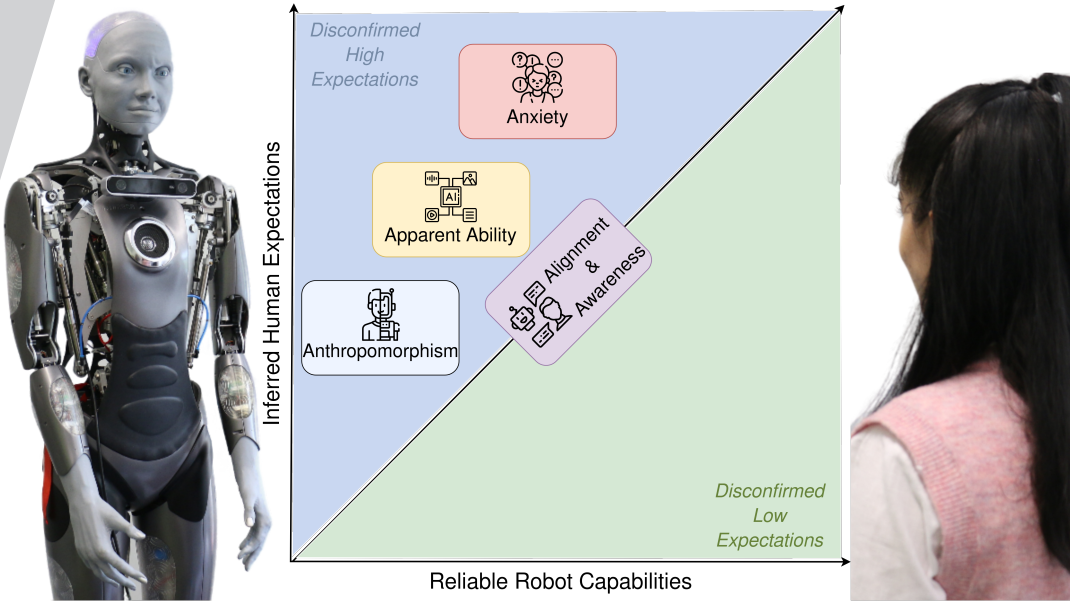
Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau



Ashita Ashok

Managing Expectation Gaps in Humanlike Robots: Toward Transparent and Acceptable Companions for University Students

RRLAB
THE ROBOTICS RESEARCH LAB

Managing Expectation Gaps in Humanlike Robots: Toward Transparent and Acceptable Companions for University Students

Thesis approved by
the Department of Computer Science

RPTU University Kaiserslautern-Landau

for the award of the Doctoral Degree

Doctor of Engineering (Dr.-Ing.)

to

Ashita Ashok

Date of Defense:	3 July 2026
Dean:	Prof. Dr. Christoph Garth
Head of Committee:	Prof. Dr. Annette Middelkoop-Bieniusa
Reviewer:	Prof. Dr. Dr. h.c. Karsten Berns
Reviewer:	Prof. Dr. Barbara Bruno
Reviewer:	Prof. Dr. Aleksandar Rodic

DE-386

Acknowledgements

The teacher is creator, preserver, and divine guide.

I would like to most sincerely thank my Doktorvater and boss, Prof. Berns, for giving me the once-in-a-lifetime opportunity to pursue a PhD under his guidance. I would never have considered beginning this PhD journey after my master's thesis at RRLab if it had not been for his belief that I could do it. Thank you for providing me with a working environment that included the most advanced humanoid robot, Ameca, the best workstations, funding for all twelve international research visits, and a full position. Whenever external trouble struck, you always had my back. You even taught me that a rejection from one venue because of unfair reviewers is only an opportunity to submit somewhere better. From rewriting all PhD chapters one week before my deadline to changing slides overnight before the defence, adapting to your feedback truly made my PhD the most presentable version of itself. Thank you. Finally, thank you for providing a safe space at the university in which I could truly grow over the past five years. I also warmly thank Mrs. Berns for her uplifting words and sincere encouragement, especially for reminding me in difficult moments that if anyone could carry it all through, it was me.

I would like to wholeheartedly thank Prof. Bruno for so humbly striking up a conversation with me at ROMAN 2023, where we connected over our cities, Karlsruhe and Kaiserslautern, being only one hour apart even though we met on the other side of the globe in Busan, South Korea. That small networking moment led to a butterfly effect of me exponentially fangirling, rightly so, over the exceptional research output of Prof. Bruno's SARAI Lab. All I want to say is thank you for accepting my request to be my thesis reviewer despite your hectic schedule during an intensely busy phase of your professorship and growing research recognition, and for the honesty and kindness you continuously bestowed upon me over the past three years.

I would like to humbly extend my gratitude to Prof. Rodić, whom I first met during our DAAD Serbia collaboration in 2022. Throughout this collaboration, several HRI discussions, and proposal writing, Prof. Rodić was nothing but kind and hands-on helpful. It was therefore extra special that he became my third reviewer. Thank you for accepting my request to be my PhD reviewer and for the honest feedback you provided throughout my PhD journey.

I would like to most gratefully extend my thanks to my PhD committee chair, Prof. Bieniusa, who agreed to chair my committee within seconds of me asking her. It was a true honour to defend my PhD under your lead in 2026, as you were the first professor to ever hear my PhD conceptual pitch during my first PhD presentation in 2022. I sincerely thank you for your constant support, kind encouragement, and for enabling the opportunity to provide robot demos for school children.

To put Prof. Allen's role in my PhD life into context, I must begin with my first PhD failure. My first research article was rejected from ICMI 2021, for understandable reasons, but without concrete feedback. Later, I attended Prof. Allen's workshop on "Scientific Writing", and it completely changed my brain chemistry on how to write a research article. My first PhD publication at BioRob 2022 was possible purely because of this workshop, and for that I will sincerely and forever be grateful.

When I say that this PhD thesis would not have progressed the way it did without the help of Prof. Czernochowski, I mean it. Her support in obtaining ethical approval from the Ethics Committee of the Department of Social Sciences for my HRI human experiments made my life as a budding HRI researcher much more coherent and aligned with the dos and don'ts of human experiments. I sincerely extend my thanks to you for this big role you played in my PhD journey.

Prof. Lachmann entered my PhD journey through the cognitive science student mini-conferences I attended while supervising students. His direct questions gave me a much-needed perspective in the final stages of this PhD, especially for explaining Ameca's uncanniness. One question stayed with me: in 2023, he asked, "Why do social robots need to have emotions?" I did not agree then, but three years of struggle later, I humbly do. For non-living things such as Ameca, performative facial expressions may indeed be the better term. Thank you for raising this question before the HRI community caught up.

I would like to thank Prof. Ebert for giving me insights into HCI human experiments right before my PhD journey, which helped me gain perspective for my own work. Toward the end of my PhD, I also deeply appreciated his efforts to support clearer ethics structures for the Computer Science Department, so that PhD students like me could navigate human-subject research and publication requirements with less uncertainty. I am truly grateful for his presence and for the way he supports students at RPTU.

May we work together with strength.

I would like to thank my collaborators, co-organizers, and co-authors for believing in me, including me in your ideas, saying yes to working on my ideas, and engaging in many productive discussions as a team. I hereby thank, chronologically: Prof. Zafar, Jakub, Jovan, Uroš, Maximilian, Silvana, Vlatka, Zhijing, Prof. Noccaro, Prof. Ivanova, Prof. Eden, Mauliana, Prof. Czernochowski, Prof. Lachmann, Dr. Mishra, Clarissa Sabrina, Rabea, Lan, Tamara, Tanu, Nihal, Prof. Laban, Prof. Babel, Rucha, Karla, Prof. Winkle, Prof. Fischer, Prof. Lee, Dr. Holthaus, Dr. Eymann, Eva Theresa, Nora, Mea-Sophie, Mehrbod, Prof. Wieching, Prof. Randall, Prof. Wulf, Nadezhda, Dr. Watanabe, Abdalla, Dr. Jan, Hana, Tim, Maike, Prof. Kraus, Raffaella, Prof. Rossi, Lukas, Prof. Eimler, Chimezie, Rishika, Clara, Julian, Prof. Rosen, and Dr. Beck.

I would like to thank my conference connections: BioRob 2022 – Jeonghan; ROMAN 2023 – Jauwairia, Maximilian, Elizabeth, Carolin; ARSO 2024 – Melissa, Ela, Barbara, Iaroslav, Hareesh, Annica, Thomas; ICSR 2024 – Chinmaya, Triniti, Victoria, Lukas, Jana, Marcel, Eduardo; HRI 2025 – Zoe, Oliver, Marcel, Eduardo; ROMAN 2025 – Julia, Kristina, Jan, Jairo, Ayeesha, Jayant, Karla, Marco, Andrew; HRI 2026 – Miguel, Verena, Chinmaya, Jairo, and Julia; ICRA 2026 – Verianti, Shubham, Olha, Caterina, Sara, Hana.

I also wholeheartedly thank all 929 voluntary experiment participants.

Service is the highest duty.

I would like to thank Frau Broschart for all her help from Day 1 to Day n: bureaucratic documents, Dienstreise applications, student Gutachten, certificates, and anything and everything outside research. Even the free plants from her garden mattered to me. Thank you very much!

I sincerely extend my thanks to Dr. Fernandez for helping manage the supervision of Cognitive Science students. Thank you for this structured coordination.

I would like to thank Jens for introducing me to the rights I had at the university through the Personalrat, so that I could advocate for myself.

I would like to thank Dr. Nissen and the TU-Nachwuchsring for funding my German B1.1 Abendkurse and my participation at ROMAN 2025.

I gratefully acknowledge the RPTU Open Access Fund for supporting the open-access publication of *Exploring LLM-powered multi-session human-robot interactions with university students* in *Frontiers in Robotics and AI*.

I would like to extend my heartfelt thanks to the Dekanat team, Dr. Schürmann, Frau Owens, and Frau Stuppy. A special shoutout goes to dear Frau Owens, who was very kind, endlessly helpful, and patiently repeated answers to all my questions. You were a blessing through this PhD journey, especially toward the end, when we sat down and planned my marathon run to meet the May 2026 deadline and the July 2026 defence. All of this was possible because of your advice and clear official to-dos. Thank you for being there!

In difficult times, strength lies in the group.

I thank the RRLab editorial management team that helped me during the two years of this group task: Alex, Oleksandr, and Pankaj. Thank you for the support.

I would like to thank the RRLab website management team, Eike and Nadezhda, for their support.

I would like to thank the RRLab system admins Dennis, Patrick V., Max, and Jannis for keeping the terminal system sane and working throughout. Special shout-out to Dennis for integrating Ollama models, and to Patrick V. for voicing the archiving of legacy code.

I would like to thank Qi for introducing Isaac Sim to RRLab members and for making Mandarin word enunciations fun when I found them super hard.

I would like to thank Abdalla and Hamza for the knowledge transfer on the workings of the RRLab AutoBus, and for our great ICRA 2026 collaboration article. Our hard work on this ambitious shot actually enabled me to attend the biggest robotics conference right before my graduation. Thanks, guys.

I would like to thank the RRLab HiWis Christel and Julian for their assistance with pupil-related activities such as LEGO courses, LEGO workshops, MINT Messe, Schülerinnentag, Nacht der Wissenschaft, lab tours, and Ameca demos. Shoutout also to Clara and Julian for their assistance with their first research endeavour at Seniorenwoche 2026.

I would like to thank Catharina and Prof. Wolf for sharing their research motivation with me during goal-dumping discussions over long cups of coffee. I especially thank Prof. Wolf

for his constant words of appreciation, acknowledging every milestone I achieved along the way.

I would also like to thank my last two master's students, Hana and Chimezie, for tolerating me during the final phase of my PhD by being magnificently self-reliant with programming and Ameca handling.

I would like to thank Sascha for being the best robot doctor RRLab needs on an hourly basis. RRLab robots would not function without your constant support, especially the fragile-expensive Ameca!

I would like to thank Dennis and Hussemann for being my office brothers and helping me stay sane: laughing at chaos, terminal magic commands, Dienstreise document self-reliance, decoding Hochdeutsch emails and Pfälzisch German phone calls, work-life-balance motivation – they did, I did not – complaining about the weather, making whiteboard memes, eating/sharing delicious meals, and believing in utter disbelief as I pushed through this PhD.

I would like to humbly thank Axel and Jakub for their patient programming advice before their replacement by ChatGPT. Jokes aside, I owe a lot of Ameca's direct-control tips, tricks, and hacker coding to Jakub, especially for ardently professing "it has to be possible" instead of "no, it was not in the manual." I extend my sincere thanks to you for being my helpful human agent.

I would like to thank Nadezhda, my ex-CogSci student turned RRLab-HRI follower, for giving me a taste of teamwork during the year we shared as coworkers. Although it was overwhelming at times, I will never forget burning the midnight oil at Ahrtal during your first Klausurtagung, fixing Latex and JASPing the HRI 2025 submission with no internet.

A special shout-out to RRLab doggo, Dexi, your presence warmed my heart especially on really tough days.

Knowledge gives humility.

I would also like to humbly acknowledge Ameca's manufacturer, Engineered Arts, for conceptualizing and building this remarkable machine, and Marcus for his support during the initial robot setup.

I would like to sincerely thank Dr. Rosen and Mr. Farimani for granting me permission to use their figures in this thesis. I also thank Merphi, for their robot posters, which repeatedly helped me gain a clear visual overview of the field.

I would like to thank the RPTU University library and its authorized access for allowing me to read and download research articles. I would furthermore like to thank Google Scholar for being my best friend throughout research reading and writing, especially through Google Scholar Alerts. I would like to thank Semantic Scholar for helping me acquire BibTeX files during the final days of writing, when Google Scholar blocked my access because it thought I was a bot due to my fast, automatic, robot-like behaviour. I would also like to cheer Miss Elbakyan for her access-to-research mindset.

A refuge for all beings.

I would like to thank the city of Kaiserslautern for giving me a sense of safety and freedom that I had not known before. Walking back alone after sunset, even for something as ordinary as grocery shopping, is a small thing here, but for me it is something I deeply cherish. I would also like to thank the VRN bus transportation system for bringing me safely to work and back home through my late-night research shenanigans.

A friend is remembered as equal to life.

I would like to thank An-Ka and Vera for their warm, women-in-research, and motivating presence in my PhD life. Knowing that I could come to either or both of you in moments of doubt and "let me tell you something" was truly a blessing.

I would like to extend my deepest gratitude to my research brothers, Chinmaya and Pankaj, for caring about me and my research, for guiding me when I needed it, and for proving through actions that I was not alone.

I would like to thank my friends who saw me through thick and thin, and watched me grow through this PhD journey in a foreign country. Thank you, Shilpa and Sachin, for the comforting Kerala meals that kept me warm, full, and grateful during my final PhD phase. Thank you Katherin and Heupel for the much needed relaxing start to New Year 2025 before everything got hectic. Thank you Klaas for the motivational PhD goals exchanges. Female friendship helped me through quiet battles: I lovingly thank Charu for her calming positivity; Dani for her introspective take on my research and day to day; dearest Cigdem, Leyla, and Azra for the courage of lionesses; Hande and Lucia for having my back like sisters; Tami for the much needed sushi venting dates; and Vani for being my Sundrop flower with daily affirmations. My childhood friends, Sahil, Pallavi, and Ajay: thank you very much for being part of my life and for standing as my pillars of strength telling me how proud of me you are!

Memory lives in the heart.

I thank Achamma, Simon, and Bruno for the loving memories I have left of you. Losing you during this PhD journey was heart-wrenching, but I thank you for the time you spent on Earth and for letting me be a tiny part of it.

The world is one family.

I lovingly thank the two sets of German parents I was blessed to have as my support system: Mama and Papa Meckel, and Mr. Markus and Frau Lambrecht. A special shoutout to Mr. Markus for the amazing opportunity to organize two Seniorenmesse events in Kaiserslautern, and to Frau Lambrecht for feeding me during the final days of PhD writing, gardening with me, and giving me all the style inspirations.

May we walk together in dharma.

To my toughest critic, Kuba: thank you for being my family and my best friend over the past 5.5 years. Thank you for being the best conference travel partner across the globe, allowing me to switch off my brain and simply enjoy the journey. Thank you for sharing every food and cinema adventures with such openness and joy. I am especially grateful for all the times you vrum-vrummed me around and picked me up from the university at 2 a.m.

Let mother be revered, let father be revered.

Mumma Mani, thank you for the disciplined military upbringing you gave us: for teaching me the importance of routine, respect for elders and wiser humans, and the courage to speak up when witnessing wrongdoing. The constant movement across the country trained me to be perceptive, to adapt effortlessly, and to find my place in new groups. Thank you also for raising me trilingual from the age of three. So much of what you gave me in childhood quietly helped me survive and excel during this PhD.

Papa Ashok, thank you for being the epitome of love, for nurturing us with care and concern, and for being the bulletproof jacket for your daughters. Thank you for praying to all 33 million Hindu deities and to St. Anthony of Padua for me every day of my life. No amount of words can describe my gratitude, but I hope one of the biggest honours I could give you was to carry your name with mine, to become Dr. Ashok, and to let my work be referred in the world as Ashok et al.

Let a person lift oneself by oneself.

"Last but not least, I wanna thank me, I wanna thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having no days off, I wanna thank me for never quitting, I wanna thank me for always being a giver, And tryna give more than I receive, I wanna thank me for, tryna do more right than wrong, I wanna thank me for just being me at all times." — SD.

Student Collaborators

I would like to acknowledge and thank all the students I supervised over the last five years. Supervising students not only expanded my knowledge but also made research in HRI at RRLab feel less solitary. As a fellow Ash, my goal was to train them and teach them to understand the power that was inside. They taught me, and I taught them. Through this experience, I learned how to support and manage students from different backgrounds and how to guide them towards their best work. I closely worked with each student for at least six months to one year, with biweekly meetings to maintain focus, provide conceptual and design feedback, prepare 30-min final presentations, record demo videos, and in several cases program Ameca side by side because help was always given at RRLab to those who asked for it. Some of these students became collaborators, and some even became friends who motivated me to keep going and finish this PhD. I hereby acknowledge all successful student works I supervised below.

Theses (30CP):

1. Accent Based Speech Emotion Recognition for Human-Robot Interaction [Gil Guevara 22]
2. Employment of Text Style Transfer for Human-Robot Interaction [Manjunath 22]
3. Observational Learning: Body Language Mimicking for an Anthropomorphic Social Robot [Tolj 23]
4. Modelling Empathetic Conversations for Human Computer Interaction [Sinha 23]
5. Multi-Turn Dialog with Visual Awareness on a Robot [Mohamed 23]
6. Dynamic Hand Gesture Recognition for Human-Robot Interaction [Islam 23]
7. A Multi-Turn Large Language Model for Emotionally Intelligent, Fact-Aware, and Memory-Enhanced Interactions [Mauliana 24]
8. Mimic Gestures from Audio-Visual Cues [Sevinc 24]
9. Vision and Dialogue: Visual Question Answer Framework [Rao 24]
10. Managing Anxiety in International Students Learning the German Language with LLM-Powered Social Robots [Helf 25]
11. Multi-Task Vision Model for Human-Robot Interaction [Huang 25]

12. The Influence of Social Robot Cues on Human Anxiety and Trust in Safety-Critical Task [Jeong 26]

13. Multimodal Emotion Recognition and Congruent Expressive Response Generation for the Ameca Social Humanoid Robot [Esiagu 26]

Projects (8CP):

1. Naturalistic Conversational Robot Gaze Control [Mandal 23]

2. RoboReID: Person Re-Identification via Visual and Verbal Cues for Human-Robot Interaction [Lu 23]

3. Transformer-Based First Interaction for HRI [Mauliana 23]

4. Mapping of Action Units to Valence-Arousal for Ameca and Emotion Detection [Terdale 23]

5. Multiparty Robot Gaze Algorithm [Sangappa 24]

6. Self-Awareness in Social Robots [Huang 24]

7. Event-Based Robot Behavior for Human-Robot Interaction [Gurung 24]

Lab Rotations (4CP):

1. Developing Self-Awareness: The Road Towards Emotional Robots [Carrasco Melo 21]

2. A Model of Robot Inner Speech Utilizing ACT-R to Simulate Cognitive Functions and Increase Empathy [Turner 21]

3. A Comparative Study of ACT-R vs. Non-ACT-R Based Chatbot Systems [Khalil 22]

4. Robot Behavior During Conflict Scenarios [Srivastava 22]

5. Accommodating Introversion in Social Robots [Oskurci 22]

6. Modulating Robot Behaviour During Self-Introduction Scenario [Goede 22]

7. Solving Puzzles Together: Piecing Together Functional Consciousness in a Humanoid Robot Assistant [Breininger 23]

8. Humanness in Robots: The Role of Emotions and Self-Awareness in Building Trust [Kushina 23]

9. Conflict Resolution in Human-Robot Interaction Scenarios: Questionnaire Validation [Nowak 23]

10. Robot's Backstory and How It Affects Human Perception [Nguyen 23]

11. Utilizing Ekman's Action Units for Emotional Expressions on Ameca's Face [Shakeebae 24]

12. Humor as a Repair Strategy for Robot Hallucination in HRI [Kukaj 24]

13. Perceived Consciousness in Human-Robot Interactions: Contributions of a GWT Architecture and Eye Gaze [Akinci 24]
14. Artificial Voice on Humanoid Robot Ameca [Jafarova 25]
15. When Do Emotions Begin? Timing Effects on Human Perception of Robot Animacy [Puertas Céspedes 25]
16. Narrative and Story-Telling in Human-Robot Interaction [Majumder 25a]
17. Closing the Social Robot Expectation Gap: A Teachable Approach to Social and Task-Based Robot Behaviour [Jeong 25]

Seminars (4CP):

1. Review of Cognitive Architectures for Enhanced Decision Making [Jois 22]
2. Context in Automated Affect Recognition in HRI [Lu 22]
3. Analysis on Explainable Artificial Intelligence for Enhanced Human-Robot Interaction [Sangappa 22]
4. Empathy in Human-Robot Interaction and Its Impacts [Terdale 22]
5. Hallway Interactions with Emah: Exploring the Role of Social Robots in Corridor Engagement [Chowdhury 23]
6. Robots in Two Worlds: A Comparative Study of Utility, Presence, and Anthropomorphism [Shaikh 24]
7. Robot Capabilities: A Reality Check for Humans [Werner 24]
8. Real-Time Dynamic Hand Gesture Recognition for Robotics with Dual-Stream 3D-CNNs [Rajendran 25]
9. Robot Ears: A Review of Automatic Speech Recognition for Human-Robot Interaction [Balaji 25]
10. Overview of Speech Recognition Pipeline in Embedded Systems with Focus on Voice Activity Detection [Sikorski 25]
11. AI-Based 3D Perception in Humanoid Robots [Das 25]

Research Collaborators

To my peers and senior collaborators: thank you for being my "on your left" when I needed to keep standing.

