating to perceived usability for example, are 'I felt frustrated while talking to Furhat.' and 'I found Furhat confusing to use.´

Furhat, first greeted the participants and asked them open ended unique questions during each session which were consistent across subjects. There were a total of 18 questions and the robot answered these questions briefly as well while the interactant was asked to observe the naturalistic behavior of the robot.

### 5.3.1 Measures

Users' subjective understanding of each interaction was assessed using the responses to the perception questionnaires. We further assessed users overall experience by analysing the responses to the user engagement scale. For both of these scales we use only average figures aggregated over the sub-items for each factor.

Figure 5.3 (Plot-1) represents the analysis of a Perception Questionnaire comparing ratings across different attributes ('Anthropomorphism', 'Animacy', 'Likeability', 'Intelligence', 'Safety') under the three experimental conditions: 'Neutral', 'Experimental', and 'Random'.

### 5.3.2 Event Attention Interface

In the *neutral* condition the system tracks the face with the users movement and blinks at regular intervals similar to the other conditions but does not react to the user's gaze. Instead the robot's attention was directed towards the user's face position. A wait key was initialized in all conditions in case of delayed response from the user. In case of no reply, the robot repeated the question and if the user did

not want to share or answer during the conversation, by default Furhat moved on the next question after a bridging sentence. This paradigm was to make sure the conversation flow did not affect the perception of the robot since the main focus in to access the non-verbal cues.

In the *experimental* condition, the gaze patterns were dynamically programmed so as to follow the verbal cues with semantic and pragmatic information, in line with evidence from gaze research in human-human interaction. The completion of the syntactic unit by the user was a cue to break mutual gaze, while the end of sentence completion of the robot was a cue to move gaze back to the listener which are cues for turn-yielding and turn-holding (Skantze 2021). Gaze aversions are used to reduce cognitive effort and modulate intimacy, hence the robot looked away during the speaking turn (Jaber 2023). During a longer speaking turn, the robot looked away in the environment for one second before (randomly initialized) to avoid eerie mutual attention (figure 5.1, block 4) (Somashekarappa et al. 2021). In order to avoid overlap, if the participants began speaking before or after turn completion, Furhat paused until the user finished their turn.

In contrast, the *random* condition was designed to randomly trigger the gaze of the robot without any predetermined contextual padding. In this case we implemented the same basic gaze movements as in the experimental condition, but the initiation of the gaze behaviors was not directly tied to the interactive context. For example, gaze could be directed towards the environment (and away from the speaker) during a users speech, towards the speaker for an indefinite length of time or the robot could appear to look away in the middle of a sentence.

In order to understand and visualize the dynamic gaze events in random and experimental conditions we utilize the *Event Attention Interface (EAI)*. Figure 5.1 shows the platform of the robots interface, indicating the attention on the user to the left. The graphical representation depicts the generated gaze behavior of Furhat during

interactions and also the status of users gaze focus on or away from the robot. During the transitions, the robot maintained mutual attention towards the user while they glanced away frequently. Similarly, the robot actively broke the eye contact when the user looked steadily at the robot.

## 5.4 Questionnaire Analysis

Participants' feedback on the overall interaction experience and the impression of the robot is presented in figure 5.3.

*Attributes Comparison*: The graph in figure 5.3 plot 1 shows the average ratings for each attribute across the three experimental conditions.

As can be seen in figure 5.3, the *experimental* condition, based on human gaze behavior, consistently leads to higher ratings across all attributes, suggesting that the experimental natural gaze manipulations positively impact user perceptions. The *random* condition generally falls below the *experimental* condition but shows comparable ratings to the *neutral* condition. This might indicate that random gaze factors have a less pronounced impact on perceived attributes.

We ran a series of Generalized Linear Mixed Models (GLMMs) with each of the five attributes as dependent variable, gaze pattern type (neutral, experimental or random) as independent variable, participant ID as a within-subject factor and age and gender as random effects. Post hoc pairwise comparisons were carried out in the case of significant effects to identify which differences were significant.[3]

---

3. All statistical analyses were run using SPSS 28. The models use a linear model with a normal distribution.

For *anthropomorphism*, there was a significant main effect of gaze pattern type ($F_{2,60} = 5.681, p = 0.006$). Post hoc pairwise analyses showed that the rating for anthropomorphism was significantly higher in the experimental than random condition ($t = 3.362_{60}, p = 0.001$) and marginally higher than the neutral condition ($t = 1.895_{60}, p = 0.063$). Neutral and random conditions were not significantly different from eah other ($t = 1.46_{60}, p = 0.148$).

For *animacy*, there was a significant main effect of gaze pattern type ($F_{2,60} = 15.666, p < 0.001$). Pairwise tests showed that animacy was rated significantly higher in the experimental condition than the neutral condition ($t = 3.283_{60}, p = 0.002$) and the random condition ($t = 5.568_{60}, p < 0.001$). Furthermore, the neutral gaze pattern was rated as significantly more animated than the random gaze pattern ($t = 2.284_{60}, p = 0.026$).

There was also a significant main effect of gaze pattern type on *likeability* ($F_{2,60} = 8.183, p < 0.001$). Likeability was rated significantly higher in the experimental condition than both random ($t = 3.990_{60}, p < 0.001$) and neutral ($t = 2.572_{60}, p = 0.013$). Likeability was not rated significantly differently in the random and neutral conditions ($t = 1.419_{60}, p = 0.161$).

The same pattern of effects was found for the ratings of *intelligence*, with a significant main effect of gaze pattern type ($F_{2,60} = 8.183, p < 0.001$), a significant difference between the experimental and random ($t = 3.883_{60}, p < 0.001$) and experimental and neutral ($t = 2.142_{60}, p = 0.036$) conditions and no significant difference between random and neutral gaze patterns ($t = 1.691_{60}, p = 0.096$).

There was no significant main effect of gaze pattern on participants' perceptions of *safety* ($F_{2,60} = 1.485, p = 0.235$).

With the exception of *safety*, therefore, user's perceptions of the robot were higher for all our measured attributes in the experimental condition than in the other two conditions, highlighting the importance of gaze behavior in users perceptions of a social robot in interaction.

*User Engagement Scale*: Utilizing a structured User Engagement Scale (fig 5.3, plot 2), four key aspects were examined: 'Focused Attention,' 'Perceived Usability,' 'Aesthetics,' and 'Rewarding.' Each aspect was assessed based on user-provided ratings, and the mean values were calculated to discern the overall perception of users. Each of these contain three sub-statements. Participants expressed a moderate level of engagement in terms of focused attention (Mean Value = 3.19). Perceived usability and aesthetics yielded a moderate mean value of approximately 2.84 & 3.0. This indicates a generally satisfactory level of usability, hence requires potential enhancements in the user experience. The aspect of 'Rewarding' exhibited a relatively higher mean value of approximately 3.5 where participants found the interaction to be rewarding, indicating a positive and fulfilling experience. Note that as we only asked our subjects these questions once after all three of their interactions, further between-subject experiments are needed to assess whether these engagement factors are also positively impacted by more human-like gaze behavior in the robot.

# Chapter 6

# Discussion

The thesis firstly, explores the dynamics of eye gaze signals in human-human interactions, delving into their role in reducing cognitive effort, regulating attention, and signaling social intentions. Through extensive annotation and analysis, it uncovers patterns of gaze behavior indicative of agreement, disagreement, and decision-making processes. Notably, mutual gaze emerges as a powerful conversational reinforcement, contributing to natural perception and interaction fluency.

Secondly, the research introduces a novel neural network architecture for gaze estimation, offering a more accurate and efficient approach compared to previous methods. Challenges such as head pose variations and occlusions are addressed, paving the way for real-time gaze tracking applications in human-robot interaction contexts. This advancement holds promise for improving the naturalness and intuitiveness of interactions with social robots, enhancing user engagement and task performance.

Finally, offers valuable insights into the nuanced role of gaze behavior in human interaction, laying the foundation for the design of more sophisticated and socially adept conversational agents. Future work may include expanding the annotated corpus, evaluating inter-rater reliability, and integrating gaze cues into robotic systems for various applications, such as therapy and education. Additionally, ongoing efforts to address ethical and legal concerns will be crucial in ensuring the responsible and ethical use of gaze technology in human-robot interaction contexts.

## 6.1 Implications of Gaze in Human Interaction Dynamics

Eye gaze signals play a role in reducing cognitive effort and balancing attention with intimacy, impacting turn-taking dynamics and indicating cognitive effort during gaze aversions. Gaze-following without speech differs from static stimuli observation, and gaze transitions in dialogue acts are influenced by speech stimuli, as reported by Admoni and Scassellati (2012).

The study discussed in Chapter 2 presents findings that suggest potential avenues for further research in gaze analysis. The identified themes include the prediction of gaze agreement/disagreement before the emergence of linguistic cues and the exploration of how decision-making influences eye gaze behavior. The study reveals that mutual gaze is established by a speaker before making remarks, but when denying or expressing a differing opinion, the gaze remains in the shared environment. This observation is consistent across various scenarios related to agreement/disagreement. The partner establishes mutual gaze before making remarks. But while denying or not sharing the same opinion as the partner, the gaze is in the shared environment as noted in *Excerpt 1*, Similar pattern was observed across scenarios pertaining to agreement/disagreement scenarios. This hypothesis of existence of a unique correlation between gaze patterns, is seconded by (Grynszpan and Nadel 2015).