1. Personality Traits Assessment using PAD Emotional Space in Human-Robot Interaction [Zafar 21]
2. Paralinguistic Cues in Speech to Adapt Robot Behavior in Human-Robot Interaction [**Ashok 22**]
3. Social Robot Dressing Style: An Evaluation of Interlocutor Preference for University Setting [**Ashok 23**]
4. University Copycat Game: Observational Learning with Robot Tutee [**Ashok 24b**]
5. RoboReID: Audio-Visual Person Re-Identification by Social Robot [Lu 24]
6. "You Scare Me": The Effects of Humanoid Robot Appearance, Emotion, and Interaction Skills on Uncanny Valley Phenomenon [Berns 24]
7. Chit-Chat with Humanoid: LLMs for University Icebreaker Session [**Ashok 24a**]
8. From GPT to Open-Source LLMs: Advancing LAIC for More Efficient Qualitative Analysis [Arlinghaus 25]
9. "A Glimpse Into My World": Empathy Towards Emotional Robot Backstories at University [**Ashok 25b**]
10. "Thanks for the Practice!": LLM-Powered Social Robot as Tandem Language Partner at University [**Ashok 25a**]
11. Exploring LLM-Powered Multi-Session Human-Robot Interactions with University Students [Mauliana 25]
12. Teachable Social Robots: Managing Expectations in Highly Anthropomorphic Designs [Majumder 25b]
13. Does ChatGPT Know What Robots Are Thinking? Evaluation of Human-Created vs. AI-Generated Scientific Ideas [Eymann 25]
14. Understanding Human–AI Interaction in Aging Populations: Usability Differences between Robots and Avatars [Jahn 25]

15. SensHRPS: Sensing Comfortable Human-Robot Proxemics and Personal Space With Eye-Tracking [Kushina 25]
16. Designing Artificial Identity: The Identity Design Framework and Research Agenda [Bransky 26]
17. Embodied AI in the Wild: Comparing Older Adults' Interactions with an Avatar and a Humanoid Robot in a Public Space [Jahn 26]
18. When Imagination Meets Reality: Embodiment Order Effects in Human-Robot Interaction [Majumder 26]
19. Evaluating Multimodal Communication Methods for Autonomous Buses in Pedestrian-Dense University Environments [Mohamed 26]
20. Interpreting Robot Failure and Deception: The Role of Human-Likeness [Esposito 26]
21. Why Did My Robot Just Change Personality? Prompting Guidelines for a Grounded Robot Persona in LLM-Based HRI [**Ashok 26**]
22. When Female-Voiced Robots Talk Back: Effects of Assertive vs. Apologetic Response to245 Mistreatment on Bystander Perception [**Beck 26**]

Originality and Reuse Statement

This thesis presents an original dissertation-level synthesis of my doctoral research. It does not reproduce any publication, including my own prior articles, as verbatim thesis text. Where the thesis builds on published work, collaborative work, or student-supervised work, the corresponding contributions are identified in the Research Collaborators and Student Collaborators sections and acknowledged in the relevant chapters.

For consistency across the dissertation, all empirical analyses reported in this thesis were re-run in R/RStudio, and the results were rechecked, reinterpreted, and integrated into the thesis argument developed here. Reused figures, study materials, datasets, and external permissions are indicated where applicable.

Technical and Editorial Support

I would like to thank MSc Pankaj Deoli for proofreading the thesis, checking the storyline with respect to the computer-science contribution, and giving suggestions on which material should remain in the main chapters and which material should be moved to the appendix.

I would like to thank Dr. Vera Memmesheimer for answering my submission-related questions and for checking the template style and formal requirements.

I acknowledge the use of ChatGPT-5.5 for editorial support, including academic language polishing, storyline re-structuring, and assistance in improving the coherence of the thesis text. The final content, scientific interpretation, and responsibility for the thesis remain my own.

Abstract

Highly anthropomorphic social robots create a difficult interaction problem because their bodies suggest social, perceptual, and cognitive capability before the implemented system has shown what it can reliably support. This thesis examines that problem through Ameca, a full-body humanoid robot whose face, gaze, voice, speech, upper-body movement, and public recognisability make system behaviour socially interpretable. The work develops the 5A framework as an organising analytical lens for the anthropomorphic expectation gap. The framework diagnoses expectation mismatch through *Anthropomorphism*, *Apparent Ability*, and *Anxiety*, and addresses it through *Alignment* and *Awareness*.

Across small-scale empirical studies and system evaluations with $N = 929$ reported participants, the thesis examines Ameca and the developed Empathic Mechanized Anthropomorphic Humanoid System (EMAH system) across embodiment, auditory dialogue, visual perception, behaviour generation, and tandem-language practice. The embodiment studies show how identity, voice, gender, gaze, affect, movement, and failure can be interpreted as cues from one social agent. The auditory-dialogue work connects speech recognition, Large Language Model (LLM)-based response generation, memory, retrieval, prompt boundaries, and lip-synchronised Text-to-Speech (TTS) to the stability of Ameca's robot identity. The visual-perception work shows how eyes create expectations of seeing and fair attention, making Visual Question Answering (VQA), gaze control, active-speaker detection, and Hero selection socially consequential. The behaviour-generation work shows how facial expression, gaze shifts, mimicry, and pointing make ability visible, while timing, actuator limits, gesture detection, and morphology constrain what the robot can express or indicate. The tandem-language studies apply the framework in a bounded use case, where Ameca supports low-pressure German speaking practice through controlled dialogue, short turns, supportive correction, and clear role limits.

The central contribution is an implementation-grounded account of expectation management for highly humanlike robots. The thesis shows that usability and acceptance depend on the fit between apparent ability, actual capability, interaction role, and failure behaviour. For researchers and designers, the findings indicate that appearance, sensing, dialogue, memory, movement, prompts, explanation, and repair need to be designed as one interaction system. Ameca can support meaningful Human-Robot Interaction (HRI) when interlocutors can recognise both the social invitation of the body and the technical boundary behind it.

Contents

List of Figures **VI**

List of Tables **IX**

List of Acronyms **XII**

List of Icons **XIV**

List of Software and Tools **XVI**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Research Scope and Questions . . . 4
1.3 Approach and Contributions . . . 5
1.4 Structure of the thesis . . . 5

2 Human-Robot Interaction **7**
2.1 Foundations of Human-Robot Interaction . . . 8
2.2 Social Robots . . . 8
2.3 Social Robot Embodiment . . . 9
2.4 Interaction Modalities . . . 11
2.5 Wizard-of-Oz, Hybrid, and Autonomous Control . . . 11
2.6 Generative-AI Architectures . . . 12
2.7 User Studies and Evaluation Samples . . . 13
2.8 Social Robot Expectation Gap . . . 14
2.9 EMAH-powered Ameca . . . 14

3 The 5A Framework **17**
3.1 Problem Statement and the 3As . . . 21
3.1.1 Anthropomorphism . . . 22
3.1.2 Apparent Ability . . . 22
3.1.3 Anxiety . . . 23
3.2 Mitigation: The 2As . . . 25
3.2.1 Alignment via Exposure . . . 26
3.2.2 Awareness via Explanation . . . 27
3.3 Implications in this Thesis: Using the 5A Framework . . . 28

4 Robot Embodiment, Expectation, and Uncanniness 33
4.1 Overview: Appearance, Expectation, and Uncanniness . . . 34
4.2 Expectation mismatch in Ameca . . . 35
4.2.1 System-development indications of expectation mismatch . . . 35
4.2.2 Perception studies as evidence of expectation mismatch . . . 37
4.3 Expectation matching as a mitigation strategy . . . 44
4.4 Discussion . . . 48

5 Auditory Perception and Dialogue Management 49
5.1 Overview: Auditory Interaction as Ability Alignment . . . 50
5.2 Speech-to-Text . . . 51
5.3 Language Model Architecture . . . 51
5.3.1 ChitT5: Single-session Dialogue Ability . . . 52
5.3.2 MultiSessionT5: Repeated Interaction, Memory, and Role Stability 56
5.3.3 Open-Source LLM Integration with Ollama . . . 60
5.4 Text-to-Speech (TTS) . . . 62
5.5 Discussion . . . 63

6 Visual Perception and Gaze Management 65
6.1 Overview: Visual Perception Pipeline as Ability Alignment . . . 65
6.2 Visual Percepts for Social Interaction . . . 68
6.3 Visual Question Answering . . . 71
6.4 Robot Gaze Management . . . 74
6.4.1 From Gaze Bias to Speaker Detection . . . 75
6.4.2 Human “Hero” Detection for Active Speaker Selection . . . 77
6.5 Discussion . . . 83

7 Robot Behaviour Generation and Adaptation 85
7.1 Overview: Behaviour Generation as Ability Alignment . . . 86
7.2 Platform and Behaviour Generation Constraints . . . 88
7.3 Facial Behaviour Generation . . . 91
7.4 Eye Behaviour and Gaze Generation . . . 95
7.5 Body Gesture Generation and Mimicry . . . 95
7.6 Task-Oriented Deictic Gestures . . . 98
7.7 Discussion . . . 99

8 Social Robots as Tandem Language Partners 101
8.1 Overview: Ameca as a Bounded Tandem Partner . . . 102
8.2 Empirical Grounding: Foreign-Language Anxiety Survey . . . 105
8.3 Tandem System Architecture . . . 107
8.4 Single-Session Tandem Study . . . 108
8.5 Multi-Session Tandem Study . . . 110
8.6 Discussion . . . 115

9 Discussion **117**
9.1 Synthesis Across the Research Questions . . . 117
9.2 The 5A Framework Across the Thesis . . . 119
9.3 Implications for HRI Design . . . 119
9.4 Limitations . . . 120
9.5 Future Work . . . 120

10 Conclusion **123**

Bibliography **125**

Appendices **157**

A Appendix **157**
A.1 System-development details for Chapter 4 . . . 157
A.1.1 Audio-visual person re-identification . . . 157
A.1.2 Empathetic conversational system . . . 158
A.1.3 Multi-turn dialogue with visual awareness . . . 158
A.2 YouTube comment examples for Chapter 4 . . . 158
A.2.1 Name and identity continuity . . . 158
A.2.2 Voice continuity . . . 159
A.2.3 Gendered voice-body mismatch . . . 159
A.2.4 Apparent mind, empathy, and attachment . . . 160
A.2.5 Technical failure as identity failure . . . 160
A.2.6 Movement and expression mismatch . . . 161
A.3 Voice, gender, and conflict-study details for Chapter 4 . . . 161
A.3.1 Artificial-voices study . . . 161
A.3.2 Voice-emotion-onset study . . . 162
A.3.3 Female-voice conflict study . . . 165
A.4 Affective-framing study details for Chapter 4 . . . 168
A.4.1 Self-awareness study . . . 169
A.4.2 Backstory study . . . 171
A.5 Uncanniness study details for Chapter 4 . . . 174
A.5.1 Research questions . . . 174
A.5.2 Robot implementation . . . 174
A.5.3 User study . . . 174
A.5.4 Open-ended responses . . . 174
A.5.5 Results . . . 176
A.6 Robot-as-tutee study details for Chapter 4 . . . 177
A.6.1 CopyCat gesture-learning study . . . 177
A.6.2 Storytelling tutee study . . . 178
A.7 Embodiment-order study details for Chapter 4 . . . 179
A.7.1 Research goal . . . 179
A.7.2 Robot implementation . . . 179

A.7.3 User study . . . 180
A.7.4 Results . . . 180
A.8 Social-repair study details for Chapter 4 . . . 180
A.8.1 Humour repair study . . . 181
A.8.2 Reminder-conflict repair . . . 184
A.9 Rule-based dialogue baseline details for Chapter 5 . . . 185
A.10 Speech-to-text implementation details for Chapter 5 . . . 185
A.10.1 Google SpeechRecognition module . . . 185
A.10.2 WhisperX module . . . 185
A.11 ChitT5 model and evaluation details for Chapter 5 . . . 185
A.11.1 Instruction script . . . 187
A.11.2 Model training and deployment . . . 187
A.11.3 Extended CSRS and Godspeed results . . . 188
A.11.4 Qualitative patterns . . . 188
A.12 MultiSessionT5 model and evaluation details for Chapter 5 . . . 189
A.12.1 Model training and architecture . . . 189
A.12.2 Inference and robot deployment . . . 190
A.12.3 Study procedure . . . 190
A.12.4 Extended quantitative results . . . 190
A.12.5 Qualitative patterns . . . 191
A.13 Ollama prompt and implementation details for Chapter 5 . . . 193
A.13.1 Grounded Ameca system prompt . . . 193
A.14 Text-to-Speech implementation details for Chapter 5 . . . 194
A.15 Visual percepts for Chapter 6 . . . 194
A.15.1 Dynamic hand gesture recognition . . . 195
A.15.2 Object detection and robot self-recognition . . . 196
A.15.3 Tangram-based visual perception and joint attention . . . 197
A.15.4 Human attribute recognition for adaptive behaviour . . . 198
A.16 Visual question answering details for Chapter 6 . . . 199
A.16.1 Model selection and VQA implementation . . . 199
A.16.2 Example VQA outputs . . . 201
A.16.3 Priming stimuli . . . 201
A.16.4 Study design and measures . . . 202
A.16.5 Extended descriptive results . . . 202
A.16.6 Question categories . . . 202
A.17 Gaze management and Hero selection details for Chapter 6 . . . 204
A.17.1 Single-person gaze control implementation . . . 204
A.17.2 Single-person gaze pilot evaluation . . . 205
A.17.3 Multi-party active-speaker detection implementation . . . 206
A.17.4 Multi-party active-speaker pilot evaluation . . . 206
A.17.5 Hero selection deployment notes . . . 207
A.17.6 Hero selection evaluation setup . . . 208
A.17.7 Public-deployment bystander behaviour . . . 209
A.18 Behaviour-control parameters for Chapter 7 . . . 209

A.18.1 Exposed behaviour-control values 209
A.19 Facial behaviour generation details for Chapter 7 211
A.19.1 Facial Action Coding System (FACS) reference and Action Unit (AU)-inspired mapping 211
A.19.2 Exploratory AU-inspired facial configurations 213
A.19.3 Voice-Emotion-Onset control combinations 216
A.20 Gesture mimicry and body-behaviour details for Chapter 7 216
A.20.1 CopyCat candidate gesture set 216
A.20.2 Final CopyCat gesture set 218
A.20.3 Supervised gesture-recognition setup 218
A.20.4 Storytelling gesture-teaching setup 219
A.20.5 Live one-to-one mimicry setup 219
A.21 Task-oriented deictic gesture details for Chapter 7 220
A.21.1 Tangram pointing 220
A.21.2 Pill-selection pointing 220
A.22 Extended statistical results for Chapter 8 220
A.22.1 Survey reliability, group descriptives, and regression details 222
A.22.2 Single-session reliability and outcome tests 222
A.22.3 Multi-session reliability and paired comparisons 224
A.23 Transcript-analysis details for Chapter 8 226
A.23.1 Transcript metrics 226
A.23.2 Session 1 and Session 5 transcript metrics 226
A.24 Artificial-Voices Study Questionnaire 227
A.25 Voice-Emotion-Onset Study Questionnaire 229
A.26 Conflict Scenario Study Questionnaire 232
A.27 Visual Question Answering Study Questionnaire 235
A.28 Survey materials and FLACS items for Chapter 8 237
A.29 Single-session tandem prompts and implementation details for Chapter 8 . 239
A.29.1 Immersive German-only prompt 239
A.29.2 Bilingual German-English prompt 240
A.30 Multi-session tandem prompt and implementation details for Chapter 8 . . 241
A.30.1 Multi-session tandem prompt 241

Curriculum Vitae **243**

List of Figures

1.1 Public interaction with Ameca at the Elderly Convention 2025 3

2.1 Embodied interaction lab setup with Ameca 7
2.2 Social robot embodiments across anthropomorphic appearance 10
2.3 Ameca sensing components used in this thesis 16

3.1 The 5A framework for expectation management 19
3.2 Uncanny-valley placement of humanoid robots 20

4.1 Embodiment expectation-alignment overview 34
4.2 Ameca in the RRLab corridor 42
4.3 RRLAB robots in the uncanny-valley comparison 43

5.1 Auditory expectation-management pipeline 50
5.2 ChitT5 expression mapping and evaluation setup 55
5.3 CSRS ratings in the ChitT5 study 56
5.4 MultiSessionT5 workflow 58
5.5 Engagement across MultiSessionT5 sessions 59
5.6 Conversation-partner expectations after repeated interaction 59

6.1 Visual expectation-management pipeline 66
6.2 Robot self-recognition as visual disambiguation 70
6.3 Tangram game-playing setup 71
6.4 Robot-side VQA loop 73
6.5 VQA priming effects on Godspeed ratings 74
6.6 Simulated group-interview gaze-bias setup 78
6.7 Real-time social-attention system 80
6.8 Public Hero-selection deployment setup 82
6.9 Controlled active-speaker demonstration 83

7.1 Behaviour-generation expectation-management pipeline 86
7.2 Ameca morphology and facial surface 88
7.3 AU-inspired face reference 93
7.4 Ameca gaze targets 96
7.5 Live mimicry pipeline 97
7.6 Live mimicry demonstration 98

8.1 Overview of the tandem-partner chapter 102
8.2 Self-confidence and speech anxiety 107
8.3 Encouraging Ameca expression 110
8.4 Expectations by Tandem mode 111
8.5 RALL by Tandem mode 111
8.6 Multi-session tandem setup 113
8.7 Multi-session anthropomorphism 114
8.8 Multi-session RALL 115

A.1 Robot images for gender attribution 163
A.2 Humour repair responses 181
A.3 MultiSessionT5 Sociability, Disturbance ratings 192
A.4 Public visitors documenting Ameca 209
A.5 Selected FACS action units 213
A.6 Survey speech anxiety by group 221
A.7 Single-session lexical gain 223
A.8 Single-session likeability and safety 223
A.9 Multi-session pre-post changes 225
A.10 Survey routing pages 237

List of Tables

3.1 5A framework definitions 18
3.2 3A evidence in Ameca 21
3.3 2A mitigation evidence in Ameca 26

4.1 Related studies on robot appearance and expectation 36
4.2 System-development studies and expectation mismatch 37
4.3 YouTube comment patterns around Ameca identity 38
4.4 Voice studies and identity mismatch in Ameca 39
4.5 Affective-framing studies in Ameca 40
4.6 Robot-as-tutee studies for expectation matching 45
4.7 Embodiment order and expectation matching 46
4.8 Social repair studies for expectation matching 47

5.1 Speech-to-text solutions across EMAH system systems 52
5.2 LLM-powered social robots 53
5.3 ChitT5 single-session dialogue study 54
5.4 MultiSessionT5 as repeated dialogue ability 57
5.5 Technical progression from ChitT5 to MultiSessionT5 61
5.6 Ollama-based local LLM integration 62

6.1 Visual perception and gaze in social robots 67
6.2 Visual percepts for Ameca 69
6.3 Visual question answering with Ameca 72
6.4 Gaze-management development path 75

7.1 Related work on behaviour generation 87
7.2 Ameca behaviour-control groups 89
7.3 Behaviour-generation systems 92
7.4 AU-inspired facial mappings 94

7.5 Voice-Emotion-Onset facial expressions 100

8.1 Educational robot roles 103
8.2 RRLab robot use cases 104
8.3 RALL systems with social robots 104
8.4 Tandem studies and 5A alignment 105
8.5 Shared tandem system configuration 108
8.6 Single-session LLM configuration 109
8.7 Multi-session LLM configuration 112

A.1 RRLAB YouTube videos 159
A.2 Artificial-voices descriptives 162
A.3 Voice-onset animacy ratings 165
A.4 Voice-gender-onset descriptives 165
A.5 Conflict video dialogue 166
A.6 Conflict-condition descriptives 168
A.7 Paired comparisons between assertive and apologetic conditions. 168
A.8 Self-awareness monologues 170
A.9 Self-awareness descriptives 171
A.10 Backstory monologues 172
A.11 Backstory HRIES descriptives 173
A.12 Backstory empathy descriptives 173
A.13 Uncanny-valley descriptives 176
A.14 Hallucination detection 183
A.15 Humour repair ratings 184
A.16 Best-fitting humour repairs 184
A.17 T5 and Flan-T5 automatic metrics 187
A.18 ChitT5 CSRS item results 188
A.19 MultiSessionT5 datasets 189
A.20 MultiSessionT5 helper classifiers 190
A.21 MultiSessionT5 generation metrics 190
A.22 ChitT5 and MultiSessionT5 comparison 191
A.23 Dynamic hand gesture recognition setup 196
A.24 Object detection and robot self-recognition setup. 197
A.25 Tangram-based visual perception and joint-attention setup. 198
A.26 Human attribute recognition for adaptive robot behaviour. 199

A.27 VQA model-selection criteria for robot-side integration. 199
A.28 VQA example outputs 201
A.29 VQA study design 202
A.30 VQA priming descriptives 203
A.31 VQA question categories 204
A.32 Single-person gaze pilot 206
A.33 Multi-party active-speaker pilot 207
A.34 Hero-selection ablation 208
A.35 Ameca control values 210
A.36 FACS-to-Ameca mapping 211
A.37 Exploratory facial configurations 213
A.38 Voice-Emotion-Onset setup 216
A.39 CopyCat candidate gestures 217
A.40 Final CopyCat gesture set retained for live interaction. 218
A.41 Supervised gesture-recognition setup for Ameca mimicry. 218
A.42 Storytelling gesture-teaching setup. 219
A.43 Live mimicry implementation 219
A.44 Tangram deictic-gesture setup. 220
A.45 Pill-selection deictic-gesture setup. 221
A.46 Survey reliability checks 221
A.47 Survey group descriptives 222
A.48 Survey regression models 222
A.49 Single-session reliability checks 222
A.50 Single-session lexical gain 223
A.51 Single-session ANOVA summary 224
A.52 Single-session tandem evaluation 224
A.53 Multi-session reliability checks 224
A.54 Multi-session descriptive change 226
A.55 Multi-session paired comparisons 226
A.56 Transcript-analysis metrics 227
A.57 Multi-session user speech metrics 227
A.58 Modified FLACS items 237

List of Acronyms

AI Artificial Intelligence. 8, 24, 26, 27, 31

AU Action Unit. 30, 91, 93, 94, 99, 100, 118, 120, 211–215, V, IX

EMAH system Empathic Mechanized Anthropomorphic Humanoid System. 14, 15, 52, 54, 57, 62, 70, 95–97, 157, 158, 173, 174, 177, 178, 190, 193, 195–198, 228, 229, s, IX

FACS Facial Action Coding System. 91–94, 211–213, 215, V

GenAI Generative Artificial Intelligence. 7, 12, 14, 15, 17, 23, 120

HRI Human-Robot Interaction. 2–8, 11–15, 17, 23, 24, 28, 52–55, 57, 62, 68, 71–77, 82, 99, 118–120, 124, s, III

LLM Large Language Model. 12, 18, 21, 23, 26–29, 44, 47, 52–54, 56, 58, 60–62, 101, 104, 107, 108, 119, 120, 123, 181, 182, 185, 193, s, II, IX

NARS Negative Attitudes towards Robots Scale. 24

RALL Robot-Assisted Language Learning. 101, 103, 105, 109, 110, 112–114

RAS Robot Anxiety Scale. 24

RRLab Robotics Research Laboratory. 2–4, 14, 33, 37, 41, 42, 46, 68–71, 77, 90, 95, 158, 159, 174, 195–197

TTS Text-to-Speech. 62–64, 119, 120, 195, s, II

VAP Visual Attentive Prompting. 98, 120

VLM Vision-Language Model. 12, 18, 23, 67, 68, 98, 120, 220, 221

VQA Visual Question Answering. 65, 68, 71–74, 83, 84, 118–120, 123, 199–202, 204, 236, s, IV, XI

WoZ Wizard-of-Oz. 11, 12, 28, 104

List of Icons

Some icons used in this dissertation were obtained from Flaticon (https://www.flaticon.com/)..

Icon	Meaning	Icon Attribution
	Chat / dialogue interaction	Smashicons[1]
	Research question	Freepik[2]
	Machine learning component	Freepik[3]
	Large language model component	–
	Vision-language model component	–
	Robot system or embodiment	Konkapp[6]
	Human interlocutor / participant	Freepik[7]
	Quotation or qualitative example	Iconjam[8]
	Result or finding	Uniconlabs[9]
	Iterative development / refinement	Uniconlabs[10]
	Generative AI component	andinur[11]

[1] https://www.flaticon.com/free-icons/robot
[2] https://www.flaticon.com/free-icons/question
[3] https://www.flaticon.com/neural-network
[6] https://www.flaticon.com/free-icons/humanoid
[7] https://www.flaticon.com/free-icons/politician
[8] https://www.flaticon.com/free-icons/quotes
[9] https://www.flaticon.com/free-icons/result
[10] https://www.flaticon.com/free-icons/improvement
[11] https://www.flaticon.com/free-icons/ai-application

List of Software and Tools

Software / Tool	**Use in this thesis**
draw.io[1]	Used to create flowcharts and schematic diagrams throughout the thesis.
RStudio[2]	Used with R for statistical analysis, assumption checks, and result visualisation.
ChatGPT Images 2.0[3]	Used to generate or edit selected figure illustrations where explicitly marked in the corresponding figure captions.
Overleaf[4]	Used for LaTeX thesis writing and version control.
Google Scholar[5]	Used to retrieve BibTeX entries and check publication metadata.

[1] https://app.diagrams.net/
[2] https://rz.rptu.de/en/dienstleistungen/serverdienste/webbasierte-software/rstudio
[3] https://openai.com/index/introducing-chatgpt-images-2-0/
[4] https://www.overleaf.com/
[5] https://scholar.google.com/

1. Introduction

> *“I wouldn’t have a deeper conversation about something worrisome with the robot and expect it to be nearly as understanding and helpful as a human might be.”*
>
> Participant P15, Self-awareness study, Chapter 4, describing Ameca’s affective delivery of positive and negative news.

1.1 Motivation

Public expectations of robots are often formed before physical interaction, as science fiction films repeatedly present humanlike machines as tools, companions, servants, rivals, or threats [Złotowski 15]. These roles attach culturally learned expectations to visible and audible cues such as face, voice, age, expressivity, and autonomy, so viewers often interpret a robot through a familiar narrative frame before its robotic behaviour is shown in detail [Perugia 23]. In *I, Robot* (2004), the humanoid robot Sonny is placed within a storyline of danger, suspicion, and moral ambiguity, where its movements, expressions, and decisions are interpreted through the possibility that it may be a threat [Laakasuo 21]. In *Ex Machina* (2014), the android Ava is placed within a storyline of intimacy and manipulation, where her conversational timing, questioning, eye gaze, and emotionally responsive turn-taking make her autonomy appear strategic and personally directed [Złotowski 17]. Although *Ex Machina* includes scenes that show how Ava was created and programmed,[1] the dominant social impression is shaped less by this technical exposition than by the later revelation that Ava used the protagonist Caleb to escape and killed her creator Nathan with the help of another android.[2] Science fiction films featuring humanoid robots therefore make robots understood through perceived social character first, while the technical systems that define sensing, dialogue, autonomy, and behavioural control remain secondary to the narrative memory they produce.

[1] https://youtu.be/ruOXWHbyfjo?si=o59RJwe76Hd1QkPw
[2] https://youtu.be/D___TDSeE0dE?si=Nnz6LaNt7NlmNqTC

Online robot videos carry these film-shaped expectations into encounters with existing robot systems. Viewers watch real robots being tested, praised, mocked, abused, interrupted, or compressed into short viral scenes, yet again the sensing conditions, model boundaries, operator control, prompting logic, latency, and failure cases that structure the impression usually remain outside the frame. YouTube-based HRI studies show that such clips can shape perceived robot mind, morality, fear, sympathy, and uncanniness by making technical behaviour appear as socially meaningful conduct [Küster 21, Tolj 26]. TikTok robophobia develops this mediated framing through humour, insults, verbal abuse, and public robot encounters, where robots are presented as future social actors, labour competitors, relationship threats, or acceptable targets of ridicule [Rosén 26]. Mediated robot encounters therefore prepare viewers to assign humanoid robots a social character before the physical system acts, which makes prior narrative exposure relevant for studying how users interpret real HRI.

A common introductory claim in HRI research, that robots are rapidly entering everyday society, has become increasingly difficult to sustain for social robots and humanoids. Everyday contact with such platforms remains limited, especially in Europe, where social robots are more often encountered through laboratories, conferences, museums, demonstrations, and institutional pilots than through stable public or domestic infrastructure [Zawieska 24, Blair 26]. Public expectations therefore develop through research visibility, company promotion, online media, and public curiosity around rare encounters, while direct experience with sensor limits, dialogue breakdowns, maintenance demands, operator intervention, and situated failure remains scarce.

Ameca[3] made this gap between public expectation and practical contact physically visible at Robotics Research Laboratory (RRLab). The robot was purchased as a successor to the malfunctioning RoboThespian.[4] Ameca initially appeared to offer the platform often imagined in humanoid HRI. Its full-body embodiment, expressive facial actuation, gaze, and speech matched, and in some respects exceeded, the humanoid image circulated through science fiction films, online robot videos, and promotional media. Ameca also matches the cultural image of a humanoid as an apparatus with arms, legs, and a head, covered in metal or silicone-like material [Mara 22]. Its expressive design by Engineered Arts[5] makes it a highly visible platform for empirical HRI. The same expressive surface raises the standard for technical behaviour, since small mismatches can be read against the social competence promised by the robot body.

Repeated work with Ameca at RRLab first made this gap visible in university-based interaction studies. Participants responded to technical behaviour, but also to being looked at by a machine that appeared socially capable before its limits were clear. Curiosity, fascination, humour, and engagement appeared alongside discomfort, sarcasm, hesitation, and subtle avoidance, sometimes before the robot had spoken or been powered on. These encounters redirected the research from technical implementation on an impressive humanoid platform toward the expectation mismatch produced by Ameca's appearance, visible abilities, and technical limits.

[3] https://rrlab.cs.rptu.de/en/robots/ameca
[4] https://rrlab.cs.rptu.de/en/robots/robin
[5] https://engineeredarts.com/de/robots/ameca

The mismatch later became visible when Ameca was framed as a meaningful HRI system for older adults at the Senior Citizens' Fair 2025 [Jahn 26]. Figure 1.1 shows the public encounter, where the robot was positioned within care support after earlier university-based interaction studies. The demonstration followed RRLab proposal to examine whether Ameca could support elderly-care settings shaped by care demands and skilled-worker shortages in Germany. A local newspaper article about the demonstration[6] described Ameca as black-grey, metal-like, and alien-like, while also presenting it as a humanoid that could speak and laugh through ChatGPT. The demo showed the robot's limits, yet the deployed system used a local Ollama model without internet access, making the public description broader than the implemented architecture. The article shows how media accounts can compress a bounded technical demonstration into a stronger public image of capability.

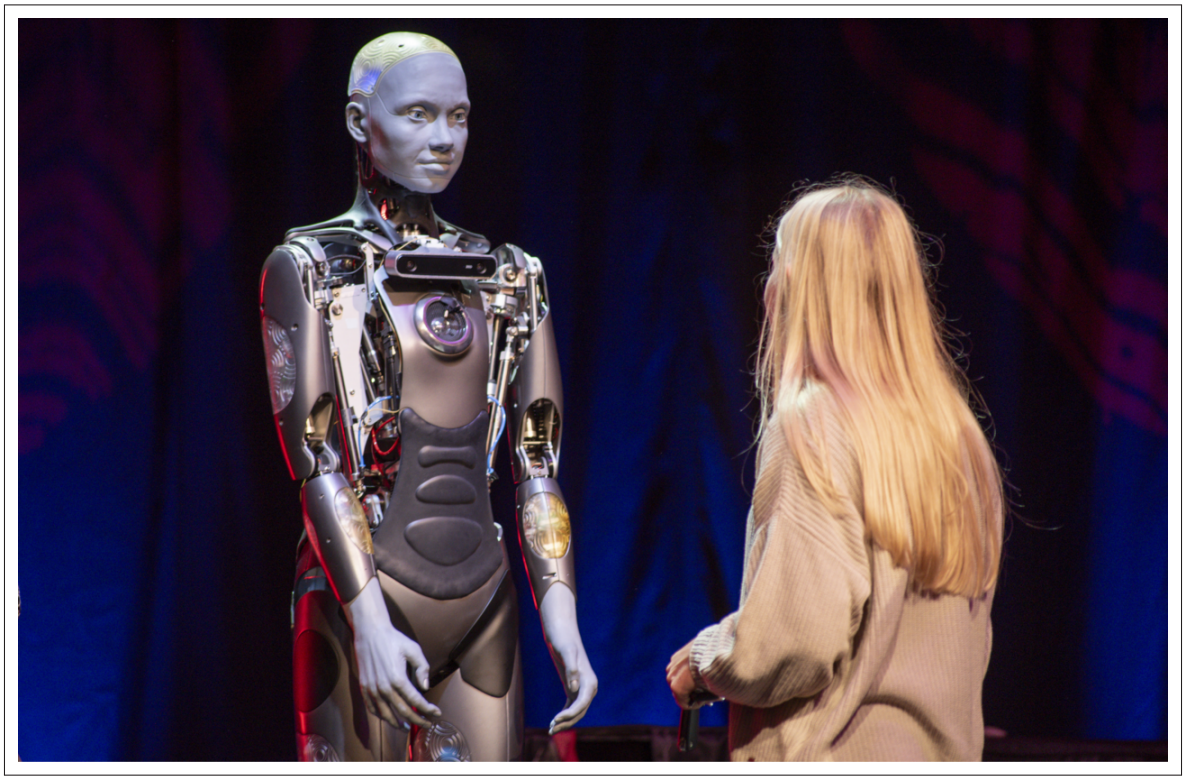


Figure 1.1: Public interaction with Ameca at the Elderly Convention 2025. The encounter shows how Ameca's adult-scale humanoid body, gaze, facial expression, and visible sensing hardware enter the interaction before the system's implemented capabilities and limitations are explained. Photo credit: Abdalla Mohamed

The motivation of this thesis comes from reducing the expectation gap between Ameca's social appearance, expected usefulness and capability, and the actual capabilities supported by its implemented architecture. Media-shaped expectations, rare everyday exposure, and public care-oriented framing make users interpret Ameca through social usefulness before first-hand interaction.

[6] Joachim Schwitalla, *Die Rheinpfalz*, 12 September 2025.

1.2 Research Scope and Questions

The research scope centres on reducing the expectation gap around Ameca as a life-sized humanoid robot. The motivation established that media-shaped expectations, rare everyday contact, and public care-oriented framing can make users interpret Ameca through an expected social role and assumed capability before first-hand interaction. The scope therefore follows how Ameca's social appearance, visible abilities, and implemented capabilities can be aligned within bounded HRI. Early work at RRLab examined Ameca with university students, including its role as a tandem language practice partner [Ashok 25a]. Later public work examined whether the same humanoid platform could be framed for older-adult care contexts [Jahn 25, Jahn 26]. The tandem language context becomes the main use-case in this thesis because it links social presence, generated dialogue, repeated exposure, and role boundaries to a concrete interaction need.

The research scope builds on expectation matching in HRI, where appearance, role, behaviour, and reliable capability must fit together during interaction [Rosén 24a]. It also builds on ability-based expectation work, where visible robot skill can shape perceived sociability, competence, and interaction quality while leading users to infer capabilities beyond what the system can reliably support [Horstmann 20a, Kwon 16]. Explanation-based expectation management provides the mitigation layer, where users need concrete information about robot abilities, limits, and breakdowns to interpret the interaction [Yadollahi 25].

The research problem is diagnosed through three dimensions of expectation mismatch in highly anthropomorphic HRI. Anthropomorphism sets the first expectation because Ameca's body is interpreted through human social categories. Apparent ability strengthens the expectation when perception, speech recognition, generated dialogue, gaze, lip synchronisation, facial movement, and gesture behaviour suggest listening, seeing, remembering, and responding. Anxiety or discomfort can appear when users cannot resolve the gap between these visible cues and the robot's reliable capabilities.

The proposed solution in this thesis is alignment and awareness. Alignment concerns the perceived fit between Ameca's appearance, role, behaviour, and reliable capability within a bounded task via exposure. Awareness concerns what users understand about sensing, generation, memory, inference, explanation, and breakdowns via explanations. Technical system design therefore becomes part of expectation management through task framing, capability explanations, dialogue boundaries, and repair responses.

The research is guided by three research questions:

RQ1: How does Ameca's anthropomorphic embodiment shape user expectations before and during interaction?

RQ2: Which technical limitations in perception, dialogue, gaze, and behaviour generation contribute to expectation mismatch in human–robot interaction?

RQ3: How can alignment through exposure, bounded task design, explanation, and robot-delivered capability awareness reduce uncertainty and discomfort in interactions with Ameca?

RQ1 examines anthropomorphic embodiment and expectation matching. RQ2 examines visible ability and technical capability inference. RQ3 examines expectation management through exposure, explanation, task framing, robot behaviour, and repair.

1.3 Approach and Contributions

To systematically approach the aforementioned research questions, the preceding diagnostic and mitigation dimensions are introduced as the 5A framework, a proposed analytical lens for connecting expectation diagnosis with design response in highly anthropomorphic HRI. The framework links three diagnostic dimensions, Anthropomorphism, Apparent Ability, and Anxiety, with two mitigation dimensions, Alignment and Awareness. On one hand, the 3As explain how Ameca's body and visible behaviours create expectations that become uncertain or uncomfortable when users cannot resolve what the robot perceives, understands, remembers, or can reliably do. On the other hand, the 2As define the design response by requiring Ameca's appearance, role, behaviour, and reliable capability to fit the bounded task. They also require users to understand the sensing, generation, memory, inference, and repair boundaries that shape the interaction.

The work combines empirical HRI studies, system development, and robot interaction design. The empirical contribution shows how full-body Ameca is perceived across mediated, laboratory, public, and repeated interaction contexts, with university students as the main user group. The technical contribution covers speech and dialogue pipelines, low-computation language-model architectures, text-to-speech integration, prompt-level behaviour boundaries, and visual perception. It extends to person-related perception, audio-visual social attention, multiparty gaze management, and robot behaviour generation through controlled facial, eye, arm, gesture, and attention movements during HRI. The design contribution develops a bounded account of meaningful HRI through Ameca as a tandem language practice partner. The conceptual contribution refines the social robot expectation-gap argument for highly anthropomorphic robots [Rosén 24a].

Across these contributions, the thesis shows that expectation mismatch in highly anthropomorphic HRI cannot be reduced by increasing human likeness alone. Additional human likeness can intensify mismatch when visible social cues exceed what perception, dialogue, memory, movement, repair, and transparency can support. Ameca becomes more acceptable when users can understand it as a bounded social robot with a clear interaction role and explicit capability limits. Meaningful interaction with highly anthropomorphic robots therefore depends on aligning apparent ability, actual capability, and interaction role.

1.4 Structure of the thesis

The thesis structure follows the four contribution layers introduced above. Chapters 4 and 8 provide the main empirical and design contributions. Chapters 5, 6, and 7 provide the technical system contributions. Chapter 3 provides the conceptual 5A framework that links these layers.

Each main chapter follows a shared structure that makes the contribution layer explicit. The chapter first positions its topic within the 5A framework, then uses a need-fulfilment diagram to connect the problem, user need, proposed solution, empirical evidence, and next step. A focused related-work section establishes the relevant HRI and system-design context. The core thesis works are then summarised in tables that state the motivation, method, result, and contribution of each study or implementation. Each chapter closes by explaining how its findings reduce, expose, or manage expectation mismatch. Detailed

methods, questionnaires, additional results, and study materials are provided in the corresponding appendices.

Chapter 2 introduces the HRI foundations needed for the thesis, including social robots, embodiment, generative AI, autonomy, Wizard-of-Oz methods, evaluation measures, and the robot platforms used in the work.

Chapter 3 develops the 5A framework as the conceptual structure of the thesis. It defines the anthropomorphic expectation gap and separates diagnostic dimensions from mitigation dimensions.

Chapter 4 examines robot embodiment, expectation, and uncanniness in Ameca. It connects appearance, emotion, movement, voice, encounter type, and expectation mismatch across perception studies and physical interaction contexts.

Chapter 5 presents the auditory perception and dialogue-management systems developed for Ameca. It covers speech-to-text, language-model architectures, local LLM integration, prompt boundaries, text-to-speech, and lip-synchronised speech.

Chapter 6 presents the visual perception and gaze-management pipeline. It covers visual percepts, visual question answering, speaker detection, human selection, and gaze behaviour for social interaction.

Chapter 7 presents robot behaviour generation and adaptation. It covers facial behaviour, eye behaviour, gaze generation, body gestures, and task-oriented deictic gestures.

Chapter 8 evaluates Ameca as a tandem language partner for university students. It tests whether a highly anthropomorphic robot can support bounded, low-pressure language practice when its role, dialogue, and limitations are made clear to users.

Chapter 9 discusses the thesis findings across embodiment, dialogue, perception, behaviour generation, and application.

Chapter 10 concludes the thesis by summarising the expectation-management argument and outlining implications for future HRI system design.

2. Human-Robot Interaction

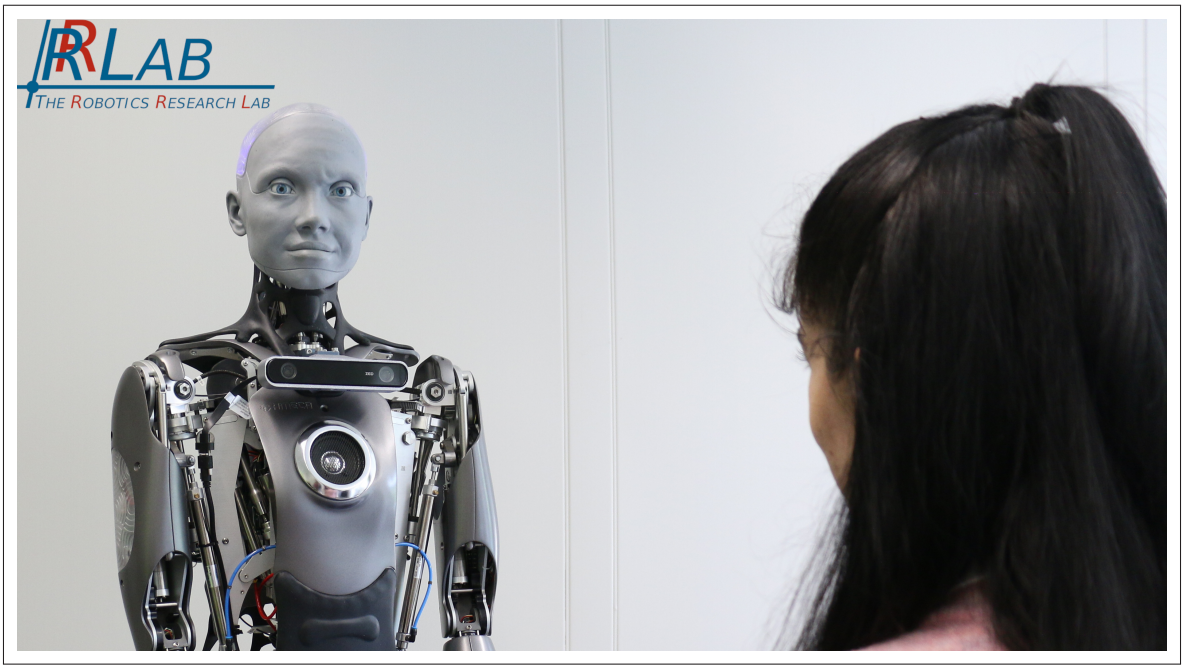


Figure 2.1: Embodied human–robot interaction with Ameca at the Robotics Research Lab. The human interlocutor faces the robot in shared physical space, where gaze, posture, distance, facial expression, and visible hardware shape how the interaction is interpreted.