Mutual attention occurrences are found to be the least common but still act as powerful conversational reinforcements. In human-robot interactions, mutual gazes contribute to natural perception, higher recall, and improved persuasiveness. Gaze duration is influenced by individual personalities, with extroverts spending more time looking at their partner.

Gaze cueing is found to influence joint attention, and in verbal interaction, joint attention can occur simultaneously without the influence of verbal utterances, suggesting the development of social cognition (Beaudoin and Beauchamp 2020). The duration of gaze during joint attention is notably spent in coordination tasks, with short fixations potentially contributing to more natural interactions. Considering the task required coordination, the maximum gaze duration during the session was spent in Joint attention, but interestingly, on average, each of these fixations lasted for about 2.9 seconds. The gaze-shifts from one object to another is quick and is not directly influenced by other factors such as speech. From a conversational robotics perspective, short duration fixations in robotic gaze could in turn make interaction more natural.

Eye gaze signals reduce cognitive effort and balance attention with intimacy when the speaker wants to maintain or relinquish the floor and the gaze aversions we observe (lack of overlap between G1 and G2) signal cognitive effort that is looking away or toward the speaker while beginning to answer a question depending on whether they were in agreement or not, which in turn, suggests that these gaze aversions are influenced by the purpose of the direction shift (Andrist et al. 2014).

Gaze-following without speech relies on following the motion based on observation of static stimuli, but it acts differently during dialogue acts (Shepherd 2010a). As noted in *Excerpt 4*, gaze transition occurs a few seconds before the actual implication of movement because of the added assistance from the speech stimuli. As reported by Admoni and Scassellati (2012) people can process gaze information, such as direction, really quickly, as shown in overlapping gaze transition (see qualitative analysis). Interactive eye gaze improves fluency and smooths task performance and subjective experience.

Establishing mutual attention can assist in reducing pragmatic overload in perceiving cues (Zhang et al. 2017). However, in our study mutual attention gaze occurrences are the least common, even though they still act as powerful conversational reinforcements. Mutlu et al. (2012), showed the human-like gaze behaviour of story

telling implemented in a humanoid robots created more natural perception where the people had higher and more effective recall when numerous mutual gazes were established with the listeners. Also, promotion of persuasiveness of the robot while telling a story with the addition of gaze shifts. But vise versa is true when only gestures were added without any gazing effects. During human robot handover interactions using eye gaze helped the robot to improve fluency and subjective experiences.

Peoples' personalities can also affect gaze duration in a conversation. People are more likely to speak when their conversation partner looks at them more often (Vertegaal and Ding 2002). The interpersonal dynamics between partners i.e the familiarity is correlated to the amount of mutual gaze not only on each partner's individual gaze behaviour. From an adaptive evolutionary perspective, attending where others attend can provide information about behaviourally relevant events in the surroundings, particularly action plans, intentions and successive action plans, through the means of joint attention (Shuai 2012).

Processing eye gaze information results from an interaction between face specific structures (involved in visual analysis) and an extended system (spatial attention) as proposed by investigations into the distributed human neural system for face perception (Haxby et al. 2000). Psychophysical interactions (PPIs) in functional magnetic resonance imaging study showed differential connectivity or correlation with core face perception structures in the posterior superior temporal sulcus and fusiform gyrus. This was noted while viewing gaze shifts relative to control eye moments such as the opening or closing of the eyes. It demonstrated the contribution of both the dorsal and the ventral core face areas to gaze perception. Hence, this network provides an interactive system focusing on spatial attention and corresponding shifts in attention (Nummenmaa et al. 2009)

A fearful facial expression with the pointing/directing of the eyes can signify the presence of danger in the surrounding which was investigated in an fMRI study (Hadjikhani et al. 2008) where the meaning of the facial expression along with the direction of the gaze was proven to compute the behavioural implications from the observer's perspective.

Discussing the aforementioned neural correlates of gaze behaviour and its direct influence on dialogue has helped expand our understanding of some linguistic phenomena in persons with disabilities. One of the important goals of the the thesis is developing an automated system with improved gaze behaviour recognition. It can also be used to assist in the diagnosis and rehabilitation of persons with communication impairment, developmental delay, cerebral palsy, quadriplegia, autism, Angelman syndrome, schizophrenia, and aphasia, to name a few.

Patients with schizophrenia displayed reduced non-verbal behaviour and increased negative symptoms (Lavelle et al. 2012) and poor coordination of turn-taking along with disfluencies using fewer self-repairs in dialogue (Howes et al. 2017). Another study presented an analysis of gaze aversion patterns distinguishing between positive and negative schizophrenia (Vail et al. 2017). Children with ASD prefer more limited social interaction compared to children without ASD, hence measurement of eye gaze as a screening tool may be an important contribution in this area (Vargas-Cuentas et al. 2017). As a result, the development of innovative assistive technologies can alleviate current challenges and improve diagnostic accuracy.

## 6.2 Implications for Multimodal Meaning Representation

Chapter 3 provides evidence for the important link that gaze behaviour has for co-ordination in interaction, and also stresses the interaction between gaze, non-verbal behaviour and dialogue acts. Our data therefore offers new material for modelling multimodal meaning construction in interaction – important not only from a theoretical linguistic perspective, but also for the implementation of ECAs able to be more pragmatically adequate and to read non-verbal cues from the user in order to refine their own behaviour (e.g. if the user laughed and looked at the ECA, a likely adequate response might be to laugh back). More complex models for semantic processing are needed in order to tune ECAs behaviour to the pragmatic functions performed by the laughter, but our results suggest that gaze might be one of the cues to be considered in order to classify the type of laughter. Chapter 3 is limited by the small sample size analysed. Extending our dataset, will allow us to employ more complex statistical models to be able to account for several variables at the same time (e.g. laughter position in relation to the speech-turn, to the laughable placement, arousal). Cross-cultural studies (e.g. Rossano et al. (2009)) showed differences in gaze behaviours between different communities (consideration relevant also for the implementation of ECAs appropriate to the user's culture). Our results therefore should not be taken as absolute, but open up the possibility for interesting comparative studies.

## 6.3 Implications of Deep Learning for Gaze Estimation

The main goal of Chapter 4 is to improve the accuracy of gaze estimation and prediction. In this chapter we propose a novel neural network architecture to simultaneously and accurately detect gaze target on the intended object for multiple people in a single scene. We compare the results firstly with manually annotated data from GHI corpus and then with the popular GazeFollow dataset. The results show an improvement in the performance compared to previous methods and provide specific information of the type of gaze in a given scene.

We faced challenges such as head pose variations, occlusions, and cluttered backgrounds, but with the help of extensive manual annotation data made available it has been possible to reduce error while also adopting open-sourced pre-trained models. Most current gaze prediction methods use visual information only (Spiller et al. 2021). However, incorporating linguistic modalities such as dialogue could lead to improved performance and more natural gaze predictions (Hakkani-Tür et al. 2014).

It is possible to identify when someone is inattentive by observing how they look at an object, following their gaze, and even identifying if they are maintaining mutual gaze. Yet, there remains a complex, open challenge in automatically detecting and quantifying these types of visual attention from images and videos.

Gaze information can be used as an input for human-Robot interaction (figure 4.2), and focus on developing more sophisticated gaze-based interaction methods that are more natural and intuitive. Gaze can provide insights into human behavior, such as attention, memory, and emotions. Hence, by developing new methods for analyzing gaze data, it is possible to gain a better understanding of human behavior and its underlying cognitive processes. The final chapter discusses the implementation of the gaze results obtained from the study on a Furhat Robot.

Many current gaze estimation and prediction methods are computationally intensive and not suitable for real-time applications. Developing fast and efficient algorithms for gaze estimation and prediction is an important area for future work.

Overall, the field of gaze estimation and prediction has the potential to revolutionize the way we interact with technology and gain insights into human behavior. There is much work to be done to reach these goals, but the potential impact is significant.

Most research on gaze estimation has been constrained to specific predetermined contexts, though the problem has long been a hot research topic (Pathirana et al. 2022). Our method offers a detailed level representation of attention for each individual in a video, in contrast to approaches that explicitly infer gaze behavior.

## 6.4 Impact of Anthropomorphic Social Robots on User Perception

Chapter 5, highlights the importance of human-like gaze behavior in positively influencing user perceptions, especially in the context of an anthropomorphic social robot. Tying the gaze behavior to the context of the interaction in progress, by looking away and returning the gaze to the interlocutor in line with how humans tend to do this in relation to turn-taking interaction and not continually gazing at one's interlocutor as is often the case in social robots, positively impacts on participants' perceptions of anthropomorphism, animacy, likeability and intelligence, but has no impact on participants' perceptions of safety. Perhaps surprisingly there was no difference in participants' perceptions between the neutral and random conditions, except in the animacy case where the neutral gaze behavior (which does not vary as much) was rated as *more* animated than the random gaze behavior. This suggests that care must be taken when implementing particular behaviors in social robots as bad algorithms might actually be worse than doing nothing.

One factor that was not yet addressed in this research is the question of how people's perceptions of robots evolve over repeated interactions, which requires the creation of measurement methods suitable for extended evaluations. Currently, assessing people's views of robots heavily relies on questionnaires and interviews, which come with inherent limitations (Fink et al. 2011; Flensborg Damholdt et al. 2020). Firstly, these tools only reflect an individual's viewpoint at a particular instance, making it challenging to link shifts in perception to specific interaction moments. Secondly, for an accurate longitudinal assessment, multiple evaluations are necessary. Yet, repeatedly completing questionnaires disrupts the natural interaction with the robot, potentially reducing engagement and task performance. Lastly, relying on self-reported measures introduces biases; individuals might recall previous responses, leading to response fatigue or inadvertently revealing experimental objectives. In future work it is necessary to address these questions by analysing the videos of the interaction to try to discover if there are behavioral cues from the users (e.g. smiling, verbal and non-verbal feedback) which are correlated with their reported perceptions of the interactions.