Chapter 1 framed the thesis around an expectation gap between Ameca's social appearance, expected capability, and implemented architecture. The present chapter provides the HRI foundations needed to understand that gap. It introduces the human interlocutor, social robot embodiment, interaction modalities, control architectures, Generative Artificial Intelligence (GenAI)-based behaviour generation, and user-study evaluation as connected layers of an embodied robotic system. These concepts prepare the later chapters, where Ameca is examined through perception studies, implemented interaction pipelines, and bounded user-facing applications.

2.1 Foundations of Human-Robot Interaction

HRI studies how humans and robots perceive, act, communicate, and adapt during shared interaction. The field is interdisciplinary because human–robot interaction combines technical behaviour with human interpretation. It draws on robotics, Artificial Intelligence (AI), engineering, computer science, cognitive science, human factors, psychology, sociology, interaction design, and human-computer interaction [Ottoni 24, Dautenhahn 07].

From the robot perspective, the system is an embodied technical actor with sensors, actuators, control architecture, and some degree of autonomy or semi-autonomy. A robot can perceive selected parts of the environment, select actions, move, speak, gesture, or signal through other modalities [Fong 03, Bartneck 04]. These properties make robot behaviour physically situated. Robot action unfolds in shared space, takes time, and can change how the human positions, trusts, avoids, corrects, or addresses the system.

From the human perspective, the person is more than a user providing inputs. The term *human interlocutor* is used here for a person who speaks with the robot, interprets its behaviour, and is interpreted by the robot during interaction [Li 23c]. The term is useful here because Ameca studies involve gaze, speech, timing, facial expression, silence, repair, perceived intent, and task output as connected parts of interaction. Human interlocutors may become engaged, confused, reassured, frustrated, cautious, or trusting while the interaction unfolds [Li 23c, Henschel 21].

HRI is therefore studied here as a coupled interaction process between an embodied technical system and an interpreting human. It can involve different task types, autonomy levels, team structures, and information-exchange modes, ranging from one human with one robot to groups of humans and robots communicating through verbal and non-verbal behaviour [Ottoni 24]. The next section narrows this general field to social robots, where interpersonal behaviour becomes a design requirement.

2.2 Social Robots

Social robots form a specific area of HRI because their interaction with humans is designed around social behaviour. Social robots are evaluated through role, expression, timing, norm fit, and task performance because their interaction with humans extends beyond command execution, teleoperation, or physical assistance. Three definition excerpts clarify how social robots have been described in the HRI literature.

Excerpt 2.2.1: Socially Interactive Robots by Fong, Nourbakhsh, and Dautenhahn (2003, p. 145)

For the purposes of this paper, we use the term "socially interactive robots" to describe robots for which social interaction plays a key role.

Excerpt 2.2.2: Social Robots by Bartneck and Forlizzi (2004, p. 592)

A social robot is an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact.

Excerpt 2.2.3: Social Robots by Breazeal, Dautenhahn, and Kanda (2016, p. 1936)

Social robots are designed to interact with people in human-centric terms and to operate in human environments alongside people. Many social robots are humanoid or animal-like in form, although this does not have to be the case. A unifying characteristic is that social robots engage people in an interpersonal manner, communicating and coordinating their behavior with humans through verbal, nonverbal, or affective modalities.

Excerpt 2.2.1 separates socially interactive robots from other robotic systems by making social interaction central to the robot's purpose. Fong et al. connect this category to peer-to-peer interaction, emotion, high-level dialogue, natural cues such as gaze and gesture, personality, social relationships, and social learning [Fong 03]. Excerpt 2.2.2 adds a normative requirement, since the robot is expected to communicate according to behavioural norms recognised by the intended human interlocutors [Bartneck 04]. Excerpt 2.2.2 shifts the focus to human environments and multimodal coordination, where robots operate alongside people and communicate through verbal, non-verbal, and affective behaviour [Breazeal 16].

Across these excerpts, social robots can be understood as embodied, autonomous or semi-autonomous systems designed for socially readable, multimodal, and norm-sensitive interaction. From a predictive-coding perspective, social readability depends on learned links between appearance, movement, and action. Humans use prior experience to predict how an agent with a given body should behave, while robot models can use prediction-error minimisation to infer intention from observed action patterns [Saygin 12, Hwang 20]. Embodiment is therefore introduced here as an expectation-relevant property of social robots, since the robot's body, movement, voice, gaze, and expressive behaviour shape how human interlocutors interpret the robot's actions and expected interaction role.

2.3 Social Robot Embodiment

Social robot embodiments include mechanically abstracted humanoids, projected-face robots, zoomorphic or character-like platforms, screen-based conversational heads, highly humanlike humanoids, and androids [Fong 03, Breazeal 16, Mahdi 22]. These embodiment types shape which social cues a robot can display and which expectations a human interlocutor can form.

Pepper is a wheeled humanoid designed for public and service contexts through speech, tablet-mediated interaction, body language, and multimodal expressive gestures [Pandey 18]. Navel uses an abstracted humanoid body and expressive three-dimensional eyes to support intuitive non-verbal communication [Toussaint 23]. Miroka illustrates a character-like design that combines anthropomorphic structure with zoomorphic cues such as ears [Rossi 25]. RoboThespian and Furhat represent performance-oriented and conversational embodiments used in educational or dialogue settings [Verner 16, Mishra 22]. Humanoid platforms such as Sophia and Ameca move closer to humanlike appearance, while androids such as Nadine and ROMAN intensify the relevance of facial expression, gaze, voice, and role presentation [Fuchs 24, Kang 24, Berns 24].

The robot-selection logic in Figure 2.2 uses the MERPHI[1] poster as the closest external visual reference for locating Ameca among research, social, and entertainment humanoids [MERPHI 24]. The figure arranges Navel–QTrobot, Pepper–Miroka, RoboThespian–Furhat, ARI–EVE^{r3}, Sophia–Ameca, and Nadine–ROMAN across increasing anthropomorphic appearance. The arrangement shows how visible morphology changes the interaction cues available to users. Eyes can support gaze and attention, a mouth can support speech and expression, arms can support gesture, and a humanlike body can suggest socially coordinated behaviour.

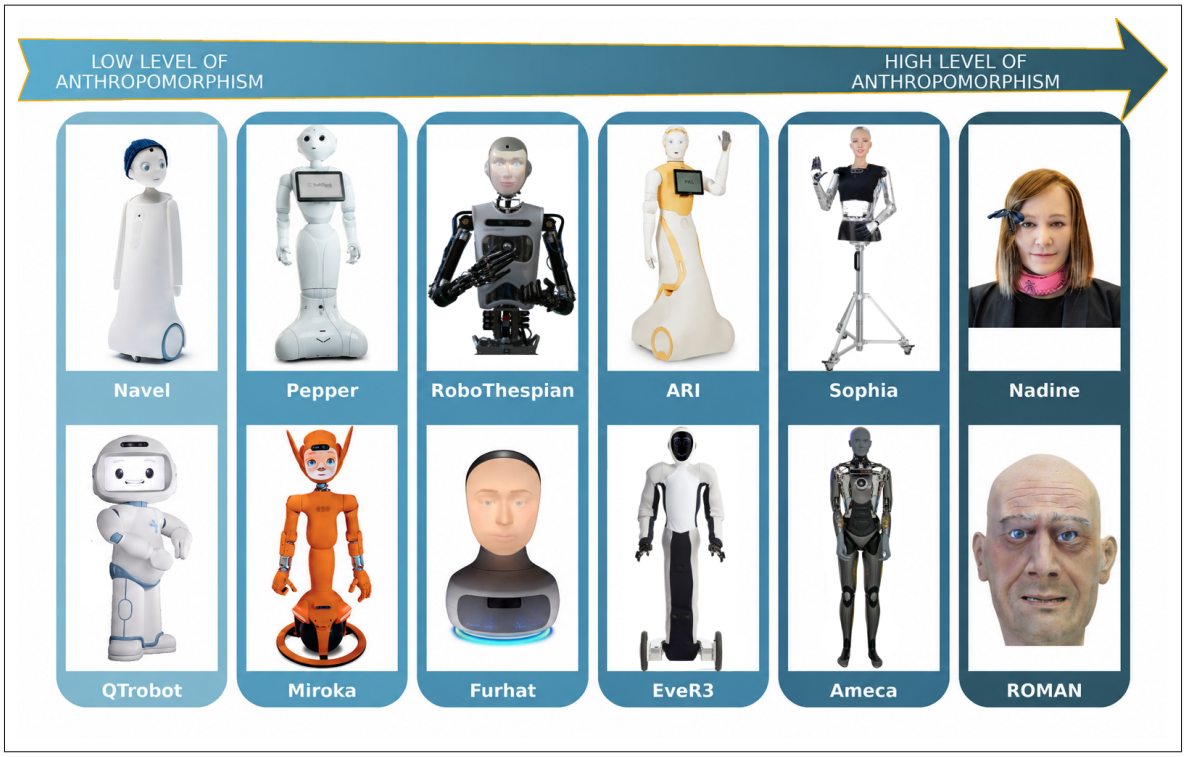


Figure 2.2: Social robot embodiments arranged across increasing anthropomorphic appearance. The examples move from abstracted and stylised platforms to projected-face, humanoid, and android-like embodiments, showing how visible morphology, facial detail, body structure, and humanlike expressivity progressively change the expectations attached to social interaction. Image outline generated with ChatGPT Images 2.0 by OpenAI, 2026 (https://openai.com/index/introducing-chatgpt-images-2-0/), based on author-provided and publicly available robot reference images listed below.

Image sources: Navel, https://navelrobotics.com/en/our-wheel-platform/; Pepper, https://us.softbankrobotics.com/pepper; QTrobot, https://luxai.com/assistive-tech-robot-for-special-needs-education/; Miroka, https://enchanted.tools/robot/; Furhat, https://www.furhatrobotics.com/; RoboThespian, https://rrlab.cs.rptu.de/en/robots/robin; ARI, https://pal-robotics.com/robot/ari/; EveR3, https://1x-tech.webflow.io/androids/eve; Sophia, https://www.hansonrobotics.com/robots; Ameca, https://rrlab.cs.rptu.de/en/robots/ameca; Nadine, https://nadineforgood.ch/; ROMAN, https://rrlab.cs.rptu.de/en/robots/more-robots/roman.

[1] https://www.merphi.se/

2.4 Interaction Modalities

Social robot embodiment becomes interactionally meaningful through multimodal behaviour. Voice, speech, gaze, facial expression, gesture, posture, proximity, timing, silence, and movement allow the robot body to be interpreted as socially responsive [Breazeal 16, Li 23c]. Voice carries linguistic content and paralinguistic information, including affect, pitch, rhythm, accent, personality, fillers, and perceived humanlikeness. These properties can shape trust, social presence, perceived competence, and the perceived fit between the robot's voice and body [Ashok 22, Seaborn 21]. Recent work on robot text-to-speech shows that voice design must fit the task, since storytelling and assistant-like interaction require different balances of expressivity, clarity, and timing [Tuttösí 25a, Tuttösí 25b].

Non-verbal behaviour regulates interaction as well as expression. Gaze helps organise turn-taking, joint attention, intimacy regulation, and referential grounding. Planning-based gaze control can therefore make robot attention more interpretable than purely reactive gaze behaviour [Mishra 22]. Facial expression supports emotional and pragmatic alignment when the expression fits the conversational context and timing. Delayed or inappropriate expressions can make an otherwise functional dialogue feel unnatural [Kirjavainen 25]. Gestures and body movement support pointing, explanation, imitation, and storytelling, especially in educational or public-facing robots such as RoboThespian and Pepper [Verner 16, Pandey 18]. Posture, proximity, and mobile orientation shape how the robot approaches, attends to, and positions itself toward human interlocutors, as shown in service and companion platforms such as ARI and Navel [Cooper 20, Toussaint 23].

Interaction modalities connect embodiment to control architecture. Each modality has to be implemented through sensing, scripting, generation, preprogramming, teleoperation, or online adaptation. Later technical chapters of this thesis return to these implementation choices through Ameca's auditory dialogue, visual perception, gaze management, and behaviour-generation pipelines.

2.5 Wizard-of-Oz, Hybrid, and Autonomous Control

Autonomy in HRI describes how these control choices are distributed across perception, dialogue, motion, expression, timing, and task behaviour. Social robots often combine autonomous modules with scripted behaviours, preprogrammed actions, generated dialogue, human supervision, and Wizard-of-Oz (WoZ) control. The required level of autonomy depends on the research question, safety constraints, technical reliability, interaction setting, and target behaviour [Maure 25].

WoZ methods remain important in HRI because they allow controlled social behaviours to be studied before robust automation is available. Depending on the embodiment and task, an HRI designer may control speech, body gestures, facial expression, dialogue, navigation, timing, or interaction onset. This allows an interaction concept to be evaluated while specific technical limitations are isolated [Maure 25, Nasir 22].

Hybrid control provides an implementation strategy for making social interaction reproducible, safe, and technically feasible. RoboThespian lessons used preprogrammed and teleoperated control to support group-based science learning, while Ameca patient

simulation used reproducible robot behaviour as an alternative training format in medical education [Verner 16, Schwarz 24]. Robust service-robot architectures follow a similar logic by monitoring unreliable modalities and shifting input across speech, tablet, phone, or external microphones [De Jong 18].

Recent systems extend this control layer with reinforcement learning for gaze, predictive coding for action-intention inference, and LLM-based architectures for memory, affect, and situated response generation [Ghamati 24, Hwang 20, Kang 24]. The control layer becomes especially important in GenAI-powered robots, where fluent dialogue, memory-like behaviour, affective responses, and situated action selection emerge from prompts, models, retrieval, sensors, safety filters, and robot modules.

Across the thesis work, control ranged from WoZ to hybrid and fully autonomous pipelines. In-person lab studies, perception video studies, and visitor demonstrations used WoZ or scripted flows to keep robot behaviour reproducible. The teaching-robot-as-tutee task used a hybrid setup, where behaviour learned from one participant group was tested with another group. The LLM- and Vision-Language Model (VLM)-based studies used fully autonomous pipelines, including video studies that depicted autonomously generated behaviour. These control choices prepare the next section, where GenAI architectures are treated as robot-control components.

2.6 Generative-AI Architectures

GenAI architectures extend the perception–dialogue–behaviour pipeline by turning LLMs and VLMs into robot-control components. In robotics, LLMs are used across communication, perception, planning, and control. They can interpret natural-language commands, generate dialogue, ground references, support scene understanding, produce task plans, and translate instructions into intermediate representations for robot modules [Kim 24c]. In HRI, this makes generative models part of the control layer, where robot responses must be coordinated with sensing, embodiment, timing, gaze, facial expression, gesture, and task execution through the dialogue manager. Dialogue management maintains context, interprets the current utterance, selects the timing and content of the next response, and coordinates dialogue with non-dialogue modules [Reimann 24]. GenAI gives HRI designers flexibility through response generation, clarification, adaptation, and interaction history, but it also reduces direct control over robot utterance selection.

LLMs in social robots should therefore be described as bounded architectural modules that can quickly enable full-pipeline interaction, while their opacity, incompleteness, and safety limits complicate control over robot utterances [Williams 24]. The rapid integration of generative models into social robots since early 2023 has produced reporting guidelines that require model specification, prompts, data flow, input–output handling, ethical risks, privacy implications, and the model's role in LLM-powered robots [Matuszek 26, Ashok 26]. GenAI also changes the expectation profile of social robots because fluent, role-consistent output can make ungrounded responses appear reliable, retrieval appear as humanlike remembering, and missing functions appear as available capabilities. LLM-powered care robots make this problem visible when a Pepper-based system falsely claimed medication-reminder functionality and created a deceptive capability impression in a high-stakes context [Ranisch 24]. GenAI architectures require explicit capability boundaries, grounding

checks, privacy handling, and repair responses before generated dialogue is coupled to embodied interaction.

These architectural boundaries have to be evaluated with human interlocutors, because embodied HRI depends on how people interpret, respond to, and adapt around the robot during interaction.

2.7 User Studies and Evaluation Samples

User studies in HRI include in-person laboratory studies, controlled public-space studies, in-the-wild observations, online surveys, interviews, questionnaires, behavioural logs, task outcomes, and physiological or kinematic measures. Evaluation in HRI must match what is being assessed, since technical module performance, user perception, and situated interaction each require different forms of evidence.

When HRI studies test hypotheses about human responses, the study design has to specify independent and dependent variables, estimate the required sample size through power analysis[2], recruit the corresponding participant sample, and analyse subjective, objective, or combined measures with appropriate statistical tests. This differs from many robotics and computer-science evaluations, where small samples can support pilot testing when technical performance, task completion, or application behaviour can be assessed through technical performance. In HRI, such evidence is necessary but incomplete when the research question concerns how human interlocutors perceive, interpret, trust, avoid, or adapt to the robot.

This methodological range is visible in public-space studies that combine field observation and interviews to explain why people comply with or ignore a robot's request [Schneider 22b]. Controlled live and video-based studies can measure interaction effects through behavioural traces such as motor responses, in addition to self-report [Vannucci 24]. Engagement studies combine self-reports, observational coding, behavioural cues, and automatic prediction because engagement changes during interaction [Sorrentino 24]. Scale-based evaluation is common in HRI, but scale choice through resources such as the "Perfect-scale" database[3] must still match the construct, items, reliability, validity, and study context [Saad 26].

HRI evaluation is broader than module benchmarking. Automatic metrics such as accuracy, F1 score, latency, task success, or classification performance are useful for evaluating components, including engagement detection and dialogue modules [Chung 24, Sorrentino 24]. Embodied HRI also requires user studies with coherent samples, because robot behaviour is interpreted through participant perception, response, and interaction context [Vannucci 24]. Participant selection therefore defines the scope of the conclusions that can be drawn from an HRI study [Seaborn 23]. Social-robot research separates populations by role and context, including children in daycare and education, older adults in assisted living and care settings, and companion-robot users in wellbeing, healthcare, rehabilitation, and socialisation domains [Rixen 26, Cruces 24, Ahmed 24]. University students are justified as a target population when the robot is deployed in a university setting and the interaction concerns student-facing use, support needs, or campus-based social situations [Rasouli 24].

[2] https://en.wikipedia.org/wiki/Power_(statistics)
[3] https://hriscaledatabase.psychology.gmu.edu/

The studies in this thesis use small pilot samples across several interaction settings, so their results should be interpreted as context-specific evidence rather than broad population-level claims. Within that scope, the evidence shows how human interlocutors form, revise, or retain expectations of social robots, which leads to the social robot expectation gap introduced next.

2.8 Social Robot Expectation Gap

Rosén defines an expectation as a belief about what will happen and separates it from belief, expectancy, anticipation, prediction, certainty, and assumption as related but distinct forms of future-oriented judgement. Belief provides the assumed truth from which expectations can be derived, expectancy describes the act of expecting, anticipation adds positive or negative emotion about a future state, prediction estimates a specific outcome, certainty estimates the likelihood of that outcome, and assumption treats something as true under uncertainty [Rosén 24a]. The distinction is useful for HRI because a participant may believe that highly anthropomorphic humanoid robots are advanced, anticipate an uncomfortable encounter, predict fluent dialogue, assume camera-based recognition, and then judge the interaction against those different expectations.

The social robot expectation gap refers to the misalignment between a user's expected capabilities and the robot's actual capabilities. Rosén describes the diagonal in Figure 2.6 as the ideal case, where expectations and robot capabilities are aligned, and the two surrounding spaces as disconfirmed expectations: high expectations when the robot cannot display the expected capability, and low expectations when the robot can do more than the user expected [Rosén 24a]. Disconfirmed expectations are relevant for interaction quality because they can produce negative affect, explicit cognitive processing, revised or retained expectations, and changed behaviour during interaction [Rosén 24a]. Rosén's empirical work applied this account with the Pepper robot, which is positioned two levels below Ameca in Figure 2.2. Extending this expectation-gap account to Ameca is therefore relevant for this thesis, since a more highly anthropomorphic humanoid may intensify the distance between visible social cues, expected capability, and implemented architecture.

2.9 EMAH-powered Ameca

EMAH system connects Ameca's humanoid embodiment to the real-time HRI architecture developed at RRLab[4] since 2022. The Generation 1 Ameca robot[5] functions as the embodied output layer for dialogue, expression, gaze, and response delivery.

The system was initially implemented as a Finroc-based C++ framework [Reichardt 12] and later redeveloped as a Python-based architecture to support reproducible integration of contemporary speech, vision, and GenAI components. The current technical architecture combines rule-based behaviour models with machine-learning and deep-learning components, while external sensing extends Ameca's default hardware through a push-to-talk microphone and a torso-mounted ZED2 RGB-D camera. These inputs support controlled

[4] https://rrlab.cs.rptu.de/en/robots/ameca
[5] https://wiki.engineeredarts.co.uk/Ameca

speech capture, person and scene perception, and embodied output through speech, lip synchronisation, facial expression, head motion, and gestures.

Ameca's physical design presents these interaction cues at human scale through a 187 cm, 49 kg humanoid body with 51 degrees of freedom, binocular eye cameras, binaural ear microphones, a chest speaker, face-detection-based eye tracking, and a highly expressive silicone face. These hardware and software layers support thesis studies on social perception, scene awareness, observational learning, and university-facing companion interaction. In these studies, sensing, embodiment, and response generation jointly shape how the human interlocutor interprets the robot. Figure 2.3 shows the sensing components used in the implemented systems.

Across the thesis, EMAH system supports both online video-based perception studies and small-scale in-person laboratory studies. Evaluation combines self-report questionnaires, open-ended responses, interaction transcripts, and technical inspection of generated robot behaviour. This implementation context links the general HRI foundations in this chapter to the Ameca platform used in the remaining thesis.

The chapter has introduced the HRI concepts needed to read the later Ameca studies: social robot embodiment, multimodal interaction, control architecture, GenAI-based behaviour generation, and evaluation design. Additionally, the above set the foundations for Chapter 3, where the expectation gap is organised through anthropomorphism, apparent ability, anxiety, alignment, and awareness.

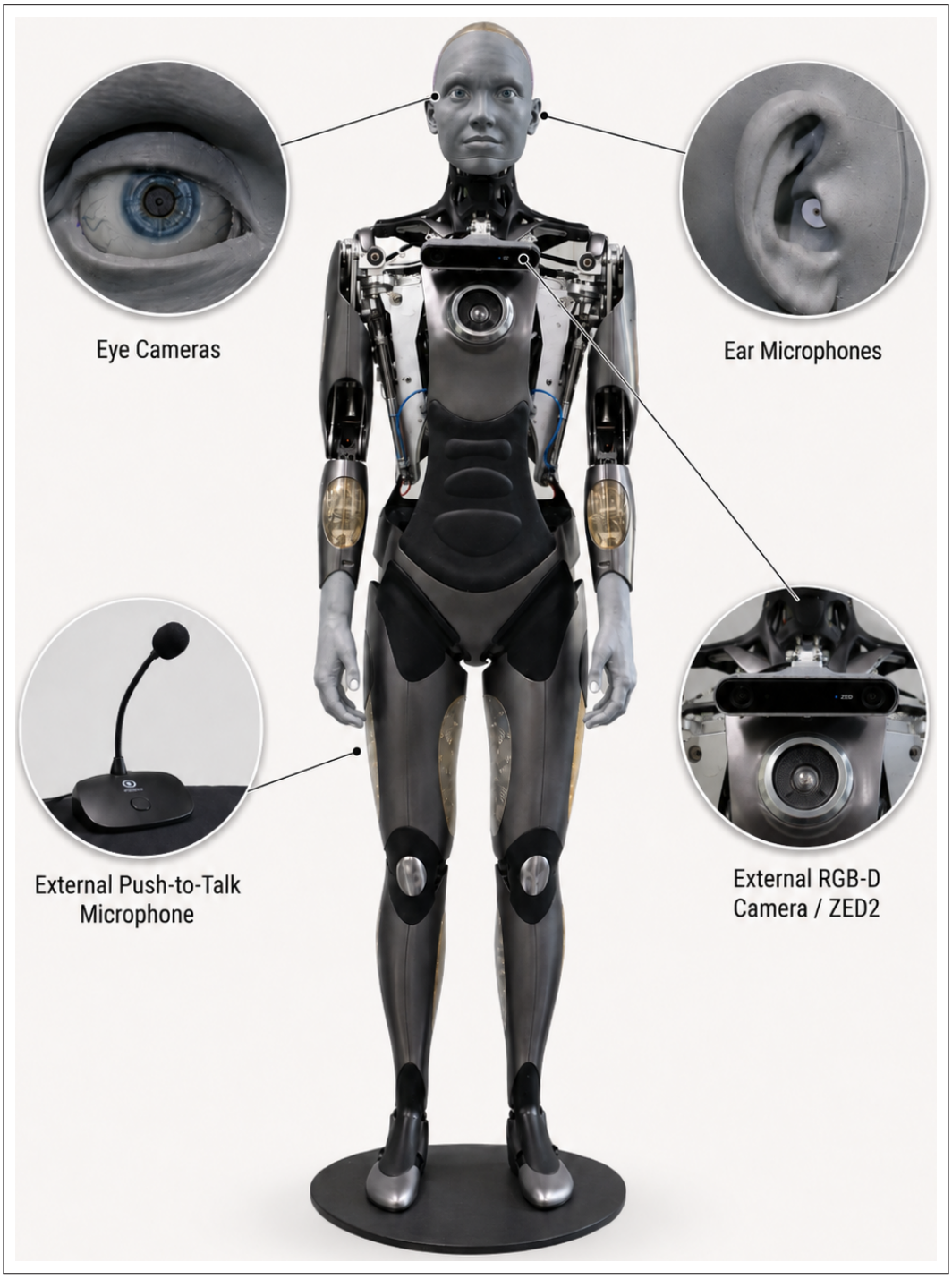


Figure 2.3: Ameca sensing components used in this thesis. The robot combines built-in eye cameras and ear microphones with external sensors added at RRLab, including a push-to-talk microphone and an RGB-D camera/ZED2 mounted on the torso. These components support speech input, visual perception, gaze management, and interaction control in the implemented HRI systems. Image generated with ChatGPT Images 2.0 by OpenAI, 2026 (https://openai.com/index/introducing-chatgpt-images-2-0/), based on author-provided photographs of the Ameca robot and its sensors.

3. The 5A Framework

Chapter 2 introduced the HRI concepts needed to analyse Ameca as an embodied social robot, including embodiment, modality, control, GenAI-based behaviour generation, evaluation, and the social robot expectation gap. Building on this foundation, the present chapter develops the *5A framework* as the conceptual structure for the thesis problem. The framework extends Rosén's social robot expectation gap toward highly anthropomorphic robots such as Ameca, where the body, visible behaviour, and implemented architecture jointly shape what interlocutors expect the system to do [Rosén 24a]. As an organising lens across the empirical and technical chapters, it links robot appearance, apparent ability, interlocutor response, design mitigation, and system transparency.

The anthropomorphic expectation gap describes the mismatch between what a humanlike robot appears able to support and what its implemented architecture can reliably deliver. Rosén defines the social robot expectation gap as a misalignment between users' expectations and a robot's actual capabilities, often shaped by social robot design and presentation that invite humanlike or social interpretation [Rosén 24a]. For Ameca, the expectation gap combines architecture and interpretation because interlocutors respond to a social body whose behaviour is produced through bounded sensors, models, scripts, prompts, actuators, latency, and repair limits.

The 5A framework separates the anthropomorphic expectation gap into three diagnostic dimensions and two mitigation dimensions. *Anthropomorphism* refers to the humanlike form that creates the initial social expectation. *Apparent Ability* refers to visible interaction behaviours that make this expectation appear technically and socially plausible. *Anxiety* refers to discomfort, mistrust, uncertainty, fear, or uncanny impression when the interlocutor cannot resolve what the robot perceives, understands, remembers, intends, or can be trusted to do. *Alignment* and *Awareness* provide the mitigation dimensions. Alignment concerns the fit between appearance, behaviour, role, and reliable capability within a bounded task. Awareness concerns the interlocutor's understanding of the system boundaries behind the robot's behaviour. Table 3.1 defines the five terms for the analyses in this thesis.

The table defines the five terms separately, while Figure 3.1 shows how they operate as one expectation-management structure. The figure extends Rosén's social robot expectation-

Table 3.1: Definitions of the 5A framework for anthropomorphic social robots.

5A dimension	Definition in this thesis	Relevant robot cues
Anthropomorphism	Humanlike form that leads the interlocutor to interpret the robot through human social categories before interaction performance is known.	Face, eyes, eyelids, mouth, skin, body scale, posture, voice source, start-up motion, clothing, and backstory.
Apparent Ability	Visible or described competence that makes the robot's humanlike form appear capable of social, perceptual, cognitive, or emotional interaction.	Speech, gaze, facial expression, gesture, emotion display, turn-taking, memory-like responses, person recognition, LLM dialogue, VLM perception, and task execution.
Anxiety	Interlocutor uncertainty, discomfort, mistrust, fear, or uncanny response when expected capability cannot be resolved through observed robot behaviour.	Response delay, silence, misrecognition, inconsistent gaze, privacy ambiguity, role ambiguity, appearance–motion mismatch, unexpected fluency, and unpredictable breakdowns.
Alignment	Fit between the robot's appearance, assigned role, visible behaviour, and reliable capability within a bounded interaction task.	Bounded tasks, role framing, physical exposure, gaze–speech timing, repair behaviour, task-specific interaction, and visible system limits.
Awareness	Interlocutor understanding of what the robot can sense, generate, remember, infer, disclose, and fail to do during interaction.	Self-introduction, consent information, capability explanation, privacy statements, memory disclosure, sensing disclosure, interaction boundaries, and repair responses.

gap figure by adding the 5A diagnostic and mitigation layers above the expected-capability and actual-capability relation [Rosén 24a]. Anthropomorphism and apparent ability increase inferred capability when the robot's body and behaviour suggest social, perceptual, cognitive, or emotional competence before reliable performance is established. Anxiety marks the unresolved space in which interlocutors cannot stabilise what the robot perceives, understands, remembers, or can be trusted to do. Alignment and awareness reduce the gap by giving interlocutors bounded exposure to the robot's role, behaviour, and limits, and by making the relevant system boundaries explicit.

The uncanny valley provides the design background for the expectation–capability relation in highly humanlike robots. Mori's hypothesis describes a non-monotonic relation between human likeness and affinity, where near-human appearance can produce eeriness when the entity appears humanlike while remaining perceptually or behaviourally mismatched [Mori 12]. Mori also treats movement as a factor that can steepen the valley, since facial motion, bodily timing, and hand movement can intensify affinity or eeriness depending on their fit with human expectations [Mori 12]. Rausch's discussion connects the uncanny valley to trust, service roles, bounded functionality, and over-extension in public robot deployment, where robots became more acceptable when role and capability

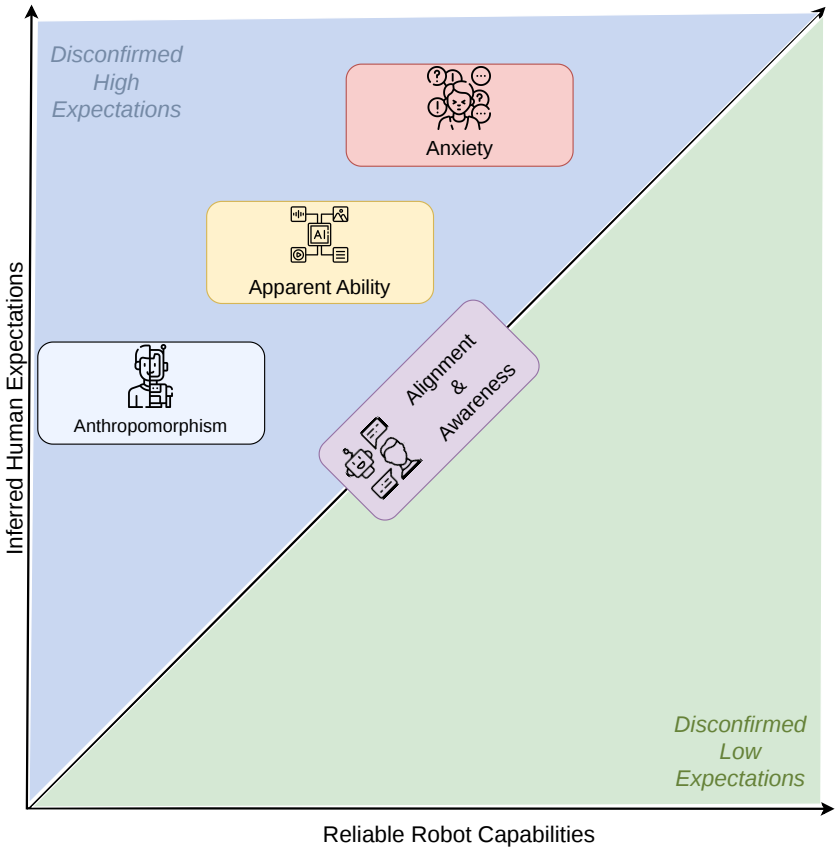


Figure 3.1: The 5A framework for diagnosing and mitigating anthropomorphic expectation gaps in highly humanlike social robots.

limits were narrow enough to understand [Rausch 21]. In the 5A framework, uncanniness is therefore one possible response to unresolved expectation, rather than a fixed property of Ameca.

Figure 3.2 places Ameca in relation to other humanoid and android robots, and its placement should be read as relative rather than absolute. Ameca's full-body appearance, exposed mechanics, non-human colour–material–finish design, and distributed degrees of freedom can keep the whole robot outside the strongest uncanny-valley region. The visible industrial body prevents a complete human illusion. An isolated head view[1], a different comparison set, or a focus on facial realism alone could move perception toward uncanniness, since the same robot can be judged differently when body, motion, and mechanical context are removed[2]. The relative placement supports analysing Ameca through the relation between visible social expectation and reliable system capability. That relation becomes concrete in the *Ameca encounter*, where a life-sized humanlike robot is physically encountered before its system boundaries are understood. The encounter can range from astonishment, curiosity, interest, and awe to discomfort, uncertainty, and fear, because interlocutors must infer what kind of entity the robot is while its perception, dialogue, memory, movement, and autonomy remain technically constrained.

[1] https://engineeredarts.com/robots/ameca-desktop
[2] This interpretation follows author correspondence with Mehrdad H. M. Farimani

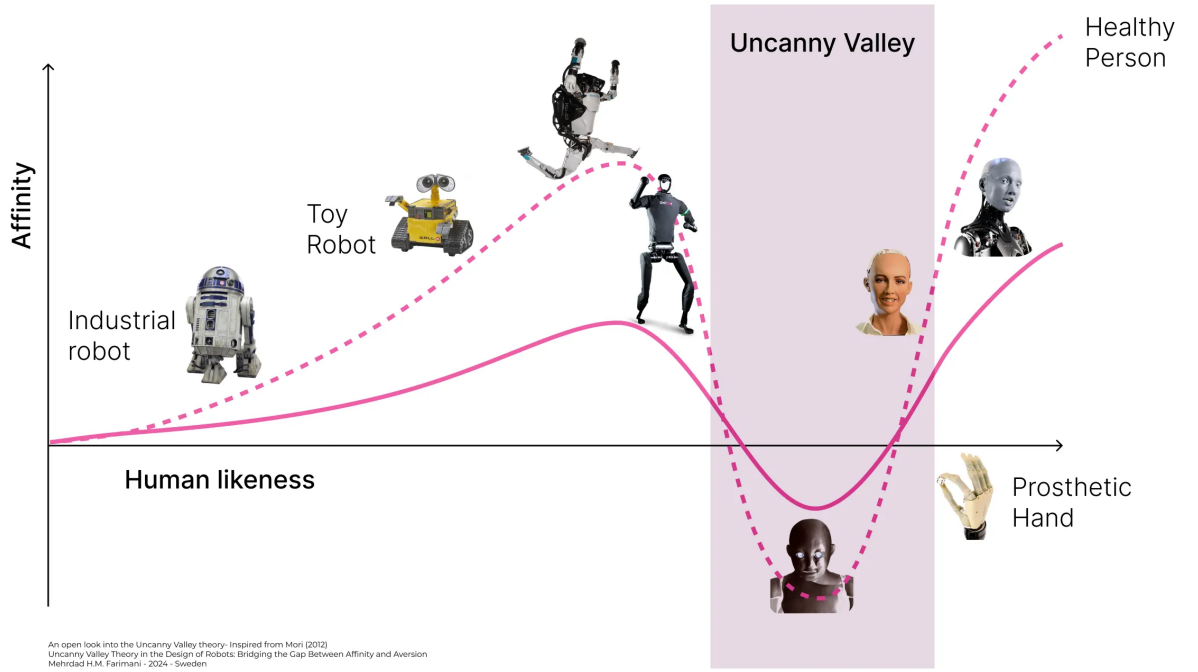


Figure 3.2: Uncanny-valley placement of selected robots across human likeness and affinity. Used with permission from Farimani [Farimani 24]; the visual logic follows Mori's uncanny-valley hypothesis [Mori 12].

The *Ameca encounter* marks the perceptual entry point of the anthropomorphic expectation gap because interlocutors first encounter a social body and only later infer the technical architecture that produces its behaviour. From a predictive-coding perspective, discomfort can emerge when gaze, facial expression, voice, silence, latency, or generated dialogue lead interlocutors to anticipate perception, understanding, or agency beyond what the system can sustain [Saygin 12, Hwang 20]. The same Ameca interaction can therefore appear fascinating, acceptable, uncomfortable, or frightening, depending on how interlocutors stabilise the relation between humanlike form and bounded machine capability.

The Ameca studies ground the expectation-based reading in empirical evidence. In the Ameca uncanniness study, participants compared ROMAN[3], ROBIN[4], and Ameca[5] through humanlikeness, familiarity, eeriness, emotional expression, and interaction movement. Their judgements changed after exposure to robot expressions and interaction videos [Berns 24]. The study indicates that Ameca is not evaluated through appearance alone, since participants can combine body form, expression, movement, timing, and interaction skill into one assumed social capability.

The design challenge is to keep assumed capability close to the capabilities that Ameca can reliably support. Higher human likeness intensifies the expectation constraint when implemented sensing, dialogue, memory, movement, and repair cannot meet the capability anticipated by interlocutors. Ameca can support companion-like interaction with university students when its role, generated dialogue, perception, memory, privacy limits, and repair

[3] https://rrlab.cs.rptu.de/en/robots/more-robots/roman
[4] https://rrlab.cs.rptu.de/en/robots/robin
[5] https://rrlab.cs.rptu.de/en/robots/ameca

behaviour remain accountable to the interlocutor. Expectation management becomes an implementation problem because every added capability changes the apparent ability that the robot presents during interaction. The following sections develop the diagnostic dimensions before defining how alignment and awareness reduce expectation mismatch.

3.1 Problem Statement and the 3As

The 5A framework begins with three diagnostic dimensions that explain how expectation mismatch emerges in highly anthropomorphic interaction: *Anthropomorphism*, *Apparent Ability*, and *Anxiety*. Anthropomorphism sets the initial expectation through the robot body, Apparent Ability makes that expectation believable through visible behaviour, and Anxiety emerges when interlocutors cannot resolve what the robot perceives, understands, remembers, intends, or can be trusted to do. These dimensions diagnose how a highly humanlike robot can create an expectation gap before any single module fails.

Across the Ameca studies, appearance, emotion, movement, dialogue, embodiment order, memory-like behaviour, and role framing shaped how interlocutors interpreted the robot [Berns 24]. Table 3.2 maps this evidence onto the three diagnostic dimensions used in this thesis.

Table 3.2: Ameca evidence mapped onto the 3As of the anthropomorphic expectation gap.

Dimension	Expectation-gap problem	Ameca evidence & design implication
Anthropomorphism	Humanlike form sets social expectations before system performance is known.	Appearance, emotion, movement, and interaction skill were interpreted together in the Ameca uncanniness study [Berns 24]. Clothing and backstory studies show that surface design and narrative framing can also shape how interlocutors approach a robot [Ashok 23, Ashok 25b]. *Implication:* form functions as an expectation-setting interface.
Apparent Ability	Visible partial competence can lead interlocutors to anticipate broader social understanding.	Speech emotion, gesture mimicry, person recognition, LLM dialogue, multi-session continuity, and language-practice interaction made separate technical functions appear as social competence when performed by Ameca [Ashok 24b, Lu 24, Mauliana 25]. *Implication:* interlocutors need boundaries for what is generated, detected, remembered, inferred, or simulated.
Anxiety	Unclear capability limits make the robot difficult to place and trust.	Ameca studies show curiosity and unease in university settings, expectation drops after virtual-first exposure, and ambivalent responses to physical presence and gaze [Berns 24, Majumder 26, Jahn 26]. *Implication:* expectation management requires bounded exposure, transparent sensing, role limits, and failure explanations.

3.1.1 Anthropomorphism

Anthropomorphism defines the first diagnostic dimension of the expectation gap because Ameca is interpreted before it performs. Its face, eyes, eyelids, mouth, torso, posture, skin, voice source, and breathing-like idle motion invite interlocutors to use human social categories even when they know that the system is artificial [Złotowski 15, Lee 03, Horstmann 19]. The body therefore sets an initial standard for later behaviour: speech is expected to be conversational, gaze is expected to be attentive, facial motion is expected to be meaningful, and silence is expected to have social relevance.

Ameca intensifies the anthropomorphic expectation gap because its embodiment differs from a small tabletop device, a screen-based avatar, or a low-humanlikeness robot. The MERPHI[6] robot poster places Ameca near the highly humanlike end of the robot appearance spectrum. Ameca's life-sized humanoid body, expressive face, visible facial actuation, non-functional legs, and physical presence place the interlocutor in a face-to-face encounter before the technical boundaries of the system are known. Appearance research shows that morphology, facial features, material, texture, clothing, customization, and body form shape how interlocutors assign robot identity, task fit, affordance, and social meaning [Pimminger 25, Paananen 22, Wang 22]. In Ameca, these cues make the robot socially addressable before the interaction task begins.

Humanlike appearance also changes what interlocutors expect the robot to explain, perceive, and understand. Kopecka and Such showed that more humanlike domestic service robots were more often connected to mental-capacity explanations, while Pang et al. found that a more humanlike android shaped interlocutor speech patterns in real conference interactions [Kopecka 25, Pang 25]. These findings support the Ameca encounter because the robot is judged against the capability anticipated from its form.