The study could also benefit from qualitative data to complement quantitative ratings, providing deeper insights into users' subjective experiences. Therefore, to effectively analyze how perceptions of robots change over time and connect these changes to specific robot actions, there's a need for more subtle and continuous assessment methods.

Developers can leverage the insights gained from this analysis to prioritize elements such as gaze that enhance anthropomorphism and animacy in similar systems. In conclusion, the analysis provides a nuanced understanding of how different experimental conditions impact user perceptions across multiple attributes, offering valuable insights for both researchers and practitioners in the field.

## 6.5 Ethical Implications

The ethics of gaze estimation involve considering the potential consequences and impacts of using the technology on individuals, groups, and society. One indispensable ethical issue is privacy, since it can be used to track and monitor individuals, potentially violating their privacy. The collected gaze data can be used for purposes beyond gaze estimation, such as advertising, surveillance, or profiling.

Another ethical issue is bias and fairness. Gaze estimation algorithms can be biased based on the training data and the demographic characteristics of the individuals in the dataset. For example, a gaze estimation algorithm trained on data from a predominantly male or white population may perform poorly on individuals from other demographic groups. This can lead to unequal treatment and discrimination based on race, gender, or other characteristics. To mitigate these biases, the following steps should be considered:

1. Data bias: Ensure that the training data used to develop the gaze estimation algorithms is diverse and representative of the population that the technology will be used on. This can help reduce the potential for bias in the algorithms.
2. Transparency: Provide clear information about the purpose and function of gaze estimation technology, how it works, and the type of data that has been collected. This can help to build trust and increase transparency around the use of the technology.
3. Data privacy: Implement strong data privacy and security measures to protect the collected gaze data. This can include encryption, anonymization, and secure storage of the data.
4. Regulation: Develop and enforce appropriate regulations and guidelines to ensure that gaze estimation technology is used in an ethical and responsible manner. This can include guidelines for the use of the technology in specific contexts, such as security or law enforcement, and for the handling of the collected gaze data.

5. User control: Allow individuals to control how their gaze data is collected and used. This can include options for opt-in and opt-out of data collection and the ability to access and delete their data.
6. Monitoring and evaluation: Regularly monitor and evaluate the performance and impact of gaze estimation technology to identify and address potential ethical concerns and biases. This can include ongoing audits and assessments of the technology and its usage.

Additionally, gaze estimation can be used in applications that have potential consequences for individual well-being and society as a whole, such as in security or law enforcement contexts. In these scenarios, it is important to ensure that the technology is used in a transparent, responsible, and ethical manner, and that appropriate safeguards are in place to protect individuals' rights and freedoms. In conclusion, it is important to consider the ethical implications of gaze estimation and to ensure that the technology is developed, used, and regulated in a responsible and ethical manner.

## 6.6 Legal implications

The collection and use of gaze data raises privacy concerns as it involves the monitoring and tracking of an individual's gaze behavior. This data can be sensitive and personal, and its collection and use must be done in compliance with privacy laws, such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States. Obtaining informed consent from individuals for gaze interaction and data collection is crucial, particularly if the interaction involves capturing and storing personal information.

The use of gaze technology can raise concerns about discrimination, particularly if the data collected is used to make decisions that affect an individual's rights and opportunities. For example, if gaze data is used to make decisions about employment, credit, or housing, there is a risk of discrimination based on race, gender,

or other protected characteristics. Gaze technology can be subject to bias, which may affect the accuracy and fairness of the data collected and used. For example, if the algorithms used to analyze gaze data are biased towards certain groups, this could result in discriminatory outcomes. Robots that interact with humans, including through gaze, must meet safety standards to prevent physical harm. Failure to do so could lead to product liability claims.

In order to address these legal concerns, it is important to implement gaze technology in a transparent manner. This includes providing clear explanations of the purpose and use of gaze data, as well as offering users the ability to control or delete the data collected. The use of gaze technology also raises ethical considerations, such as the impact on privacy, autonomy, and dignity. For example, the use of gaze tracking to monitor and control the behavior of individuals raises concerns about individual autonomy and privacy rights. This includes compliance with privacy laws, addressing the risk of discrimination, mitigating bias, promoting transparency, and considering ethical implications. By doing so, designers and developers can ensure that gaze technology is used in a responsible and ethical manner that protects the rights and interests of individuals.

## 6.7 Conclusion and Future Work

Detailed gaze annotation helps to unearth hidden layers in human interactions which can further help build automated dialogue systems. For future work, inter-rater reliability will be measured for all the videos that are annotated by the first author followed by automatic annotation of the manually-coded data will be conducted which would allow us to expand the corpus for multimodal interactions. The rapid advancements in the field of robotic technologies increases the importance of social robots that are built for interacting with people and are designed for various

contexts such as therapy, education, and industrial applications. Depending on the degree to which they would need the autonomous capacity to display socially acceptable behaviour for human comfort, the results of this thesis can be used in the implementation of gaze cues in avatars/robots such as Furhat.

# Bibliography


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

Adetunji, Raji Ridwan and Koh Sze (2012). 'Understanding Non-Verbal Communication across Cultures: A Symbolic Interactionism Approach'. In: *i-Come International Conference on Communication and Media*.

Admoni, Henny and Brian Scassellati (Mar. 2012). 'Robot Gaze Is Different From Human Gaze: Evidence that robot gaze does not cue reflexive attention'. In.

Aftab, Abdul Rafey (2019). 'Multimodal driver interaction with gesture, gaze and speech'. In: *2019 International Conference on Multimodal Interaction*, pp. 487–492.

Akechi, Hironori et al. (2013). 'Attention to eye contact in the West and East: Autonomic responses and evaluative ratings'. In: *PloS one* 8.3, e59312.

Akinyelu, Andronicus A and Pieter Blignaut (2020). 'Convolutional neural network-based methods for eye gaze estimation: A survey'. In: *IEEE Access* 8, pp. 142581–142605.

– (2022). 'Convolutional Neural Network-Based Technique for Gaze Estimation on Mobile Devices'. In: *Frontiers in artificial intelligence* 4, p. 796825.

Aliasghari, Pourya et al. (2020). 'Implementing a gaze control system on a social robot in multi-person interactions'. In: *SN Applied Sciences* 2, pp. 1–13.

Alnajar, Fares et al. (2013). 'Calibration-free gaze estimation using human gaze patterns'. In: *Proceedings of the IEEE international conference on computer vision*, pp. 137–144.

Andersson, Richard and Olof Sandgren (2016). 'ELAN Analysis Companion (EAC): A Software Tool for Time-course Analysis of ELAN-annotated Data'. In: *Journal of Eye Movement Research* 9.3.

Andrist, Sean et al. (Mar. 2014). 'Conversational Gaze Aversion for Humanlike Robots'. In: DOI: 10.1145/2559636.2559666.

Argyle, Michael and Mark Cook (1976a). *Gaze and mutual gaze*. Cambridge University Press.

– (1976b). 'Gaze and mutual gaze.' In.

Argyle, Michael, Mark Cook and Duncan Cramer (Dec. 1994). 'Gaze and Mutual Gaze'. In: *British Journal of Psychiatry* 165, pp. 848–850. DOI: 10.1017/S0007125000073980.

Azad, Pedram, Tamim Asfour and Ruediger Dillmann (2006). 'Combining appearance-based and model-based methods for real-time object recognition and 6d localization'. In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 5339–5344.

Bailey, Reynold et al. (2009). 'Subtle gaze direction'. In: *ACM Transactions on Graphics (TOG)* 28.4, pp. 1–14.

Bailly, Gérard, Stephan Raidt and Frédéric Elisei (June 2010). 'Gaze, conversational agents and face-to-face communication'. In: *Speech Communication* 52, pp. 598–612. DOI: 10.1016/j.specom.2010.02.015.

Bates, Douglas et al. (2015). 'Parsimonious mixed models'. In: *arXiv preprint arXiv: 1506.04967*.

Bavelas, Janet Beavin, Linda Coates and Trudy Johnson (2002). 'Listener responses as a collaborative process: The role of gaze'. In: *Journal of Communication* 52.3, pp. 566–580.

Beattie, Geoffrey W (1978). 'Sequential temporal patterns of speech and gaze in dialogue'. In: *Semiotica* 23.1-2, pp. 29–52.

– (1981). 'A further investigation of the cognitive interference hypothesis of gaze patterns during conversation'. In: *British Journal of Social Psychology* 20.4, pp. 243–248.

Beaudoin, Cindy and Miriam Beauchamp (Jan. 2020). 'Social cognition'. In: *Handbook of clinical neurology* 173, pp. 255–264. DOI: 10.1016/B978-0-444-64150-2.00022-8.

Becker-Asano, Christian and Hiroshi Ishiguro (2009). 'Laughter in social robotics-no laughing matter'. In: *Intl. Workshop on Social Intelligence Design*. Citeseer, pp. 287–300.

Bedford, Rachael et al. (2012). 'Precursors to social and communication difficulties in infants at-risk for autism: gaze following and attentional engagement'. In: *Journal of autism and developmental disorders* 42, pp. 2208–2218.