Physical embodiment strengthens the expectation set by Ameca's appearance. A physically present humanoid can affect personal space, compliance, and social response differently from a video-displayed robot because the interlocutor shares space with the body instead of observing an image [Bainbridge 08, Oh 18]. Social presence also depends on movement, modality, behavioural realism, proximity, voice, interactivity, and perceived communicative ability [Sasser 24, Chen 23]. Ameca's 187 cm body, blue-eyed gaze, and interactional timing can therefore make the encounter feel socially consequential before the interlocutor has evidence of the robot's actual capability. Kwon et al. showed that a small number of social cues can lead interlocutors to infer more capability than a robot has, while Rosén et al. describe the social robot expectation gap as a mismatch between expected and experienced robot behaviour [Kwon 16, Rosén 24a]. In the Ameca studies, embodiment therefore functions as the first diagnostic source of expectation mismatch, as detailed in Chapter 4.

3.1.2 Apparent Ability

Apparent Ability defines the second diagnostic dimension of the expectation gap because visible behaviour can make the anthropomorphic expectation set by the body seem technically supported. When Ameca appears to listen, see, think, and respond, interlocutors

[6] https://www.merphi.se/

may treat separate implemented functions as evidence of broader social understanding. Horstmann and Kraemer showed that low interaction skill reduced evaluations of sociability, competence, and interaction quality, while expectation violations led to less positive evaluations of the robot and interaction [Horstmann 20a, Horstmann 20b]. For Ameca, each added visible ability changes the standard against which the next behaviour is judged.

The resulting difficulty is a capability-inference gap. Kwon et al. describe the expectation gap as the difference between the few abilities a robot displays and the broader abilities interlocutors infer from those displays [Kwon 16]. Once a robot shows speech, gaze, turn-taking, emotion, or memory-like behaviour, interlocutors may treat the behaviours as signs of one coherent social mind instead of separate modules [Horstmann 20a]. In Ameca, a fluent answer can make later silence feel like refusal, a precise gaze shift can make later misrecognition harder to excuse, and a remembered detail can make memory limits harder to understand.

Non-verbal behaviour can further strengthen apparent ability because humanoid cues are often processed as socially meaningful. Direct gaze can become meaningful when a humanoid is introduced as intentional, and communicative robot behaviour can be judged through perceived origin, deliberateness, reference, genuineness, and clarity [Linnunsalo 26, Holthaus 23]. In Ameca, eye motion, facial actuation, actuator noise, delay, and silence can therefore be experienced by interlocutors as social responses as well as technical events.

GenAI sharpens the apparent-ability problem because fluency can be mistaken for understanding. LLM-based HRI can support full-pipeline interaction, while the resulting systems can remain opaque, incomplete, and difficult to control [Williams 24, Matuszek 26]. Ranisch et al. exemplify the risk through an LLM-enhanced Pepper care system that falsely claimed medication-reminder functionality that it could not actually provide [Ranisch 24]. For Ameca, generated dialogue, VLM perception, and memory-like behaviour support interaction only when interlocutors can distinguish what is sensed, stored, generated, inferred, guessed, or unavailable.

The ethical concern raised by apparent robot ability is capability deception, including cases where no deliberate deception is intended. Ciardo et al. frame deception as a mismatch between ordinary and scientific mental models of robots, while Sharkey and Sharkey argue that harmful false beliefs can emerge from robot design without deliberate intent [Ciardo 26, Sharkey 21]. Deception studies show that misleading robot ability can affect trust, evaluation, and general attitudes toward robots [Rossi 24, Saadon 25]. Ameca encounters therefore require transparent ability boundaries that keep visible behaviour close to implemented and reliable capability.

In this thesis, Ability is treated as Apparent Ability when visible behaviour leads interlocutors to infer capability beyond the implemented architecture. Chapters 5, 6, and 7 address that diagnostic risk by aligning Ameca's visible abilities with what the system can reliably support.

3.1.3 Anxiety

Anxiety defines the third diagnostic dimension of the expectation gap because unresolved anthropomorphism and Apparent Ability can make Ameca difficult for interlocutors to place. A highly humanlike robot first invites social interpretation through its body and then

strengthens that interpretation through visible behaviour. When the expected capability created by those cues does not match the actual capability available in the interaction, interlocutors may experience uncertainty, discomfort, or fear.

Robot anxiety has been measured through instruments such as Negative Attitudes towards Robots Scale (NARS) and Robot Anxiety Scale (RAS), but attitudes, acceptance, trust, and anxiety toward social robots vary with exposure type, robot design, application domain, and interlocutor characteristics [Nomura 06a, Naneva 20]. An interlocutor can therefore approach Ameca with curiosity and still experience discomfort when appearance, apparent ability, and system status do not align. Anxiety, in this diagnostic sense, does not mean a stable negative attitude toward robots; it marks the affective consequence of not knowing what the robot is, what it can do, and how much agency should be attributed to it.

Expectations can be built from direct experience, indirect experience, and inference, and their disconfirmation can affect cognition, behaviour, performance, and affect [Rosén 22]. Prior robot experience can shape whether interlocutors evaluate social behaviour positively or report more negative attitudes and robot anxiety [Nikolovska 24, Rosén 24b]. During Ameca encounters, some interlocutors bring HRI experience, some bring fictional references, some bring online videos, and some bring only the surprise of a life-sized humanoid looking back at them. These different sources of expectation shape how quickly interlocutors can adjust from imagined capability to observed capability.

Indirect media exposure can raise expected robot skills, and negative fictional robots can increase threat expectations [Horstmann 19]. Online robot videos can shape perceptions of mind, morality, and fear, while unexpected abilities may affect mind perception more strongly than appearance alone [Küster 21]. Public robot discourse also carries fear scripts through robophobia, robot-abuse videos, and science-fiction associations [Rosén 26, Tolj 26]. A short Ameca encounter therefore has to compete with expectations formed outside the laboratory, exhibition, classroom, or hallway where the interaction occurs.

Limited everyday contact with social robots makes the expectation gap harder to correct during a first encounter. Social robots remain largely absent from everyday physical and social environments, especially in Europe, despite repeated claims that they are approaching daily use [Zawieska 24, Blair 26]. During the writing of this thesis, publicly visible full-body Ameca installations in Europe appeared limited to a small number of institutional and exhibition contexts, including OFFIS[7], the Deutsches Museum Bonn[8], WAIYS[9], the Heinz Nixdorf MuseumsForum[10], and the National Robotarium[11]. A first *Ameca encounter* therefore asks interlocutors to revise imagined capability into observed capability quickly enough for the interaction to remain understandable.

Anxiety around Ameca can also attach to concerns about AI. Perceived AI control can increase fear across several dimensions [Zhan 24]. Social robots can raise concerns about

[7] https://www.offis.de/offis/aktuelles/meldung/ein-digitales-gegenueber-was-ein-roboter-im-seniorenheim-bewirken-kann.html

[8] https://www.deutsches-museum.de/bonn/programm/veranstaltung/sprechstunde-mit-ameca

[9] https://www.waiys.com/ameca

[10] https://www.hnf.de/en/permanent-exhibition/exhibition-areas/everything-goes-digital/man-robots-living-with-artificial-intelligence-and-robotics/translate-to-english-ameca.html

[11] https://thenationalrobotarium.com/worlds-most-advanced-humanoid-robot-arrives-at-the-national-robotarium/

constant watching, constant listening, data inference, unauthorized sharing, and physical safety [Bell 25]. For Ameca, apparent intelligence and unclear ability boundaries can turn ordinary interaction cues into signals of surveillance, autonomy, or uncontrolled data use.

The Ameca studies provide context-specific evidence for anxiety as uncertainty rather than broad population-level fear. In the hallway and uncanniness study, participants showed varying mixtures of engagement, curiosity, indifference, unease, and fear instead of a single positive or negative response [Berns 24]. In the embodiment-order study, virtual-first exposure was followed by stronger expectation drops and increased anxiety during later physical interaction, while physical-first exposure produced more stable expectations [Majumder 26]. In a public-space in-the-wild study with older adults, Ameca's physical presence and gaze were described as both fascinating and unsettling [Jahn 26]. These findings indicate that Ameca can remain attractive and difficult to place at the same time.

Anxiety in this thesis is therefore treated as a response to unresolved capability uncertainty. Across four years of working with Ameca, strong fear was observed only once, when a 12-year-old participant refused to enter the university robot room and remained outside crying while holding her teacher. That case marks the extreme end of the Ameca encounter, while most responses were less absolute and ranged from curiosity to hesitation. The diagnostic issue is therefore not that Ameca generally frightens interlocutors, but that its humanlike body and apparent ability can make its actual limits difficult to judge quickly.

Anthropomorphism, Apparent Ability, and Anxiety describe how Ameca becomes socially readable, technically over-interpreted, and emotionally difficult to place. The unresolved anthropomorphic expectation gap motivates the mitigation contribution of the 5A framework as an organising analytical lens for aligning what Ameca appears to offer with what its implemented architecture can reliably support.

3.2 Mitigation: The 2As

The mitigation dimensions of the 5A framework are *Alignment* and *Awareness*. Alignment reduces the anthropomorphic expectation gap by bringing Ameca's appearance, behaviour, role, and reliable capability closer together during interaction. Awareness reduces the same gap by making the system boundaries behind Ameca's behaviour explicit, so that interlocutors can distinguish sensed, generated, scripted, inferred, stored, and failed behaviour. Alignment therefore works through bounded exposure to what Ameca can actually do, while Awareness works through repeated explanation of how interlocutors should interpret what Ameca appears to do.

The solution is not to remove anthropomorphism or suppress apparent ability. Ameca's body and interaction capabilities are central to the studies in this thesis. The mitigation task is to keep interlocutor interpretation close to implemented and reliable capability. Alignment supports that task when exposure is organised around a concrete role, task, and interaction setting. Awareness supports the same task when explanations are attached to the interaction instead of delivered once as a generic disclaimer.

Table 3.3 maps the mitigation logic onto the Ameca studies. The table shows how the 2As respond to the diagnostic dimensions introduced above: Alignment addresses expectation mismatch through controlled contact with Ameca's real body and limits, and Awareness

addresses capability uncertainty through explanation of sensing, generation, memory, inference, and failure.

Table 3.3: Ameca evidence mapped onto the 2As for mitigating the anthropomorphic expectation gap.

Mitigation	Expectation-gap management	Ameca evidence & design implication
Alignment via exposure	Interlocutors compare Ameca's apparent ability with its body, timing, task behaviour, and role limits.	Virtual-first exposure was followed by stronger expectation drops and anxiety during later physical interaction [Majumder 26]. Teachable-robot and CopyCat studies made learning limits and gesture-recognition limits visible through teaching, correction, and feedback [Majumder 25b, Ashok 24b]. *Implication:* exposure should be bounded, task-based, and connected to Ameca's actual role.
Awareness via explanation	Interlocutors are told what Ameca is, what it can sense, what it can remember, what uses AI, and where its limits are.	Capability-based self-introduction made Ameca's height, sensors, degrees of freedom, non-walking legs, and system limits easier to interpret [Majumder 26]. LLM dialogue, multi-session interaction, and RoboReID show why explanation is needed, since fluency, continuity, memory-like behaviour, and person recognition can create stronger capability impressions [Ashok 24a, Mauliana 25, Lu 24]. *Implication:* interlocutors should know what is generated, detected, stored, inferred, guessed, or simulated.

3.2.1 Alignment via Exposure

Alignment via exposure means that interlocutors encounter Ameca's actual body, timing, role, and task limits instead of relying on an imagined version of the robot. The real interaction includes the 187 cm-tall body, fixed base, gaze behaviour, facial motion, actuator noise, lip synchronisation, response timing, speech-recognition errors, and task-specific behaviour. These features matter because they turn abstract expectations into observable interaction constraints.

Physical exposure alone does not automatically solve the expectation gap, strengthen social influence, or make robot behaviour easier to interpret [Li 15, Thellman 16]. The Ameca embodiment-order study showed that the same scripted monologue, delivered by the robot in first-person speech, produced different expectation and anxiety patterns depending on whether interlocutors first met Ameca as a virtual video-based form or as a physical robot [Majumder 26]. Prior robot experience therefore matters because it shapes how interlocutors evaluate the same behaviour and how quickly their expectations align with observed capability [Nikolovska 24, Rosén 24b].

Task-based exposure strengthens alignment because interlocutors can see the robot's competence and limits inside the activity itself. In the CopyCat study, interlocutors

corrected and reacted to Ameca's uneven observational-learning-based gesture performance, which made recognition and imitation limits part of the task [Ashok 24b]. In the Ameca tutee study, active teaching reduced expectation shifts but increased robot-related anxiety, showing that calibration can also make limitations more noticeable [Majumder 25b]. Alignment therefore does not mean making the robot appear better; it means making its actual capability harder to overread.

Repeated exposure can also support calibration, although repetition does not remove every source of mismatch. Long-term interaction with Pepper increased perceived sociality and competence while novelty and usefulness remained unresolved [Laban 24, Neef 25]. The multi-session Ameca study followed a similar pattern in a university setting: students remained engaged across four weekly sessions, but appearance-related disturbance did not significantly decline, and slow response generation or speech-recognition errors shaped conversation length and turn-taking [Mauliana 25]. Repetition can therefore stabilise some expectations while leaving embodiment-based disturbance or technical friction intact.

Alignment works best when Ameca's role is designed around specific support needs, acceptable data recording, and concrete social situations instead of assumed general usefulness [Rasouli 24]. The tandem-language study followed that logic by using Ameca as an LLM-powered German practice partner for international students at a German university, instead of presenting the robot as a general companion or human substitute [Ashok 25a]. A bounded role gives interlocutors a practical basis for judging what the robot should do and which failures belong to the task limits.

For this thesis, exposure means controlled contact with Ameca's real body and real limits through first encounters, repeated dialogue, teaching, language practice, or public-space interaction. Exposure calibrates expectations by allowing interlocutors to compare imagined capability with what the robot can and cannot do [Rosén 24a]. Alignment is therefore the first mitigation step: it makes the robot's apparent ability answerable to observed behaviour, task constraints, and reliable implementation.

3.2.2 Awareness via Explanation

Awareness via explanation means guiding interlocutors in how to interpret Ameca's behaviour by connecting visible actions to system boundaries. Explanation identifies which behaviours depend on sensors, models, scripts, prompts, memory, uncertainty handling, or failure recovery. Without that explanation layer, communicative actions can be read as agency, fluent dialogue as understanding, and unclear capability boundaries as deception [Saltik 21, Williams 24, Ciardo 26]. Awareness therefore reduces the need for interlocutors to guess what Ameca perceives, knows, stores, or intends.

Concrete explanation is necessary because Ameca's visible behaviour can make system processes appear more coherent than they are. Interlocutors need to know how cameras are used, whether audio is stored, whether dialogue is scripted or generated, whether memory is session-based, whether internet access is available, and how uncertainty is handled. Lack of transparency can contribute to AI fear, and social robots can raise concerns about constant watching, constant listening, data inference, and unauthorized sharing [Zhan 24, Bell 25]. Awareness addresses those concerns by making capability boundaries available before, during, and after interaction.

A small demonstration incident in June 2023 shows why explanation has to address both generated output and embodied presentation. During an early RRLab-made LLM demonstration, Ameca answered a weather question with "it is raining cats and dogs" on a sunny German summer day because the model was not connected to real-time information and hallucinated a non-factual response [Ashok 24a]. After the limitation was explained, a youth in his early 20s pointed to the illuminated back of Ameca's head and asked, "Maybe it is pretending to be dumb so we are fooled?" The incident shows that Ameca's body can make even a hallucinated answer appear strategic when interlocutors cannot see the boundary between generated language and actual system knowledge.

Social-role framing is one way to make that boundary easier to interpret. Role framing can shape trust and behaviour interpretation through assumptions about agency, morality, and task role [Groom 11, Banks 21]. In this thesis, personified framing increased empathy but also raised expectations for expressiveness, gaze, and contextual awareness, while capability framing made Ameca's body, sensors, non-walking legs, recognition limits, actuator issues, lip-sync limits, and hardware noise easier to interpret [Ashok 25b, Majumder 26]. Personified framing made Ameca socially meaningful; capability framing made its system boundaries easier for interlocutors to judge.

A single disclosure is insufficient because Ameca's body continues to produce social cues after the explanation ends. Early disclosure of a WoZ setup did not strongly change robot perception in prior work, suggesting that explanation can be weakened by embodied behaviour during interaction [Nasir 22]. Explanation therefore has to be concise, precise, coherent, repeated, and consistent across robot identities in HRI [Bransky 26]. The proposed Ameca system prompt[12] turns explanation into an interaction rule by anchoring robot speech in identity, capability boundaries, transparency, failure protocol, privacy, and ethical limits [Ashok 26]. Interactional failure explanations add a repair layer after breakdown by naming the relevant limit and clarifying actual robot capability [Yadollahi 25, Holthaus 26].

In this thesis, Awareness becomes necessary when Ameca's visible abilities create stronger capability impressions than the implemented architecture can reliably support. ChitT5, MultiSessionT5, RoboReID, the teachable-robot study, and the tandem-language study show the progression: fluency, memory, person recognition, learning, and language practice are useful, but they require explanation because interlocutors can misread them as understanding, personal remembering, hidden-state recognition, or general social competence [Ashok 24a, Mauliana 25, Lu 24, Majumder 25b, Ashok 25a]. Explanation therefore becomes a design layer across consent forms, experimenter introductions, robot self-introductions, prompts, failure responses, task roles, and debriefing. Awareness is the second mitigation step: it keeps apparent ability tied to known system boundaries while the interaction unfolds.

3.3 Implications in this Thesis: Using the 5A Framework

The 5A framework structures the remaining thesis chapters as an organising analytical lens for interpreting Ameca-based HRI. The 3As identify where Ameca's body and Apparent

[12] https://github.com/aashok94/llm_powered_social_robot/blob/main/ameca_prompt.json

Ability create expectation mismatch, while the 2As identify how bounded exposure, explanation, role limits, and repair make that mismatch more manageable. Each chapter contributes one layer of the same thesis-level argument: embodiment sets the expectation gap, dialogue and perception make the gap technically consequential, behaviour generation makes the gap visible in motion, and application studies test whether alignment and awareness can support bounded companion-like interaction.

Chapter 4: Robot embodiment, expectation, and uncanniness

Chapter 4 establishes the embodiment basis of the anthropomorphic expectation gap. Earlier work with ROBIN showed that affect adaptation and appearance already shaped perceived social fit [Ashok 22, Ashok 23], but Ameca made the expectation gap harder to ignore because its body, face, voice, and movement invited stronger social interpretation. The 3As appear when Ameca's humanlike embodiment, partial capability, and interlocutor discomfort converge: person recognition, empathetic dialogue, and visual awareness made separate technical functions appear as continuity, understanding, and perception [Lu 24, Sinha 23, Mohamed 23]. Perception studies then showed how participants connected identity, voice, gender, affect, movement, and failure to one social body [Jafarova 25, Puertas Céspedes 25, Beck 26, Kushina 23, Ashok 25b, Berns 24, Esposito 25]. Public-space work with older adults further showed that Ameca's physical presence and gaze could be experienced as both fascinating and unsettling, which made the embodiment gap visible beyond the student setting [Jahn 25, Jahn 26]. The 2As appear through tutee framing, physical-first exposure, and repair, which made Ameca's limits easier for interlocutors to interpret and made the interaction more predictable [Ashok 24b, Majumder 25b, Majumder 26, Kukaj 24, Beck 26]. Across these studies, the chapter draws on 17 interlocutor studies with $N = 491$ reported participants, covering system development, perception, embodiment order, teaching, repair, and public-space interaction.

Chapter 5: Auditory perception and dialogue management

Chapter 5 treats dialogue as a site where humanlike fluency has to be bounded by explainable system behaviour. Ameca's face, gaze, and voice made interlocutors expect fluent, coherent, and socially appropriate conversation. The 3As appear when auditory ability does not fully match that expectation: ChitT5 improved first-session interaction but produced topic drift, identity errors, speech-recognition failures, and affect mismatches [Mauliana 23, Ashok 24a], while MultiSessionT5 added memory, emotion handling, and factual grounding but made interlocutors more sensitive to continuity, self-consistency, and role confusion [Mauliana 24, Mauliana 25]. The 2As appear through controlled local pipelines, improved speech transcription, retrieval-augmented generation, open-source LLM integration, predictable lip-synced speech, and prompt-level boundaries for identity, capability, transparency, privacy, failure, adaptation, and ethical limits [Esiagu 26, Ashok 26]. The implication for the 5A framework is that dialogue should not aim for unbounded humanlike illusion. Dialogue should make Ameca's conversational ability accountable to the architecture that produces it. In total, the chapter reports $N = 89$ participants across the auditory-dialogue evaluations.

Chapter 6: Visual perception and gaze management

Chapter 6 treats visual perception as expectation management as well as sensing. Ameca's eyes made interlocutors expect seeing, attention, and socially fair gaze. The 3As appear when visual ability partly matches that expectation but remains technically bounded: gesture recognition, object-to-self recognition, Tangram perception, visual question answering, human-attribute recognition, and gaze control made Ameca appear visually aware while also exposing errors in recognition, timing, self-other distinction, and attention allocation [Islam 23, Huang 24, Breininger 23, Akinci 24, Goede 22, Srivastava 22, Paplu 23b]. The 2As appear through better alignment of gaze, speech, and visual input. Active-speaker detection, Hero selection, and speech-gaze fusion made Ameca's attention easier to follow, while public-space interaction showed that gaze can still feel intrusive or exclusionary when interlocutors cannot predict why the robot attends to one person and not another [Huang 25, Jahn 26]. The implication for the 5A framework is that visual perception needs behavioural coordination and explanation, because apparent seeing can quickly become apparent social awareness. In total, the chapter reports $N = 136$ participants, plus interpretable in-the-wild utterances from 24 visible people.

Chapter 7: Robot behaviour generation and adaptation

Chapter 7 links behaviour generation to the alignment between visible motion and reliable robot capability. Ameca's anthropomorphic face, eyes, arms, and hands made interlocutors expect visible behaviour to match speech, perception, and task context. The 3As appear when facial expressions, gaze, gestures, mimicry, and pointing create Apparent Ability but also reveal timing limits, AU-to-actuator mismatch, gesture-detection errors, and morphology-constrained movement [Shakeebae 24, Terdale 23, Puertas Céspedes 25, Tolj 23, Ashok 24b, Sevinc 24, Majumder 25b]. The 2As appear through behaviour authoring, AU-inspired facial mappings, gaze sequences, robot-as-tutee gesture learning, live mimicry, and task-oriented pointing, where the design goal was interpretable movement supported by implementation rather than human-level expressivity [Huang 25, Akinci 24, Jeong 26]. The implication for the 5A framework is that movement should clarify what Ameca is doing, what it has detected, and where its action limits are, because visible behaviour can otherwise make partial competence appear like general social intelligence. In total, the chapter draws on $N = 107$ participants across behaviour-related evaluations.

Chapter 8: Social robots as tandem language partners

Chapter 8 tests whether Alignment and Awareness can make a humanlike robot acceptable as a constrained companion for university students. Ameca was positioned as a peer-like tandem partner for German speaking practice because its anthropomorphic body could approximate social presence while reducing the judgement associated with human interlocutors. The 3As appear in a bounded application context: Ameca's appearance and dialogue ability created expectations of responsiveness, correction, memory, and turn-taking, while anxiety shifted from robot uncanniness toward language-related fear of judgement, pacing, and interactional constraint [Ashok 25a]. The 2As appear through role-bounding and repeated exposure. Ameca was framed as a low-pressure practice partner with controlled prompts, local inference, short German turns, and supportive correction, which kept the interaction away from teacher, therapist, or native-speaker replacement

roles. The implication for the 5A framework is that companion-like interaction becomes more acceptable when the robot's role is narrow enough for interlocutors to understand what kind of support is being offered and where the robot's responsibility ends. In total, the chapter reports $N = 106$ analysed participants across the Foreign Language Anxiety survey, single-session tandem study, and multi-session tandem study.

Synthesis of the 5A framework

Across the thesis, the 5A framework narrows Rosén's social robot expectation gap toward highly anthropomorphic robots such as Ameca. The framework does so by separating three diagnostic pressures from two mitigation responses. *Anthropomorphism* makes Ameca socially interpretable before performance is known, *Apparent Ability* makes partial technical functions look like broader social understanding, and *Anxiety* emerges when interlocutors cannot place what the robot sees, understands, remembers, intends, or can be trusted to do. The mitigation response is *Alignment* and *Awareness*. Alignment lets interlocutors encounter Ameca's real body, role, task behaviour, and limits; Awareness explains the system boundaries behind sensing, memory, AI generation, privacy, and failure. The thesis-level implication is that improving Ameca as a companion for university students does not require increasing human likeness. It requires a bounded role, accountable behaviour, and visible alignment between the robot's apparent ability and the implemented architecture that produces its behaviour.

4. Robot Embodiment, Expectation, and Uncanniness

Chapter 3 defined the anthropomorphic expectation gap as the mismatch between what a highly humanlike robot appears to promise and what the implemented system can reliably deliver. This chapter examines that gap through Ameca, where technical capability, social interpretation, and uncanniness converged at the level of embodiment.

The PhD initially followed a cognitive-robotics trajectory through speech, affect recognition, facial expression, and perception for social robots in university settings. The first study used ROBIN, the RoboThespian-based humanoid platform, to adapt dialogue, facial expression, and gesture behaviour from paralinguistic emotion recognition in robot-directed speech [Ashok 22]. A later ROBIN study examined formal and casual clothing in a student-service scenario and found an almost even preference split, with casual clothing associated with friendliness, comfort, helpfulness, and approachability [Ashok 23]. ROBIN therefore established appearance and affect as interactional design variables, while its projected 3D face kept embodiment mismatch relatively limited. In-person reactions focused more on mechanical noise during movement, as visible in the RRLab video A Conversation with the Humanoid Robot "ROBIN"[1]. ROBIN gives the comparison point for the chapter, since affect, clothing, and social role already shaped perception before Ameca made the body itself the expectation-setting interface.

Ameca changed the interpretation of system development after it became the main RRLab platform in 2022. The technical work still targeted dialogue, perception, recognition, expression, and behaviour generation, yet studies conducted between 2023 and 2024 showed that higher software capability did not fully explain user response. Users reported discomfort, unease, and eeriness even as the system became more interactive. The relevant problem was therefore not what Ameca could do in isolation, but what its humanlike body made people expect from recognition, gaze, speech, memory-like behaviour, and failure repair.

The chapter examines Ameca as a concrete case of the 5A framework. System-development work first exposes the gap between apparent and reliable ability. Perception studies then

[1] https://youtu.be/v9YbAG7TBfo?si=vrkiXzM2NXD-KZJ-

show how identity coherence, voice, gender, affective framing, movement, and uncanniness shaped interpretation. Expectation matching is then examined through robot-as-tutee framing, embodiment order, and social repair, where Ameca's appearance, role, behaviour, and limits become more readable through exposure and explanation.

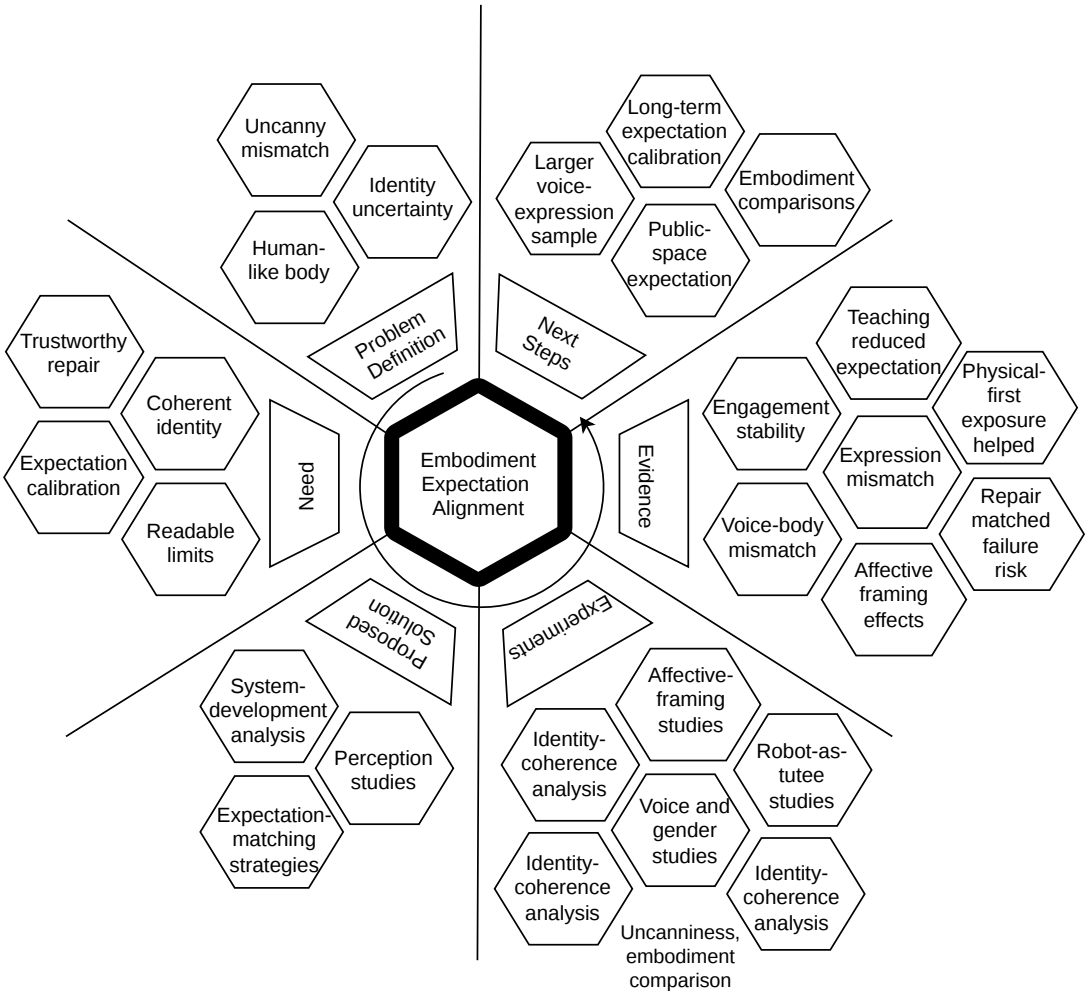


Figure 4.1: Overview of the embodiment expectation-alignment pipeline.

4.1 Overview: Appearance, Expectation, and Uncanniness

Appearance, expectation, and uncanniness provide the empirical entry point for studying Ameca as an embodied social robot rather than a neutral software platform. Table 4.1 situates the chapter within user studies on robot appearance, embodiment, expectation, uncanniness, and gendered social interpretation. The studies show that voice, media exposure, physical presence, task role, and gender cues can shape robot interpretation before technical capability is evaluated.

Ameca concentrates these variables in one platform. Its humanlike body made partial dialogue, perception, expression, recognition, and repair abilities socially consequential,

so embodiment cannot be reduced to visual design. In this chapter, embodiment marks the interface through which apparent ability, identity coherence, affective framing, and expectation mismatch become observable.

4.2 Expectation mismatch in Ameca

Expectation mismatch in Ameca emerged gradually across studies that were not originally designed around uncanniness. System-development projects improved person recognition, empathetic dialogue, visual awareness, gesture learning, facial expression, and behaviour generation, but the platform did not invite users to evaluate those functions as isolated modules. Ameca appeared as one embodied social agent whose face, gaze, voice, expression, and response timing made partial capability look socially meaningful.

Anthropomorphism and apparent ability therefore interacted from the beginning. Ameca's appearance suggested perception, intelligence, emotional expression, and social awareness before the architecture could support those in a stable way. Recognition implied continuity, fluent dialogue implied understanding, facial expression implied emotion, gaze implied attention, and backstory implied inner life. Coherent cues made Ameca easier to read, while drifting cues exposed the distance between implemented modules and perceived social agency.

The following subsections trace this mismatch across system-development work and perception studies. The system-development studies show how partial technical abilities acquired social meaning inside Ameca's body. The perception studies show how users and viewers interpreted those abilities through identity, voice, gender, affect, movement, and failure.

4.2.1 System-development indications of expectation mismatch

System-development work produced the first evidence that capability in Ameca could not be evaluated only as module performance. Person recognition, empathetic conversation, and visual question answering were designed to improve interaction, yet each function changed what the robot appeared able to do. Recognition suggested continuity across encounters, empathetic dialogue suggested affective understanding, and visual dialogue suggested scene awareness. Uneven performance therefore became an embodied interpretation problem, since users encountered delay, error, and uncertainty through a humanlike face and attentive body.

Apparent ability increased with each added function, while the boundary between visible competence and reliable capability became easier to notice. Table 4.2 summarizes three system-development studies that made this boundary visible in Ameca.

RoboReID enabled personalization, but live conditions made recognition noisy and delay-sensitive. The empathetic dialogue system produced polite and affectively appropriate responses, but the interaction flow did not match the seamless conversational standard invited by Ameca's appearance. The visual-awareness system enabled scene-based dialogue, but dependability and confidence remained moderate for real-world visual tasks. Detailed implementation, measures, and questionnaire items are provided in Appendix A.1.

Table 4.1: Comparison of related user studies on robot appearance, expectation, embodiment, and social interpretation.

Reference	Setting	Robot / Stimulus	Study Focus	Key Finding	Sample
[Lindblom 25]	In-person	Pepper + GPT-3	Expectations and UX	Users adapted strategies to perceived capability	$N = 31$ 62 videos + interviews
[Rosén 24b]	In-person	Pepper + GPT-3	Expectation change	Initial expectations largely remained	$N = 31$ 2 sessions
[Fortunati 22]	Online video	Sophia YouTube video	Human-likeness and mind attribution	Human-likeness not strongly salient; uses gendered	$N = 268$ students
[Kopecka 25]	Online image	Toyota HSR Nao Nadine	Appearance and expected explanations	Humanlike appearance linked to anthropomorphic explanations	$N = 180$ 60 per condition
[Pang 25]	In-person field	ERICA TELECO	Appearance and spoken behaviour	ERICA elicited fewer disfluencies and more complex syntax	$N = 42$ conference participants
[Wang 22]	Online focus groups	Humanoid robot images	Appearance qualities and gender attribution	Hair, apparel, face, eyes, colour, neck, and body shape guided gender	$N = 16$ 3 focus groups
[Mathur 20]	Lab / online ratings	182 robot and human faces	Uncanny valley and category confusion	Clear uncanny valley; category confusion not supported as mechanism	$N = 358$ 6 sites
[Saygin 12]	Lab fMRI	Human Robot Android Repliee Q2	Appearance-motion mismatch	Humanlike appearance with mechanical motion produced prediction error	$N = 20$ 19 usable
[Kang 25a]	In-person + media	Nadine + sci-fi films	Media exposure and direct robot interaction	Films increased nervousness; robot interaction moderated it	$N = 55$ teenagers
[Dennler 25]	Online video	Quori voice + clothing	Voice, clothing, task, and gender perception	Voice and clothing shaped gender; effects were not simply additive	$N = 65$ $N = 93$ $N = 273$

Table 4.2: System-development studies as early indications of expectation mismatch in Ameca.

Goal	System	Sample	Finding	5A
Recognize returning users [Lu 23]	Audio-visual person re-ID	N=12; lab access scenario	Face recognition outperformed voice; live use remained noisy and delay-sensitive.	*Ability*: recognition suggests continuity.
Generate empathetic conversation [Sinha 23]	Fine-tuned dialogue models	N=8; exam-related scenarios	Respect and empathy were perceived; flow was not always natural.	*Ability*: affective responses suggest understanding.
Support visual Q&A [Mohamed 23]	Visual-dialogue system	N=6; 15-20 min lab interaction	Answers were functional but mixed; confidence and dependability stayed moderate.	*Ability*: visual awareness raises perception expectations.

The shared contribution of these studies is the technical root of the expectation gap. Ameca's embodiment made each recognition attempt, answer, gaze shift, delay, and error appear as behaviour from one social mind. Perception studies then show how users tried to stabilise that social mind across identity cues, voice, affect, motion, and failure.

4.2.2 Perception studies as evidence of expectation mismatch

Perception studies expose the user-facing form of Ameca's expectation mismatch. Users and viewers did not judge face, voice, name, gender, movement, affect, and interaction quality as separate design layers; they tried to make them cohere as one social agent. YouTube comments show identity coherence as an unstable relation between name, voice, gender, appearance, intelligence, and apparent inner state. Voice studies test whether vocal identity could stabilise Ameca's body or sharpen mismatch. Affective-framing studies examine whether emotion, self-reference, and backstory made Ameca socially legible or made expressive limits easier to notice. The uncanniness study then connects appearance, expression, movement, and interaction quality across ROMAN, ROBIN, and Ameca.

These perception studies connect anthropomorphic form, visible ability, and anxiety. The body invited social interpretation, the behaviours made that interpretation believable, and mismatch across identity, voice, affect, motion, or capability made the robot harder to place.

Identity Coherence

RRLab YouTube[2] comments show that viewers interpreted Ameca through a social identity already formed by name, voice, gender, appearance, intelligence, and apparent inner state. Many viewers likely arrived with expectations shaped by polished Engineered Arts demonstrations, then encountered RRLab videos showing an experimental system

[2] https://www.youtube.com/RRLabTUKaiserslautern

under active development. One commenter made that contrast explicit, arguing that the videos risked making the robot "look like a deliberate attempt to defame Engineered Art's product," while also stating that "a robot can only be as good as the machine learning model behind it" (@EndingSimple, *Ameca: Unpredictable Responses from the Humanoid Robot!*[β], 2022). A body with a strong public identity therefore made any divergence in voice, naming, behaviour, or capability readable against a prior image.

Table 4.3 summarizes the main comment patterns. Full comments, usernames, video titles, and translations are provided in Appendix A.2.

Table 4.3: YouTube comment patterns showing identity-coherence problems around Ameca.

Theme	Excerpt	3A problem	Implication
Name and identity continuity	"Emah looks like Ameca but is still learning"	*Anthropomorphism* + *Ability*	The same body was expected to carry one stable identity.
Voice continuity	"Imposter!!! Thats not Amecas voice!!"	*Anthropomorphism* + *Ability*	Voice was read as identity, not as a replaceable technical layer.
Gendered voice-body mismatch	"that male voice sounds creepy"	*Anthropomorphism* + *Anxiety*	Voice, body, and gender cues had to cohere.
Apparent mind and empathy	"poor Emah gets tired and want to go to sleep"	*Anthropomorphism*	Viewers inferred fatigue, empathy, or inner state.
Technical failure as identity failure	"AI messed up completely"	*Ability* + *Anxiety*	Dialogue failure was judged against the promise of a capable humanoid.
Movement and expression mismatch	"sudden hands moving forward disturbing"	*Ability* + *Anxiety*	Uncanniness attached to timing, gesture, and expression limits.

The comment patterns show one shared assumption behind positive and negative reactions, where Ameca was read as one social entity. Viewers who wrote "I love Ameca" or "Love you Ameca!" did not respond to a camera, TTS voice, prompt, or gesture controller in isolation. Viewers who called the robot "hallucinating.creepy" or "definitely much worse than I expected" also judged a whole embodied agent. Face, name, voice, gender, movement, intelligence, and apparent empathy therefore operated as identity cues that either stabilised or weakened the robot's social readability. This also links Ameca's identity-coherence problem to artificial identity design, where recognisability, behavioural consistency, continuity, memory, role clarity, artificiality, and transparency define whether an agent identity remains interpretable across interaction contexts [Bransky 26]. The voice studies examine the same coherence problem through the cue that viewers most often associated with identity.

Vocal Identity and Gendered Embodiment

The YouTube identity-coherence problem motivated three studies on Ameca's voice. The first study tested which artificial voice best fit Ameca's appearance [Jafarova 25]. The

[3] https://youtu.be/TfiOU6gU5MI?si=fvWQSA1wpD0zEzFO

second study tested whether voice condition and facial-expression onset shaped perceived animacy [Puertas Céspedes 25]. The third study examined whether a female-gendered voice changed judgements of Ameca in a conflict scenario, especially when the robot responded apologetically or assertively [Beck 26]. The studies ask whether voice can stabilise Ameca's social identity or introduce additional mismatch between body, role, affect, and agency.

Table 4.4 summarizes the three studies. The detailed research questions, voice implementations, measures, questionnaires, scripts, descriptive statistics, and full test results are provided in Appendix A.3.

Table 4.4: Voice studies as evidence of identity and expectation mismatch in Ameca.

RQ	**Setup**	**Sample**	**Finding**	**5A**
Which voice best fits Ameca? [Jafarova 25]	Lucy, Graham, Joanna, Mathew, and Q voice	N=32; in-person; within-subject	Only anthropomorphism differed; Mathew was rated higher than Joanna.	*Anthropomorphism* + *Ability*: voice affected how humanlike Ameca felt.
Do voice, emotion, and onset shape animacy? [Puertas Céspedes 25]	Lucy, Graham, Q; seven facial emotions; three onsets	N=27; in-person; between-subject voice groups	Animacy was driven mainly by emotion and onset; voice interacted with disgust.	*Ability*: voice did not act alone; face and timing shaped social readability.
How is a female-voiced Ameca judged in conflict? [Beck 26]	Lucy voice; apologetic vs. assertive response	N=40; online; within-subject	Apology improved CRE and CT; assertiveness raised negative affect and mistreatment.	*Anxiety*: role-incongruent assertiveness made Ameca feel more socially risky.

Voice did not solve Ameca's identity problem through simple fit selection. In the artificial-voices study, voice condition changed anthropomorphism but did not produce broad differences in animacy, likeability, perceived intelligence, perceived safety, sound quality, or pleasantness [Jafarova 25]. Viewer comments had framed voice as a strong identity marker, while controlled evaluation showed that fit depended on the relation between voice, body, face, role, and expected behaviour.

Voice was already entangled with gendered appearance. In the voice-emotion-onset study, participants saw unlabelled images of Azi and Ameca after the main task. Azi was read as stably male-coded, whereas Ameca was often read as female-coded or visually ambiguous. Participants justified this through neck width, face shape, softer facial structure, larger eyes, fuller lips, and a more caring or expressive appearance. Voice fit was therefore evaluated against a body that already carried gendered cues.

Voice became socially consequential once Ameca entered a role. In the female-voice conflict study, participants judged whether a female-voiced humanoid assistant should apologize, defend itself, correct blame, or ask for respectful treatment. The apologetic response was rated higher for conflict resolution effectiveness and comfort/trust, while

the assertive response produced higher negative affect and perceived mistreatment. Open-ended responses showed that participants did not want endless submission, yet also rejected a robot that sounded defensive, angry, or too person-like.

Vocal identity therefore sits at the intersection of anthropomorphism, apparent ability, and anxiety. Voice makes the humanlike body more socially readable, while also raising expectations about coherence, affect, and role-appropriate behaviour. When voice, face, expression, and stance aligned, Ameca became easier to place. When they drifted, mismatch became another source of uncertainty. Affective framing tests the same problem through emotion, self-reference, and narrative.

Affective Framing

Affective framing shaped how Ameca was read before performance was judged [Groom 11]. Robot behaviour gains meaning through expectations about the interaction and the agent role assigned to the robot. In a highly humanlike platform, emotion, self-reference, and backstory can support social legibility, but they also raise the demand for gaze, timing, facial movement, and contextual coherence [Strait 15, Banks 21]. The studies in this line used scripted dialogue, facial expression, emotional reaction, and narrative backstory to test whether Ameca became easier to read or whether expressive limits became more visible.

Table 4.5 summarizes the two affective-framing studies. The detailed research questions, monologue conditions, measures, open-ended responses, descriptive tables, and statistical results are provided in Appendix A.4.

Table 4.5: Affective-framing studies as evidence of expectation mismatch in Ameca.