Berez, Andrea (Jan. 2007). 'Eudico linguistic annotator (Elan)'. In: *Lang. Document. Conserv.* 1.

Bertrand, Roxane et al. (Jan. 2006). 'Le CID - Corpus of Interactional Data: protocoles, conventions, annotations'. In: *Travaux Interdisciplinaires du Laboratoire Parole et Langage (TIPA)* 25, pp. 31–60.

Bhattacharyya, Pushpak (Aug. 2018). 'Applications of Eye Tracking in Language Processing and Other Areas: An Investigation Based on Eye-tracking'. In: pp. 23–46. ISBN: 978-981-13-1515-2. DOI: 10.1007/978-981-13-1516-9_2.

Blattgerste, Jonas, Patrick Renner and Thies Pfeiffer (2018). 'Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views'. In: *Proceedings of the Workshop on Communication by Gaze Interaction*, pp. 1–9.

Bonin, Francesca, Nick Campbell and Carl Vogel (2012). 'Laughter and topic changes: Temporal distribution and information flow'. In: *2012 IEEE 3rd international conference on cognitive infocommunications (CogInfoCom)*. IEEE, pp. 53–58.

Bousmalis, Konstantinos, Marc Mehu and Maja Pantic (2009). 'Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools'. In: *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, pp. 1–9.

Breazeal, Cynthia (2004). 'Social interactions in HRI: the robot view'. In: *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)* 34.2, pp. 181–186.

Brône, Geert (2020). 'Multimodal (di)stance in interaction. Eye gaze, irony and joint pretence'. In: *Multimodal Communication Seminar Series-Oxford University*.

Brône, Geert et al. (2017a). 'Eye gaze and viewpoint in multimodal interaction management'. In: *Cognitive Linguistics* 28.3, pp. 449–483.

– (2017b). 'Eye gaze and viewpoint in multimodal interaction management'. In: *Cognitive Linguistics* 28.3, pp. 449–483.

Brooks, Rechele and Andrew N Meltzoff (2005). 'The development of gaze following and its relation to language'. In: *Developmental science* 8.6, pp. 535–543.

Brown, Paula M (1991). 'Mechanisms for listener-adaptation in language production: Limiting the role of the model of the listener'. In: *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman* 105, pp. 117–142.

Brugman, Hennie and Albert Russel (2004). 'Annotating Multi-media/ Multi-modal Resources with ELAN'. In: *Proceedings of the Language Resources and Evaluation Conference*.

Brône, Geert and Bert Oben (Mar. 2014). 'InSight Interaction: a multimodal and multifocal dialogue corpus'. In: *Language Resources and Evaluation* 49, pp. 195–214. DOI: 10.1007/s10579-014-9283-2.

Burgoon, Judee K and Aaron E Bacue (2003). 'Nonverbal communication skills'. In: *Handbook of communication and social interaction skills*, pp. 179–219.

Burgoon, Judee K and Beth A Le Poire (1999). 'Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality'. In: *Communications Monographs* 66.2, pp. 105–124.

Bustos, Pablo et al. (2019). 'The CORTEX cognitive robotics architecture: Use cases'. In: *Cognitive systems research* 55, pp. 107–123.

Campana, Ellen et al. (Jan. 2002). 'Using Eye Movements to Determine Referents in a Spoken Dialogue System'. In: *Proceedings of the 2001 Workshop on Perceptive User Interfaces*. DOI: 10.1145/971478.971489.

Cassell, Justine et al. (2000). 'Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents'. In: *Embodied conversational agents* 1.

Cazzato, Dario et al. (2020). 'When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking'. In: *Sensors* 20.13, p. 3739.

Chawarska, Katarzyna, Ami Klin and Fred Volkmar (July 2003). 'Automatic Attention Cueing Through Eye Movement in 2-Year-Old Children With Autism'. In: *Child development* 74, pp. 1108–22. DOI: 10.1111/1467-8624.00595.

Cheng, Yihua et al. (2021). 'Appearance-based gaze estimation with deep learning: A review and benchmark'. In: *arXiv preprint arXiv:2104.12668*.

Chevalier, Pauline et al. (Dec. 2019). 'Examining joint attention with the use of humanoid robots-A new approach to study fundamental mechanisms of social cognition'. In: *Psychonomic Bulletin & Review* 27. DOI: 10.3758/s13423-019-01689-4.

Chong, Eunji et al. (2018). 'Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency'. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 383–398.

Chong, Eunji et al. (2020). 'Detecting attended visual targets in video'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5396–5406.

Chung, Adrian J et al. (2005). 'Extraction of visual features with eye tracking for saliency driven 2D/3D registration'. In: *Image and Vision Computing* 23.11, pp. 999–1008.

Cruces, Alejandro et al. (2022). 'Multimodal object recognition module for social robots'. In: *Iberian Robotics conference*. Springer, pp. 489–501.

Dale, Rick et al. (2013). 'The self-organization of human interaction'. In: *Psychology of learning and motivation*. Vol. 59. Elsevier, pp. 43–95.

Das, Dipankar et al. (2015). 'Supporting human–robot interaction based on the level of visual focus of attention'. In: *IEEE Transactions on Human-Machine Systems* 45.6, pp. 664–675.

Dawkins, Marian Stamp (2002). 'What are birds looking at? Head movements and eye use in chickens'. In: *Animal Behaviour* 63.5, pp. 991–998.

De Ruiter, JP (2005). 'If eye-gaze frequency drops, its relationship with turn-taking disappears'. In: *Poster presented at AMLAP*.

Degutyte, Ziedune and Arlene Astell (2021). 'The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings'. In: *Frontiers in Psychology* 12, p. 616471.

Ding, Yu et al. (2014). 'Laughter animation synthesis'. In: *Proc. AAMS 2014*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 773–780.

Doherty-Sneddon, Gwyneth and Fiona G Phelps (2005). 'Gaze aversion: A response to cognitive or social difficulty?' In: *Memory & cognition* 33.4, pp. 727–733.

Drewes, Heiko (2010). 'Eye gaze tracking for human computer interaction'. PhD thesis. lmu.

Duchowski, Andrew T (2018). 'Gaze-based interaction: A 30 year retrospective'. In: *Computers & Graphics* 73, pp. 59–69.

Duncan, Starkey (1972). 'Some signals and rules for taking speaking turns in conversations.' In: *Journal of personality and social psychology* 23.2, p. 283.

Duncan, Starkey and Donald W Fiske (1979). 'Dynamic patterning in conversation: Language, paralinguistic sounds, intonation, facial expressions, and gestures combine to form the detailed structure and strategy of face-to-face interactions'. In: *American Scientist* 67.1, pp. 90–98.

El Haddad, Kevin, Sandeep Nallan Chakravarthula and James Kennedy (2019). 'Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles'. In: *2019 International Conference on Multimodal Interaction*, pp. 259–263.

Emery, Nathan (Sept. 2000). 'The Eyes Have It: The Neuroethology, Function and Evolution of Social Gaze'. In: *Neuroscience and biobehavioral reviews* 24, pp. 581–604. DOI: 10.1016/S0149-7634(00)00025-7.

Farroni, Teresa et al. (Aug. 2002). 'Eye contact detection in humans from birth'. In: *Proceedings of the National Academy of Sciences of the United States of America* 99, pp. 9602–5. DOI: 10.1073/pnas.152159999.

Fink, Julia et al. (2011). 'People's perception of domestic service robots: same household, same opinion?' In: *Social Robotics: Third International Conference, ICSR 2011, Amsterdam, The Netherlands, November 24-25, 2011. Proceedings 3*. Springer, pp. 204–213.

Flensborg Damholdt, Malene et al. (2020). 'Towards a new scale for assessing attitudes towards social robots: The attitudes towards social robots scale (ASOR)'. In: *Interaction Studies* 21.1, pp. 24–56.

Flom, Ross, Kang Lee and Darwin Muir (2017). *Gaze-following: Its development and significance*. Psychology Press.

Frischen, Alexandra, Andrew P Bayliss and Steven P Tipper (2007). 'Gaze cueing of attention: visual attention, social cognition, and individual differences.' In: *Psychological bulletin* 133.4, p. 694.

Fröhlich, Marlen et al. (2016). 'Unpeeling the layers of language: Bonobos and chimpanzees engage in cooperative turn-taking sequences'. In: *Scientific reports* 6.1, p. 25887.

Fusaroli, Riccardo and Kristian Tylén (2012). 'Carving language for social coordination: A dynamical approach'. In: *Interaction studies* 13.1, pp. 103–124.

Gillberg, Christopher (Nov. 1998). 'Chromosomal disorders and autism'. In: *Journal of autism and developmental disorders* 28, pp. 415–25. DOI: 10.1023/A:1026004505764.

Ginzburg, Jonathan, Chiara Mazzocconi and Ye Tian (2020). 'Laughter as language'. In: *Glossa* 5.1.

Gironzetti, Elisa (2017). 'Multimodal and Eye-tracking Evidence in the Negotiation of Pragmatic Intentions in Dyadic Conversations: The Case of Humorous Discourse'. PhD thesis. Texas A&M University–Commerce.

Gironzetti, Elisa et al. (2016). 'Smiling synchronicity and gaze patterns in dyadic humorous conversations'. In: *Humor* 29.2, pp. 301–324.

Glenberg, Arthur M, Jennifer L Schroeder and David A Robertson (1998). 'Averting the gaze disengages the environment and facilitates remembering'. In: *Memory & Cognition* 26.4, pp. 651–658.

Glenn, Phillip (2003). *Laughter in Interaction*. Cambridge, UK: Cambridge University Press.

Gobel, Matthias S, Heejung S Kim and Daniel C Richardson (2015). 'The dual function of social gaze'. In: *Cognition* 136, pp. 359–364.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Goodwin, Charles (1980). 'Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning'. In: *Sociological inquiry* 50.3-4, pp. 272–302.

Goodwin, Charles et al. (1980). 'Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning'. In: *Sociological inquiry* 50.3-4, pp. 272–302.

Grammer, Karl (1990). 'Strangers meet: Laughter and nonverbal signs of interest in opposite-sex encounters'. In: *Journal of Nonverbal Behavior* 14, pp. 209–236.

Grynszpan, O. and Jacqueline Nadel (Jan. 2015). 'An eye-tracking method to reveal the link between gazing patterns and pragmatic abilities in high functioning autism spectrum disorders'. In: *Frontiers in Human Neuroscience* 8. DOI: 10.3389/fnhum.2014.01067.

Gu, Erdan and Norman Badler (Aug. 2006). 'Visual Attention and Eye Gaze During Multiparty Conversations with Distractions'. In: vol. 4133, pp. 193–204. ISBN: 978-3-540-37593-7. DOI: 10.1007/11821830_16.

Gullberg, Marianne and Kenneth Holmqvist (1999). 'Keeping an eye on gestures: Visual perception of gestures in face-to-face communication'. In: *Pragmatics & Cognition* 7.1, pp. 35–63.

Hadjikhani, Nouchine et al. (July 2008). 'Pointing with the eyes: The role of gaze in communicating danger'. In: *Brain and cognition* 68, pp. 1–8. DOI: 10.1016/j.bandc.2008.01.008.

Haensel, Jennifer X, Tim J Smith and Atsushi Senju (2022). 'Cultural differences in mutual gaze during face-to-face interactions: A dual head-mounted eye-tracking study'. In: *Visual Cognition* 30.1-2, pp. 100–115.

Hakkani-Tür, Dilek et al. (2014). 'Eye gaze for spoken language understanding in multi-modal conversational interactions'. In: *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 263–266.

Hanna, Joy E. and Susan E. Brennan (2007). 'Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation'. In: *Journal of Memory and Language* 57.4. Language-Vision Interaction, pp. 596 –615. ISSN: 0749-596X. DOI: https://doi.org/10.1016/j.jml.2007.01.008. URL: http://www.sciencedirect.com/science/article/pii/S0749596X07000174.

Harness Goodwin, Marjorie and Charles Goodwin (1986). 'Gesture and coparticipation in the activity of searching for a word'. In.

Harwerth, Ronald S et al. (1986). 'Multiple sensitive periods in the development of the primate visual system'. In: *Science* 232.4747, pp. 235–238.

Haxby, James, Elizabeth Hoffman and Maria Gobbini (July 2000). 'The Distributed Human Neural System for Face Perception'. In: *Trends in cognitive sciences* 4, pp. 223–233. DOI: 10.1016/S1364-6613(00)01482-0.

He, Kaiming et al. (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hessels, Roy S (2020). 'How does gaze to faces support face-to-face interaction? A review and perspective'. In: *Psychonomic bulletin & review* 27.5, pp. 856–881.

Heyselaar, Evelien, David Peeters and Peter Hagoort (2021). 'Do we predict upcoming speech content in naturalistic environments?' In: *Language, Cognition and Neuroscience* 36.4, pp. 440–461.

Ho, Simon, Tom Foulsham and Alan Kingstone (Aug. 2015a). 'Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions'. In: *PloS one* 10, e0136905. DOI: 10.1371/journal.pone.0136905.

– (Aug. 2015b). 'Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions'. In: *PloS one* 10, e0136905. DOI: 10.1371/journal.pone.0136905.

Hoffman, Guy and Cynthia Breazeal (2004). 'Collaboration in human-robot teams'. In: *AIAA 1st intelligent systems technical conference*, p. 6434.

Holler, Judith and Kobin H Kendrick (2015a). 'Unaddressed participants' gaze in multi-person interaction: optimizing recipiency'. In: *Frontiers in psychology* 6.98, pp. 1–14.

– (2015b). 'Unaddressed participants' gaze in multi-person interaction: optimizing recipiency'. In: *Frontiers in psychology* 6.98, pp. 1–14.

Holt, Liz (2012). 'Using laugh responses to defuse complaints'. In: *Research on Language & Social Interaction* 45.4, pp. 430–448.

Hortensius, Ruud and Emily S Cross (2018). 'From automata to animate beings: the scope and limits of attributing socialness to artificial agents'. In: *Annals of the new York Academy of Sciences* 1426.1, pp. 93–110.

Howes, Christine et al. (Jan. 2017). 'Disfluencies in dialogues with patients with schizophrenia'. In: ISBN: 9780991196760.

Hunyadi, László (June 2019). 'Agreeing/Disagreeing in a Dialogue: Multimodal Patterns of Its Expression'. In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.01373.

Ishii, Ryo, Shiro Kumano and Kazuhiro Otsuka (2015). 'Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings'. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 99–106.

Ishii, Ryo et al. (2016). 'Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings'. In: *ACM Transactions on Interactive Intelligent Systems (TIIS)* 6.1, pp. 1–31.

Jaber, Razan (2023). 'Towards Designing Better Speech Agent Interaction: Using Eye Gaze for Interaction'. PhD thesis. Department of Computer and Systems Sciences, Stockholm University.

Jefferson, Gail (1984). 'On the organization of laughter in talk about troubles'. In: *Structures of Social Action: Studies in Conversation Analysis*, pp. 346–369.

Jin, Tianlei et al. (2022). 'Depth-aware gaze-following via auxiliary networks for robotics'. In: *Engineering Applications of Artificial Intelligence* 113, p. 104924.

Jokinen, Kristiina et al. (2010). 'Turn-alignment using eye-gaze and speech in conversational interaction'. In: *Eleventh Annual Conference of the International Speech Communication Association*.

Jokinen, Kristiina et al. (July 2013). 'Gaze and turn-taking behavior in casual conversational interactions'. In: *ACM Transactions on Interactive Intelligent Systems* 3, p. 1. DOI: 10.1145/2499474.2499481.

Kamitani, Yukiyasu and Frank Tong (2005). 'Decoding the visual and subjective contents of the human brain'. In: *Nature neuroscience* 8.5, pp. 679–685.

Kar, Anuradha and Peter Corcoran (2017). 'A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms'. In: *IEEE Access* 5, pp. 16495–16519.

Kendon, Adam (1967a). 'Some functions of gaze-direction in social interaction'. In: *Acta psychologica* 26, pp. 22–63.

– (1967b). 'Some functions of gaze-direction in social interaction'. In: *Acta Psychologica* 26, pp. 22–63.

Kendrick, Connah et al. (2018). 'Towards real-time facial landmark detection in depth data using auxiliary information'. In: *Symmetry* 10.6, p. 230.

Kendrick, Kobin H and Judith Holler (2017). 'Gaze direction signals response preference in conversation'. In: *Research on Language and Social Interaction* 50.1, pp. 12–32.

Khan, M.s.L., Haibo Li and Shafiq Réhman (May 2016). 'Gaze Perception and Awareness in Smart Devices'. In: *International Journal of Human-Computer Studies* 92. DOI: 10.1016/j.ijhcs.2016.05.002.

Khan, Muhammad Qasim and Sukhan Lee (2019). 'Gaze and eye tracking: Techniques and applications in ADAS'. In: *Sensors* 19.24, p. 5540.

Kim, Aleksandr, Aljoša Ošep and Laura Leal-Taixé (2021). 'Eagermot: 3d multi-object tracking via sensor fusion'. In: *2021 IEEE International conference on Robotics and Automation (ICRA)*. IEEE, pp. 11315–11321.

Kobayashi, H and S Kohshima (1997). 'Unique Morphology of the Human Eye'. In: *Nature*. DOI: 10.1038/42842.

Kontogiorgos, Dimosthenis et al. (Oct. 2018). 'Multimodal Reference Resolution In Collaborative Assembly Tasks'. In: DOI: 10.1145/3279972.3279976.

Koochaki, Fatemeh and Laleh Najafizadeh (2018). 'Predicting intention through eye gaze patterns'. In: *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, pp. 1–4.

Koonce, Brett and Brett Koonce (2021). 'ResNet 50'. In: *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72.

Korkiakangas, Terhi (May 2018). 'Gaze Aversion and the Progress of Interaction'. In: pp. 208–239. ISBN: 9781315621852. DOI: 10.4324/9781315621852-7.

Koyasu, Hikari et al. (2020). 'The gaze communications between dogs/cats and humans: recent research review and future directions'. In: *Frontiers in psychology* 11, p. 613512.

Kuhn, Gustav et al. (2008). 'Misdirection in magic: Implications for the relationship between eye gaze and attention'. In: *Visual Cognition* 16.2-3, pp. 391–405.

Laube, Inga et al. (2011). 'Cortical processing of head-and eye-gaze cues guiding joint social attention'. In: *Neuroimage* 54.2, pp. 1643–1653.

Lavelle, Mary, Patrick Healey and Rosemarie McCabe (Sept. 2012). 'Is Nonverbal Communication Disrupted in Interactions Involving Patients With Schizophrenia?' In: *Schizophrenia bulletin* 39. DOI: 10.1093/schbul/sbs091.

Lee, Jina and Stacy Marsella (Aug. 2006). 'Nonverbal Behavior Generator for Embodied Conversational Agents'. In: pp. 243–255. ISBN: 978-3-540-37593-7. DOI: 10.1007/11821830_20.

Lemley, Joseph et al. (2019). 'Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems'. In: *IEEE Transactions on Consumer Electronics* 65.2, pp. 179–187.

Lian, Dongze et al. (2018). 'Multiview multitask gaze estimation with deep convolutional neural networks'. In: *IEEE transactions on neural networks and learning systems* 30.10, pp. 3010–3023.

Liu, Chaoran et al. (2012). 'Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction'. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 285–292.

Lombardi, Maria et al. (2022). 'Toward an attentive robotic architecture: Learning-based mutual gaze estimation in Human–Robot Interaction'. In: *Frontiers in Robotics and AI* 9, p. 770165.

Lücking, Andy, Sebastian Ptock and Kirsten Bergmann (2011). 'Assessing agreement on segmentations by means of Staccato, the Segmentation Agreement Calculator according to Thomann'. In: *International Gesture Workshop*. Springer, pp. 129–138.

Luiten, Jonathon, Tobias Fischer and Bastian Leibe (2020). 'Track to reconstruct and reconstruct to track'. In: *IEEE Robotics and Automation Letters* 5.2, pp. 1803–1810.

Marchesi, Serena et al. (2023). 'Cultural differences in joint attention and engagement in mutual gaze with a robot face'. In: *Scientific Reports* 13.1, p. 11689.

Marois, Alexandre et al. (2020). 'Real-time gaze-aware cognitive support system for security surveillance'. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 64. 1. SAGE Publications Sage CA: Los Angeles, CA, pp. 1145–1149.

Mason, Ian (2012). 'Gaze, positioning and identity in interpreter-mediated dialogues'. In: *Coordinating participation in dialogue interpreting* 102, p. 177.

Mason, Malia F, Elizabeth P Tatkow and C Neil Macrae (2005). 'The look of love: Gaze shifts and person perception'. In: *Psychological science* 16.3, pp. 236–239.

Massé, Benoit, Silèye Ba and Radu Horaud (2016). 'Simultaneous estimation of gaze direction and visual focus of attention for multi-person-to-robot interaction'. In: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6.

Matsumoto, David and Hyisung C Hwang (2016). 'The cultural bases of nonverbal communication.' In.

Mazzocconi, Chiara, Ye Tian and Jonathan Ginzburg (2020). 'What's your laughter doing there? A taxonomy of the pragmatic functions of laughter.' In: *IEEE Trans. on Affective Computing*.

McCarthy, Anjanie et al. (2008). 'Gaze display when thinking depends on culture and context'. In: *Journal of Cross-Cultural Psychology* 39.6, pp. 716–729.

McKay, Kate T et al. (2021). 'Visual attentional orienting by eye gaze: A meta-analytic review of the gaze-cueing effect.' In: *Psychological Bulletin* 147.12, p. 1269.

McMahan, Peter and James Evans (2018). 'Ambiguity and engagement'. In: *American Journal of Sociology* 124.3, pp. 860–912.

Mirenda, Patricia, Anne Donnellan and David Yoder (Jan. 1984). 'Gaze behavior: A new look at an old problem'. In: *Journal of Autism and Developmental Disorders* 13, pp. 397–409. DOI: 10.1007/BF01531588.

Mishra, Chinmaya and Gabriel Skantze (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)*. IEEE, pp. 1201–1208.

Mitterer, Holger and Eva Reinisch (2017). 'Visual speech influences speech perception immediately but not automatically'. In: *Attention, Perception, & Psychophysics* 79, pp. 660–678.

Moore, Monica M (2010). 'Human nonverbal courtship behavior—a brief historical review'. In: *Journal of Sex Research* 47.2-3, pp. 171–180.

Mori, Masahiro (Jan. 2020). 'THE UNCANNY VALLEY'. In: pp. 89–94. ISBN: 9781517905255. DOI: 10.5749/j.ctvtv937f.7.

Morillo-Mendez, Lucas et al. (2022). 'Age-Related Differences in the Perception of Robotic Referential Gaze in Human-Robot Interaction'. In: *International Journal of Social Robotics*, pp. 1–13.

Mosconi, Matthew et al. (Sept. 2005). 'Taking an "intentional stance" on eye-gaze shifts: A functional neuroimaging study of social perception in children'. In: *NeuroImage* 27, pp. 247–52. DOI: 10.1016/j.neuroimage.2005.03.027.

Müller, Philipp, Ekta Sood and Andreas Bulling (2020). 'Anticipating averted gaze in dyadic interactions'. In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–10.

Mundy, Peter (Sept. 2017). 'A Review of Joint Attention and Social-Cognitive Brain Systems in Typical Development and Autism Spectrum Disorder'. In: *European Journal of Neuroscience* 47. DOI: 10.1111/ejn.13720.

Murthy, LRD and Pradipta Biswas (2021). 'Appearance-based gaze estimation using attention and difference mechanism'. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 3137–3146.

Mutlu, Bilge, Nicholas Roy and Selma Šabanović (2016). 'Cognitive human–robot interaction'. In: *Springer handbook of robotics*, pp. 1907–1934.

Mutlu, Bilge et al. (2012). 'Conversational gaze mechanisms for humanlike robots'. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1.2, pp. 1–33.

Newell, Ben R and David R Shanks (2014). 'Unconscious influences on decision making: A critical review'. In: *Behavioral and brain sciences* 37.1, pp. 1–19.

Niewiadomski, Radoslaw et al. (2009). 'Greta: an interactive expressive eca system'. In: *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. Citeseer, pp. 1399–1400.

Nijenhuis, Jan and Thomas Bouchard Jr. (Feb. 2007). 'Replication of the hierarchical visual-perceptual-image rotation model in de Wolff and Buiten's (1963) battery of 46 tests of mental ability'. In: *Intelligence* 35, pp. 69–81. DOI: 10.1016/j.intell.2006.05.002.

Nijholt, Anton (2002). 'Embodied agents: A new impetus to humor research'. In: *The April Fools Day Workshop on Computational Humour*. Vol. 20. In: Proc. Twente Workshop on Language Technology, pp. 101–111.

Norris, Sigrid (Jan. 2004). *Analyzing Multimodal Interaction: A Methodological Framework*. ISBN: 041532856X. DOI: 10.4324/9780203379493.

Nummenmaa, Lauri et al. (Dec. 2009). 'Connectivity Analysis Reveals a Cortical Network for Eye Gaze Perception'. In: *Cerebral cortex (New York, N.Y. : 1991)* 20, pp. 1780–7. DOI: 10.1093/cercor/bhp244.

Ochs, Magalie and Catherine Pelachaud (2013). 'Socially Aware Virtual Characters: The Social Signal of Smiles [Social Sciences]'. In: *IEEE Signal Processing Magazine* 30.2, pp. 128–132. ISSN: 1053-5888. DOI: 10.1109/msp.2012.2230541. URL: http://dx.doi.org/10.1109/msp.2012.2230541.

Onyeulo, Eva Blessing and Vaibhav Gandhi (2020). 'What makes a social robot good at interacting with humans?' In: *Information* 11.1, p. 43.

Ozdem, Ceylan et al. (Jan. 2017). 'Believing androids - fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents'. In: *Social Neuroscience* 12.

O'Brien, Heather L, Paul Cairns and Mark Hall (2018). 'A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form'. In: *International Journal of Human-Computer Studies* 112, pp. 28–39.

Palmero, Cristina et al. (2018). 'Recurrent cnn for 3d gaze estimation using appearance and shape cues'. In: *arXiv preprint arXiv:1805.03064*.

Park, Seonwook, Adrian Spurr and Otmar Hilliges (2018). 'Deep pictorial gaze estimation'. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 721–738.

Parks, Daniel, Ali Borji and Laurent Itti (2015). 'Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes'. In: *Vision research* 116, pp. 113–126.

Pathirana, Primesh et al. (2022). 'Eye gaze estimation: A survey on deep learning-based approaches'. In: *Expert Systems with Applications* 199, p. 116894.

Petitjean, Cécile and Esther González-Martínez (2015). 'Laughing and smiling to manage trouble in French-language classroom interaction'. In: *Classroom Discourse* 6.2, pp. 89–106.

Phillips, Wendy, Simon Baron-Cohen and Michael Rutter (July 1992). 'The role of eye contact in goal detection: Evidence from normal infants and children with autism or mental handicap'. In: *Development and Psychopathology* 4, pp. 375 – 383. DOI: 10.1017/S0954579400000845.

Phutela, Deepika (2015). 'The importance of non-verbal communication'. In: *IUP Journal of Soft Skills* 9.4, p. 43.

Pinheiro, Ana P et al. (2017). 'Laughter catches attention!' In: *Biological psychology* 130, pp. 11–21.

Pomerantz, Anita and John Heritage (2012). 'Preference'. In: *The Handbook of Conversation Analysis*. Wiley Online Library, pp. 210–228.

Prasov, Zahar (2011). *Eye gaze for reference resolution in multimodal conversational interfaces*. Michigan State University. Computer Science.

Raclaw, Joshua and Cecilia E Ford (2017). 'Laughter and the management of divergent positions in peer review interactions'. In: *Journal of Pragmatics* 113, pp. 1–15.

Ragni, Marco and Frieder Stolzenburg (May 2015). 'Higher-Level Cognition and Computation: A Survey'. In: *KI - Künstliche Intelligenz* 29. DOI: 10.1007/s13218-015-0375-y.

Ramdane-Cherif, Z et al. (2004). 'Performance of a computer system for recording and analysing eye gaze position using an infrared light device'. In: *Journal of clinical Monitoring and Computing* 18, pp. 39–44.

Recasens, Adria et al. (2015). 'Where are they looking?' In: *Advances in neural information processing systems* 28.

Reddy, Vasudevi, Emma Williams and Amy Vaughan (2002). 'Sharing humour and laughter in autism and Down's syndrome'. In: *British journal of psychology* 93.2, pp. 219–242.

Rich, Charles et al. (2010). 'Recognizing engagement in human-robot interaction'. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 375–382.

Romaniuk, Tanya (2009). 'The 'Clinton Cackle': Hillary Rodham Clinton's Laughter in News Interviews'. In: *Crossroads of Language, Interaction, and Culture* 7, pp. 17–49.

Rossano, Federico (2012). 'Gaze behavior in face-to-face interaction'. PhD thesis. Radboud University Nijmegen.

– (2013). 'Gaze in Conversation'. In: *The handbook of conversation analysis*. Ed. by Jack Sidnell and Tanya Stivers. John Wiley & Sons. Chap. 15, p. 308.

Rossano, Federico, Penelope Brown and Stephen C Levinson (2009). 'Gaze, questioning and culture'. In: *Conversation Analysis: Comparative Perspectives*, pp. 187–249.

Ruffman, Ted et al. (Dec. 2001). 'Does Eye Gaze Indicate Implicit Knowledge of False Belief? Charting Transitions in Knowledge'. In: *Journal of experimental child psychology* 80, pp. 201–24. DOI: 10.1006/jecp.2001.2633.

Ruhland, Kerstin et al. (2015). 'A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception'. In: *Computer graphics forum*. Vol. 34. 6. Wiley Online Library, pp. 299–326.

Sandgren, Olof et al. (2012). 'Timing of gazes in child dialogues: A time-course analysis of requests and back channelling in referential communication'. In: *International Journal of Language & Communication Disorders* 47.4, pp. 373–383.

Sangeetha, SKB (2021). 'A survey on deep learning based eye gaze estimation methods'. In: *Journal of Innovative Image Processing (JIIP)* 3.03, pp. 190–207.

Saran, Akanksha et al. (2018). 'Human gaze following for human-robot interaction'. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 8615–8621.

Schneier, Franklin R et al. (2011). 'Fear and avoidance of eye contact in social anxiety disorder'. In: *Comprehensive psychiatry* 52.1, pp. 81–87.

Sekicki, Mirjana and Maria Staudte (2018). 'Eye'll Help You Out! How the Gaze Cue Reduces the Cognitive Load Required for Reference Processing'. In: *Cognitive Science* 42.8, pp. 2418–2458. DOI: 10.1111/cogs.12682. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12682. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12682.

Selting, Margret et al. (Jan. 1998). 'Gesprächsanalytisches Transkriptionssystem (Gat) [Conversational Analytic Transcription System]'. In: *Linguistische Berichte* 173, pp. 91–122.

Senju, Atsushi and Mark Johnson (Mar. 2009). 'The eye contact effect: Mechanisms and development'. In: *Trends in cognitive sciences* 13, pp. 127–34. DOI: 10.1016/j.tics.2008.11.009.

Shepherd, Stephen (Mar. 2010a). 'Following Gaze: Gaze-Following Behavior as a Window into Social Cognition'. In: *Frontiers in integrative neuroscience* 4, p. 5. DOI: 10.3389/fnint.2010.00005.

Shepherd, Stephen V (2010b). 'Following gaze: gaze-following behavior as a window into social cognition'. In: *Frontiers in integrative neuroscience* 4, p. 5.

Shi, Peiteng, Markus Billeter and Elmar Eisemann (2020). 'SalientGaze: Saliency-based gaze correction in virtual reality'. In: *Computers & Graphics* 91, pp. 83–94.

Shuai, Lan (Sept. 2012). 'Modelling the coevolution of joint attention and language'. In: *Proceedings. Biological sciences / The Royal Society* 279, pp. 4643–51. DOI: 10.1098/rspb.2012.1431.

Sidenmark, Ludwig and Hans Gellersen (2019). 'Eye, head and torso coordination during gaze shifts in virtual reality'. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 27.1, pp. 1–40.

Skantze, Gabriel (2021). 'Turn-taking in conversational systems and human-robot interaction: a review'. In: *Computer Speech & Language* 67, p. 101178.

Smith, Stephanie M and Ian Krajbich (2019). 'Gaze amplifies value in decision making'. In: *Psychological science* 30.1, pp. 116–128.

Somashekarappa, Vidya, Christine Howes and Asad Sayeed (2020). 'An annotation approach for social and referential gaze in dialogue'. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 759–765.

– (2021). 'A deep gaze into social and referential interaction'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 43. 43.

Somashekarappa, Vidya, Asad Sayeed and Christine Howes (2023). 'Neural Network Implementation of Gaze-Target Prediction for Human-Robot Interaction'. In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 2238–2244.

Spiller, Moritz et al. (2021). 'Predicting visual search task success from eye gaze data as a basis for user-adaptive information visualization systems'. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.2, pp. 1–25.

Sporer, Siegfried L and Barbara Schwandt (2007). 'Moderators of nonverbal indicators of deception: A meta-analytic synthesis.' In: *Psychology, Public Policy, and Law* 13.1, p. 1.

Stanley, Gordon and Donald S Martin (1968). 'Eye-contact and the recall of material involving competitive and noncompetitive associations'. In: *Psychonomic Science* 13.6, pp. 337–338.

Stevens, Catherine J. et al. (2016). 'Mimicry and expressiveness of an ECA in human-agent interaction: familiarity breeds content!' In: *Computational Cognitive Science* 2.1. ISSN: 2195-3961. DOI: 10.1186/s40469-016-0008-2. URL: http://dx.doi.org/10.1186/s40469-016-0008-2.

Stevenson, Marguerite B et al. (1986). 'The beginning of conversation: Early patterns of mother-infant vocal responsiveness'. In: *Infant behavior and Development* 9.4, pp. 423–440.

Sun, Zhong et al. (2020). 'One-step regression and classification with cross-point resistive memory arrays'. In: *Science advances* 6.5, eaay2378.

Taha, Kamal (2023). 'Semi-supervised and un-supervised clustering: A review and experimental evaluation'. In: *Information Systems*, p. 102178.

Terzioğlu, Yunus, Bilge Mutlu and Erol Sahin (Mar. 2020). 'Designing Social Cues for Collaborative Robots: The Role of Gaze and Breathing in Human-Robot Collaboration'. In: pp. 343–357. DOI: 10.1145/3319502.3374829.

Thepsoonthorn, Chidchanok, Ken-ichiro Ogawa and Yoshihiro Miyake (Jan. 2021). 'The Exploration of the Uncanny Valley from the Viewpoint of the Robot's Nonverbal Behaviour'. In: *International Journal of Social Robotics*, pp. 1–13. DOI: 10.1007/s12369-020-00726-w.

Tian, Ye, Chiara Mazzocconi and Jonathan Ginzburg (2016). 'When do we laugh?' In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 360–369.

Torres, Obed, Justine Cassell and Scott Prevost (1997). 'Modeling gaze behavior as a function of discourse structure'. In: *First International Workshop on Human-Computer Conversation*. Citeseer.

Trampert, Patrick et al. (2021). 'Deep neural networks for analysis of microscopy images—synthetic data generation and adaptive sampling'. In: *Crystals* 11.3, p. 258.

Uono, Shota and Jari K Hietanen (2015). 'Eye contact perception in the west and east: A cross-cultural study'. In: *Plos one* 10.2, e0118094.

Urbain, Jérôme et al. (2010). 'AVLaughterCycle'. In: *J. Multimodal User Interfaces* 4.1, pp. 47–58. DOI: 10.1007/s12193-010-0053-1. URL: https://doi.org/10.1007/s12193-010-0053-1.

Vail, Alexandria et al. (Oct. 2017). 'Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions'. In: pp. 490–497. DOI: 10.1109/ACII.2017.8273644.

Valenti, Roberto, Nicu Sebe and Theo Gevers (2011). 'Combining head pose and eye location information for gaze estimation'. In: *IEEE Transactions on Image Processing* 21.2, pp. 802–815.

Vargas-Cuentas, Natalia Indira et al. (Nov. 2017). 'Developing an eye-tracking algorithm as a potential tool for early diagnosis of autism spectrum disorder in children'. In: *PLOS ONE* 12, e0188826. DOI: 10.1371/journal.pone.0188826.

Vertegaal, Roel and Yaping Ding (Jan. 2002). 'Explaining effects of eye gaze on mediated group conversations: Amount or synchronization?' In: pp. 41–48. DOI: 10.1145/587078.587085.

Vickers, Joan N (2011). *Mind over muscle: the role of gaze control, spatial cognition, and the quiet eye in motor expertise.*

Wan, Zhonghua et al. (2021). 'Pupil-contour-based gaze estimation with real pupil axes for head-mounted eye tracking'. In: *IEEE Transactions on Industrial Informatics* 18.6, pp. 3640–3650.

Wang, Tian et al. (2023). 'Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey'. In: *Advanced Intelligent Systems*, p. 2300359.

Wang, Xinyao et al. (2020). 'Face manipulation detection via auxiliary supervision'. In: *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part I 27*. Springer, pp. 313–324.

Wood, Erroll et al. (2016). 'Learning an appearance-based gaze estimator from one million synthesised images'. In: *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 131–138.

Wykowska, Agnieszka et al. (Apr. 2014). 'Beliefs about the Minds of Others Influence How We Process Sensory Information'. In: *PloS one* 9, e94339. DOI: 10.1371/journal.pone.0094339.

Yuki, Masaki, William W Maddux and Takahiko Masuda (2007). 'Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States'. In: *Journal of Experimental Social Psychology* 43.2, pp. 303–311.

Yun, Sang-Seok (Oct. 2016). 'A gaze control of socially interactive robots in multiple-person interaction'. In: *Robotica* 35, pp. 1–17. DOI: 10.1017/S0263574716000722.

Zaraki, Abolfazl et al. (Apr. 2014). 'Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot'. In: *IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans* 44, pp. 157–168. DOI: 10.1109/THMS.2014.2303083.

Zhang, Chuang et al. (2011). 'Gaze estimation in a gaze tracking system'. In: *Science China Information Sciences* 54, pp. 2295–2306.

Zhang, Qiaohui et al. (Jan. 2004). 'Overriding errors in a speech and gaze multimodal architecture'. In: *Proceedings of the 9th international conference on intelligent user interfaces*, pp. 346–348. DOI: 10.1145/964442.964527.

Zhang, Xucong, Yusuke Sugano and Andreas Bulling (2019). 'Evaluation of appearance-based methods and implications for gaze-based applications'. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13.

Zhang, Xucong et al. (2020). 'Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation'. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, pp. 365–381.

Zhang, Yanxia, Jonas Beskow and Hedvig Kjellström (Oct. 2017). 'Look but Don't Stare: Mutual Gaze Interaction in Social Robots'. In: pp. 556–566. ISBN: 978-3-319-70021-2. DOI: \url{https://doi.org/10.1007/978-3-319-70022-9_55}.

Zohary, Ehud et al. (2022). 'Gaze following requires early visual experience'. In: *Proceedings of the National Academy of Sciences* 119.20, e2117184119.

# Appendix A

# Related Documents


**Consent Form**

**Human-Robot Interaction**

I , agree to participate in the study ''Human-Robot Interaction'', conducted by Vidya Somashekarappa who has (have) discussed the research project with me.

I have received, read, and kept a copy of the information letter/plain language statement. I have had the opportunity to ask questions about this research and I have received satisfactory answers. I understand the general purposes, risks, and methods of this research.

I consent to participate in the research project and the following has been explained to me:

- the research may not be of direct benefit to me
- my participation is completely voluntary
- my right to withdraw from the study at any time without any implications to me
- the risks including any possible inconvenience, discomfort or harm as a consequence of my participation in the research project
- the steps that have been taken to minimise any possible risks
- public liability insurance arrangements
- what I am expected and required to do
- whom I should contact for any complaints with the research or the conduct of the research
- I am able to request a copy of the research findings and reports
- security and confidentiality of my personal information.

In addition, I consent to:

- audio-visual recording of any part of or all research activities (if applicable)
- publication of results from this study on the condition that my identity will not be revealed.

**Name:** _____ (please print)

Signature: _____

Date: _____

Figure A.1: Consent Form


**Samtyckesformulär**

**Interaktion mellan människa och robot**

Jag , samtycker till att delta i studien ``Interaktion mellan människa och robot'', utförd av Asad Sayeed som har (har) diskuterat forskningsprojektet med mig.

Jag har tagit emot, läst och bevarat en kopia av informationsbrevet/förklaringen på klarspråk. Jag har haft möjlighet att ställa frågor om denna forskning och jag har fått tillfredsställande svar. Jag förstår de allmänna syftena, riskerna och metoderna för denna forskning.

Jag samtycker till att delta i forskningsprojektet och följande har förklarats för mig:

• forskningen kanske inte är till direkt nytta för mig
• mitt deltagande är helt frivilligt
• min rätt att när som helst dra mig ur studien utan att det påverkar mig
• riskerna inklusive eventuella besvär, obehag eller skada till följd av mitt deltagande i forskningsprojektet
• de åtgärder som har vidtagits för att minimera eventuella risker
• offentliga ansvarsförsäkringar
• vad jag förväntas och måste göra
• vem jag ska kontakta för eventuella klagomål med forskningen eller genomförandet av forskningen
• Jag kan begära en kopia av forskningsresultaten och rapporterna
• säkerhet och konfidentialitet för min personliga information.

Dessutom samtycker jag till:
• audiovisuell inspelning av någon del av eller all forskningsverksamhet (om tillämpligt)
• publicering av resultat från denna studie under förutsättning att min identitet inte avslöjas.

**Namn:** _____

Signatur: _____

Datum: _____

Figure A.2: Consent Form Swedish

**Anthropomorphism**

| | | | | | | |
|---|---|---|---|---|---|---|
| Fake | 1 | 2 | 3 | 4 | 5 | Natural |
| Machinelike | 1 | 2 | 3 | 4 | 5 | Humanlike |
| Unconscious | 1 | 2 | 3 | 4 | 5 | Conscious |
| Artificial | 1 | 2 | 3 | 4 | 5 | Lifelike |
| Moving rigidly | 1 | 2 | 3 | 4 | 5 | Moving elegantly |

**Animacy**

| | | | | | | |
|---|---|---|---|---|---|---|
| Dead | 1 | 2 | 3 | 4 | 5 | Alive |
| Stagnant | 1 | 2 | 3 | 4 | 5 | Lively |
| Mechanical | 1 | 2 | 3 | 4 | 5 | Organic |
| Artificial | 1 | 2 | 3 | 4 | 5 | Lifelike |
| Inert | 1 | 2 | 3 | 4 | 5 | Interactive |
| Apathetic | 1 | 2 | 3 | 4 | 5 | Responsive |

**Likeability**

| | | | | | | |
|---|---|---|---|---|---|---|
| Dislike | 1 | 2 | 3 | 4 | 5 | Like |
| Unfriendly | 1 | 2 | 3 | 4 | 5 | Friendly |
| Unkind | 1 | 2 | 3 | 4 | 5 | Kind |
| Unpleasant | 1 | 2 | 3 | 4 | 5 | Pleasant |
| Awful | 1 | 2 | 3 | 4 | 5 | Nice |

**Perceived Intelligence**

| | | | | | | |
|---|---|---|---|---|---|---|
| Incompetent | 1 | 2 | 3 | 4 | 5 | Competent |
| Ignorant | 1 | 2 | 3 | 4 | 5 | Knowledgeable |
| Irresponsible | 1 | 2 | 3 | 4 | 5 | Responsible |
| Unintelligent | 1 | 2 | 3 | 4 | 5 | Intelligent |
| Foolish | 1 | 2 | 3 | 4 | 5 | Sensible |

**Perceived Safety**

| | | | | | | |
|---|---|---|---|---|---|---|
| Anxious | 1 | 2 | 3 | 4 | 5 | Relaxed |
| Calm | 1 | 2 | 3 | 4 | 5 | Agitated |
| Still | 1 | 2 | 3 | 4 | 5 | Surprised |

Participant number: Age/Gender:

Figure A.3: Godspeed Questionnaire

| Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly agree |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | |
|---|---|
| **FA-S.1** | I lost myself in this experience. |
| **FA-S.2** | The time I spent using Furhat just slipped away. |
| **FA-S.3** | I was absorbed in this experience. |
| **PU-S.1** | I felt frustrated while talking to Furhat. |
| **PU-S.2** | I found furhat confusing to use. |
| **PU-S.3** | Using Furhat was taxing. |
| **AE-S.1** | This chat was interesting. |
| **AE-S.2** | Furhat was aesthetically appealing. |
| **AE-S.3** | Furhat appealed to my senses. |
| **RW-S.1** | Using Furhat was worthwhile. |
| **RW-S.2** | My experience was rewarding. |
| **RW-S.3** | I felt interested in this experience. |

Participant Number: Age/Gender:

Figure A.4: User Engagement Questionnaire

Dnr 2023-03044-01

Stockholm avdelning övrig

**BESLUT OCH YTTRANDE**

2023-06-29

**Sökande forskningshuvudman**
Göteborgs universitet

**Forskare som genomför projektet**
Asad Sayeed

**Projekttitel**
Mänskligt samtal med en social robot

**Uppgifter om ansökan**
Ansökan inkom till Etikprövningsmyndigheten 2023-05-11 och blev valid 2023-05-29.

---

Etikprövningsmyndigheten beslutar enligt nedan. Etikprövningsmyndigheten lämnar samtidigt ett rådgivande yttrande enligt 4 a § förordningen (2003:615) om etikprövning av forskning som avser människor.



**BESLUT**

Etikprövningsmyndigheten avvisar ansökan, det vill säga tar inte upp ansökan till prövning.

**Skäl för beslutet**
I det aktuella projektet kommer det inte att göras något ingrepp på en forskningsperson eller annan intervention på sätt som anges i 4 § etikprövningslagen. Det kommer inte att ske någon behandling av personuppgifter på så sätt som anges i 3 § etikprövningslagen. Mot bakgrund härav omfattas inte studien av bestämmelserna i 3-4 §§ etikprövningslagen och ska därför inte etikprövas.

**RÅDGIVANDE YTTRANDE**

Etikprövningsmyndigheten har inte några etiska invändningar mot forskningsprojektet.

---



Figure A.5: Ethical Approval