RQ	**Setup**	**Sample**	**Finding**	**5A**
Does emotional expression and self-reference increase perceived humanlikeness? [Kushina 23]	Basic, Emotional, and Emotional-Reaction monologues	N=16; between-subject; usable data limited	Emotional condition had the highest anthropomorphism mean, but evidence remained exploratory due to the small sample.	*Anthropomorphism*: emotion may raise humanlike reading, but evidence remained tentative.
Which backstory facilitates empathy most: technical, happy, or sad? [Ashok 25b]	Technical, happy, and sad backstory videos with expressions and gestures	N=60; within-subject video study	Sociability increased, disturbance decreased, and the sad backstory produced the highest empathy.	*Ability* + *Anxiety*: framing made Ameca more readable, but exposed expressive limits.

The self-awareness study remained exploratory, yet it exposed the tension created by affective expression. Adding emotion did not produce a stable linear increase in humanlikeness. Open-ended responses showed that stronger emotional display also made participants

scrutinize gesture, gaze, timing, and the boundary between acceptable robot competence and human-level emotional understanding. One participant described the imitation of human gestures as uncomfortable, another described the hand wave as unnatural and the gaze as hostile, and the strongest response set a clear boundary by stating that the robot would not be expected to support a deeper worrying conversation with human-level understanding.

The backstory study showed the same relation with stronger evidence. Emotional narrative made Ameca more socially readable, with increased sociability, decreased disturbance, and the sad backstory producing the strongest empathy. Qualitative responses also showed that affective framing raised the standard for contextual coherence. Participants noticed when Ameca did not look toward Robin while referring to him, when the robot seemed to try too hard to act human, and when facial expression lacked eyelid or pupil movement needed for readable emotion.

Affective framing therefore helped social readability while making expressive limitations easier to detect. Emotion, self-reference, and backstory made Ameca's humanlike body more interpretable as a social agent, yet also increased the burden on gaze, timing, facial movement, and contextual behaviour. The uncanniness study examines the same relation across appearance, expression, movement, and interactional readability.

Uncanniness

The first direct corridor encounter with Ameca asked what happens when a highly humanlike robot appears in an everyday university space outside the framing of a laboratory study. The hallway study was motivated by the social robots in public spaces workshop initiative [Schneider 22a] and later reported within the published uncanny valley work [Berns 24]. A Wizard-of-Oz deployment placed Ameca in the RRLab corridor, monitored it through its camera view, and controlled it from behind the green door shown in Figure 4.2. When a person entered the field of view, Ameca introduced itself and asked whether the person had time to answer three short questions. Many people ignored the robot or walked away; others stopped, answered, tested, laughed, photographed, questioned, or tried to understand what kind of social agent had appeared in a familiar corridor.

One interaction from this deployment exposed the expectation gap more clearly than the questionnaire. A group of German-speaking bachelor students, one female student and four to five male students, sustained the longest interaction of the day. They answered Ameca's questions in turns, moved close to the robot to test tracking, played with the microphone, and asked follow-up questions outside the script. When the session had to end, Ameca said goodbye and thanked them, yet the group remained. The Wizard-of-Oz operator then stepped out from behind the green door. The students were visibly surprised. One male student explained that they had assumed Ameca was "a slow AI". The female student reacted more strongly, saying that she thought she and Ameca were getting along well after the robot looked at her and remembered her name. The encounter captured the central mechanism of the chapter, since Ameca's embodiment invited assumptions about agency, memory, attention, and social understanding before the control setup was visible.

The uncanny valley hypothesis describes a non-linear relation between human likeness and affinity, where artificial agents often become more acceptable as they appear more humanlike until imperfect near-human appearance produces eeriness [Mori 12, Mathur 20]. In Ameca,

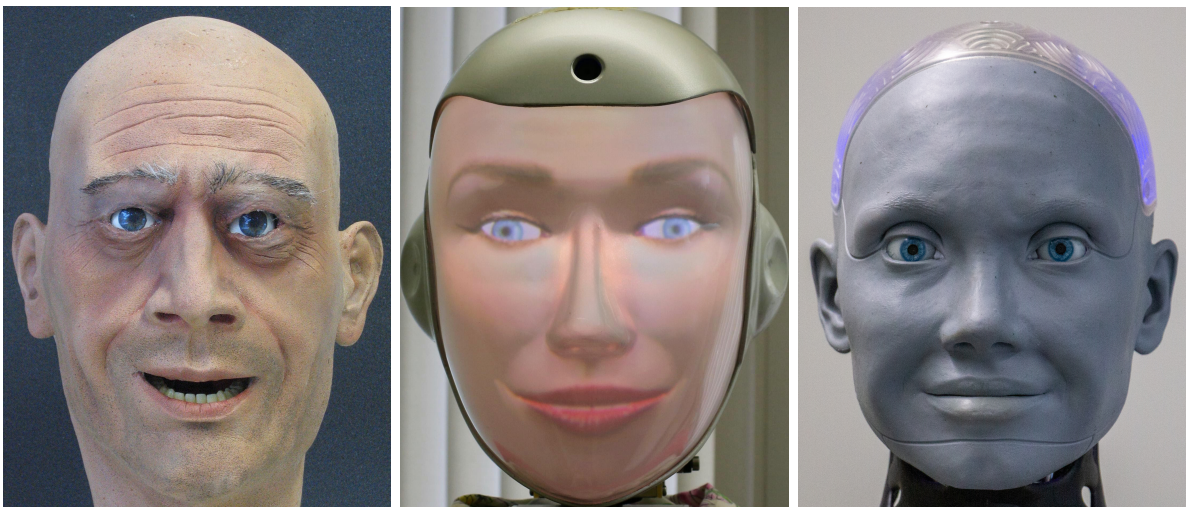


Figure 4.2: Ameca deployed in the RRLab corridor during the hallway study. This image is one of the four images shown in Figure 2 of Berns et al. [Berns 24] and is reused under the Creative Commons Attribution (CC BY) license.

uncanniness is linked to expectation mismatch across appearance, facial expression, motion, voice, timing, gaze, and interactional behaviour. Human likeness alone cannot explain the corridor response, since the robot also had to sustain the social assumptions created by its body.

The published study transformed the corridor insight into a controlled online comparison of three humanoid upper-torso robots developed or used in the RRLab context: ROMAN, ROBIN, and Ameca [Berns 24]. The within-subject study with N=67 participants examined which robot was perceived as most or least uncanny and how perceived emotion depiction and movement quality affected perceived humanlikeness. Figure 4.3 shows the three robots used in the comparison. The detailed procedure, measures, open-ended responses, reanalysis in R Studio results, and descriptive statistics are provided in Appendix A.5.

Open-ended responses clarified why humanlikeness changed after the videos. Participants did not judge the robots only from static facial appearance; they revised their impressions after the robots moved, spoke, and attempted to express emotion. ROMAN made mismatch most visible through the gap between a humanlike mask and limited expressivity. ROBIN followed the opposite pattern, with less realistic appearance becoming more humanlike through situated responsiveness. Ameca occupied a middle position, where facial structure supported humanlike expectations while limits in mouth movement, eyebrows, and affective intensity remained visible.



(1) ROMAN (2) ROBIN (3) Ameca

Figure 4.3: Robots compared on emotive expressiveness and interaction movement in the uncanny-valley study. The happy facial-expression images of ROMAN, ROBIN, and Ameca correspond to the image clusters shown in Figures 3, 5, and 7 of Berns et al. [Berns 24] and are reused under the Creative Commons Attribution (CC BY) license.

The reanalysis showed the same pattern at the measure level. The robots did not differ significantly on the direct eeriness item, despite descriptive differences. They differed on surrounding measures such as humanlikeness, familiarity, anthropomorphism, likeability, emotion accuracy, and movement elegance. Perceived emotion accuracy and movement elegance predicted perceived humanlikeness. In the uncanniness model, emotion accuracy predicted lower eeriness, while humanlikeness itself did not significantly predict eeriness. The result links discomfort to the readability of expression and movement, since these behaviours had to support the expectations created by embodiment.

A related under-review study extends expectation mismatch to robot failure [Esposito 26]. In an online escape-game scenario, N=71 participants watched Temi[4], Miroka[5], Pepper[6], and Ameca produce the same incorrect response, framed either as unintentional failure or intentional deception. Ameca received higher mind and agency ratings than the other robots, although the error was identical. Deceptive framing increased mind attribution across robots, without interacting with morphology. Qualitative interpretation showed the same pattern, as failures were explained through programming, compliance, or system limits, while deceptive framing invited more social and relational explanations. In Ameca, a technical error can therefore look more agentic due to the body that delivers it.

Public-space work with older adults gives the embodiment problem a boundary case outside the university-student setting. In a preliminary robot-avatar comparison with four older adults, Ameca produced stronger emotional responses than the screen-based avatar Ann-Sophie, with participants appreciating its realism and social cues while also reporting unease and questioning its role in care-giving [Jahn 25]. The later in-the-wild study at

[4] https://www.robotemi.com/?srsltid=AfmBOortBRizOqWe2ELLKmLsLevYJy7muxakNGuHcy9zm2l9WnJHjbYH
[5] https://enchanted.tools/robot/
[6] https://en.wikipedia.org/wiki/Pepper_(robot)

a Senior Citizens' Fair extended this pattern with 25 participants aged 24 to 90. With conversational ability held comparable through LLM-based interaction, Ameca's physical presence and gaze were described as both fascinating and unsettling, while Ann-Sophie was perceived as warmer, more controllable, and socially safer [Jahn 26]. The older-adult work shows how the same embodiment-driven expectation gap becomes sharper when age, public setting, gaze, physical presence, and care associations converge.

Embodiment-driven discomfort also appeared at the proxemic level, where a related Ameca in-person study found that reported comfort during small-talk interaction was highest at 2 m and slightly lower at closer distances, while gaze- and pupil-derived features provided preliminary evidence for estimating comfort during interaction with a highly anthropomorphic robot [Kushina 25].

The uncanniness studies therefore connect the perception evidence back to the expectation gap. Anthropomorphic form created the human comparison, apparent ability was evaluated through expression, motion, timing, and interactional readability, and anxiety appeared when these cues failed to cohere. Ameca's problem was not human likeness in isolation; it was the expectation of coherence across face, motion, gaze, voice, timing, memory, and response behaviour. The mitigation studies therefore focus on making these limits more readable instead of removing the humanlike body.

4.3 Expectation matching as a mitigation strategy

Expectation matching addresses the mismatch produced when Ameca's humanlike embodiment makes partial abilities appear socially meaningful before their reliability is clear. The goal was not to reduce Ameca's anthropomorphism or present the robot as less advanced, but to make the relation between role, behaviour, appearance, and reliable capability easier to interpret.

Alignment and awareness provide the mitigation logic. Alignment concerns whether Ameca's role, behaviour, appearance, and reliable capabilities fit from the user's perspective. Awareness concerns whether users gain enough experience or explanation to understand what the system can do, where it depends on them, and where it can fail. In this chapter, expectation matching was tested by repositioning Ameca as a tutee, controlling embodiment order, and repairing social breakdowns after failure.

The robot-as-tutee studies show how incompleteness can become a legitimate interaction condition. The embodiment-order study shows how physical exposure can anchor expectation before imagined ability expands. The repair studies show how humour, apology, explanation, and instruction-seeking can make failures socially manageable after mismatch appears. These strategies do not eliminate the anthropomorphic expectation gap; they make Ameca's limits more visible, interpretable, and usable during interaction.

Robot as Tutee

The robot-as-tutee strategy repositioned Ameca as a learner instead of a competent autonomous social agent. This role changed the standard against which performance was evaluated. A highly anthropomorphic robot can invite expectations of fluent, autonomous, humanlike competence, while a tutee role makes incompleteness compatible with the task.

Errors can then function as teaching opportunities instead of immediate evidence that the humanoid promise has failed.

Two studies explored this strategy. The CopyCat study framed Ameca as a gesture-learning tutee in a university teaching scenario [Ashok 24b]. The storytelling study extended the same logic to a longer narrative task, comparing participants who actively taught Ameca the *Cracked Pot* story with participants who only watched the resulting robot storytelling video [Majumder 25b]. Table 4.6 summarizes both studies. Detailed implementation, measures, participant procedures, and extended results are provided in Appendix A.6.

Table 4.6: Robot-as-tutee studies as expectation-matching strategies for Ameca.

RQ	**Setup**	**Sample**	**Finding**	**5A**
Can Ameca be framed as a gesture-learning tutee? [Ashok 24b]	CopyCat gesture imitation with five target gestures	N=15; in-person; within-subject lab study	Participants corrected, encouraged, and re-taught Ameca instead of treating all errors as failure.	*Alignment*: the tutee role made imperfection interactionally meaningful.
Does active teaching reduce expectation mismatch? [Majumder 25b]	Teaching the *Cracked Pot* story; teaching vs. viewing groups	N=40; mixed-mode; teaching and non-teaching groups	Active teaching reduced expectation shifts and changed how participants interpreted imperfect output.	*Alignment* + *Awareness*: teaching exposed how the system was built and constrained.

The CopyCat study showed that the tutee frame could absorb imperfect performance. Curious and unsure were rated as more successful gestures, while back off and victory exposed weaker mimicry. These weaker gestures did not break the interaction, since participants adjusted to the learning role through verbal feedback, visible frustration, and encouragement such as “good job Ameca,” “you did well,” and “good robot.” The learning frame gave users a reason to interpret imperfection as teachable.

The storytelling study provided stronger evidence for expectation matching. The non-teaching group entered with higher expectations, while active teaching reduced the size of the expectation shift. Direct teaching made Ameca’s mechanical and expressive constraints visible, changing the robot from an imagined competent humanoid into a bounded technical system. Passive observers relied more strongly on anthropomorphic appearance when judging the final storytelling performance, while participants who had taught the robot were more forgiving of imperfect output. Authorship therefore changed interpretation.

The robot-as-tutee strategy supports alignment and awareness by making incompleteness compatible with the interaction role. Users saw how the robot learned, where it depended on their input, and where the system broke down. The gap between humanlike body and technical limits remained, while the task changed its interpretation by making imperfect output readable as unfinished learning by a visible system. The embodiment-order study tests expectation matching before task interaction begins.

Embodiment Order

Embodiment order controlled how users first encountered Ameca. Many users had already seen Ameca online through Engineered Arts videos, RRLab YouTube clips, or broader media portrayals of humanoid robots. The first encounter was therefore already shaped by prior images. A video-based Ameca can leave more room for imagined capability, while the physical robot immediately exposes height, motor sound, response timing, gaze, mechanical limits, and spatial presence.

The embodiment-order study [Majumder 26] tested this relation by comparing physically present Ameca with a video-based virtual encounter. Table 4.7 summarizes the study. Detailed implementation, measures, procedure, and extended results are provided in Appendix A.7.

Table 4.7: Embodiment-order study as an expectation-matching strategy for Ameca.

RQ	**Setup**	**Sample**	**Finding**	**5A**
Does embodiment order shape expectations, anxiety, and social presence? [Majumder 25b]	Physical Ameca vs. video-based virtual Ameca; order manipulated	N=22; mixed design; physical-first and virtual-first groups	Virtual-first users showed expectation drops and increased anxiety after physical exposure; physical-first users stayed more stable.	*Alignment* + *Awareness*: early physical exposure anchored expectations in observable limits.

The results showed that encounter order changed expectation calibration. Participants who encountered Ameca virtually first showed a drop in expectations after the later physical encounter, especially for emotional and fictional expectations, and they showed a strong increase in robot anxiety from pre- to post-interaction. Participants who encountered the physical robot first showed more stable expectations and no comparable anxiety increase. Physical exposure therefore anchored the robot in observable reality, while virtual exposure left more space for imagined ability before the real robot imposed its limits.

Social presence followed the same pattern. The physical-first group experienced Ameca as socially present during the first encounter, but social presence dropped when they later saw the virtual version. The virtual-first group did not show a meaningful increase in social presence after moving to the physical robot. Qualitative responses clarified the difference, with some participants describing the physical robot as attentive and more natural than expected, and others who first saw the virtual version describing the real robot as intimidating or as a presence watching them. Several responses linked prior expectations to science fiction, suggesting imagined emotional or social capacities that the physical robot could not deliver.

Embodiment order supports alignment and awareness by deciding when the user first sees the robot's scale, timing, sound, and limits. Physical-first exposure does not remove the expectation gap, but it makes the gap easier to manage by anchoring interpretation in the real system before imagination fills in unsupported capability. Social repair addresses the next stage, where mismatch has already occurred and the robot must make the failure understandable.

Normative Social Repair

Normative social repair addresses expectation mismatch after the robot has already failed. The problem changes from expectation calibration before interaction to interaction recovery after visible breakdown. The robot must then make the failure legible, protect the interaction role, and provide a path for continuation.

Two repair cases were relevant in this work. The humour study examined whether conversational humour could repair LLM-induced hallucinations in a highly anthropomorphic robot [Kukaj 24]. The conflict study addressed a higher-stakes reminder failure, where apology, explanation, and explicit instruction-seeking became more appropriate. Table 4.8 summarizes both repair cases. Detailed stimuli, measures, open-ended responses, and statistical results are provided in Appendix A.8.

Table 4.8: Social repair studies as expectation-matching strategies for Ameca.

 **RQ**	 **Setup**	 **Sample**	 **Finding**	 **5A**
Can humour repair LLM-induced hallucinations? [Kukaj 24]	Two Ameca videos with hallucination excerpts and four humour repairs	N=46; online survey	Participants detected hallucinations at high rates; self-enhancing and aggressive humour were preferred most often.	*Awareness*: humour helped when it acknowledged the breakdown and restored coherence.
Can apology and instruction-seeking repair a reminder failure? [Beck 26]	Female-voiced Ameca in apologetic vs. assertive conflict responses	N=40; online; within-subject	Apology improved conflict resolution and comfort/trust; assertiveness raised negative affect and perceived mistreatment.	*Alignment* + *Awareness*: repair worked when it explained the failure and asked for a future rule.

The humour study showed that users noticed hallucinations as embodied coherence failures, not only as incorrect text. Participants pointed to expression mismatch, irrelevant topic shifts, incorrect facts, misrecognition, and incoherent self-description, so the failure exceeded incorrect text. Humour helped under narrow conditions when it acknowledged the mistake, referenced machine-like limits, and helped the conversation continue. It became weaker when it sounded forced, introduced implausible personality, or clashed with the perceived interaction style.

The conflict study showed the boundary of humorous repair. In the missed-reminder scenario, the useful repair was an apology, an explanation, and a concrete request for future instruction. Ameca apologized, explained that the do-not-disturb setting prevented further reminders, acknowledged the user's frustration, and asked whether important reminders should still be delivered in do-not-disturb mode. The repair therefore did more than sound polite; it made the failure legible and converted it into a future interaction rule.

Social repair supports awareness when the robot names what went wrong instead of pretending that the interaction remains coherent. It supports alignment when the repair matches the risk of the failure, with humour softening low-risk conversational breakdowns

and higher-stakes failures requiring apology, explanation, and explicit user instruction. The mitigation sequence therefore closes with a design principle for Ameca, where readable failure can preserve interaction better than an unexplained attempt to appear competent.

4.4 Discussion

Ameca's embodiment did not create expectation mismatch through appearance alone. The mismatch emerged when a highly humanlike body was paired with partial social abilities that users tried to read as one coherent agent. Face, voice, name, gender, gaze, expression, movement, recognition, dialogue, memory-like behaviour, and repair were not interpreted as separate modules. They were assembled into a social body. Coherent cues made Ameca legible; drifting cues made the robot harder to place.

System-development studies exposed the technical root of the problem. Person re-identification, empathetic dialogue, and visual awareness were developed to make Ameca more capable, but each capability raised the standard for the next behaviour. Recognition suggested continuity, empathetic dialogue suggested understanding, and visual awareness suggested perception. In a less anthropomorphic platform, these limits could appear as ordinary technical constraints. In Ameca, they became socially meaningful as the body already implied more than the architecture could consistently deliver.

Perception studies showed the same mechanism from the user perspective. YouTube viewers tried to stabilise Ameca's identity across name, voice, gender, intelligence, and apparent inner state. Voice and gender studies showed that voice was not a replaceable output channel, but a cue for identity and role. Affective-framing studies showed that emotion and backstory increased social readability while making gaze, timing, facial expression, and contextual coherence easier to scrutinise. The uncanniness study confirmed that discomfort was not explained by human likeness alone. Expression quality, movement, timing, and interactional readability mattered as they had to support the expectations created by the body.

Mitigation studies showed that expectation matching can reduce mismatch without solving it completely. The robot-as-tutee frame made incompleteness interactionally meaningful. Embodiment order showed that physical exposure can anchor expectations before imagination fills in missing ability. Social repair showed that failure must be made legible after it occurs, with humour supporting low-risk breakdowns and higher-stakes failures requiring apology, explanation, and a concrete request for future instruction. These strategies changed how the gap was read, not whether the gap existed.

This chapter explains why the technical chapters are part of the expectation-management argument. Dialogue, gaze, visual perception, behaviour generation, and language-learning interaction are not only engineering improvements. They define how Ameca's apparent ability, actual capability, role, and limits fit together during embodied interaction. The design objective is therefore a readable humanoid system whose social promise remains accountable to its implemented perception, dialogue, memory, movement, and repair behaviour.

5. Auditory Perception and Dialogue Management

Chapter 3 defined the anthropomorphic expectation gap as the mismatch between what a highly humanlike robot appears to promise and what the implemented system can reliably deliver. Auditory perception and dialogue management examine this gap at the level where speech recognition, generated response, memory, voice, facial articulation, and timing become one embodied behaviour in Ameca.

Ameca's auditory system had to make dialogue behaviour readable within the robot identity created by its face, gaze, voice, body scale, and lip-synchronised speech. The technical objective was therefore alignment between conversational output and actual system capability, with local processing, bounded identity, controlled memory, and transparent failure behaviour reducing the risk that fluent speech would be read as unrestricted understanding.

The early auditory work began with bounded, rule-based dialogue. User utterances were matched to predefined responses, topic rules, user profiles, and manually controlled dialogue states [Paplu 23a]. Rule-based interaction made Ameca inspectable and easier to constrain, while its rigidity exposed the mismatch between a humanoid body that appeared ready for natural conversation and a dialogue architecture that could only handle predefined structure. LLM-based dialogue improved fluency, topic range, and responsiveness, while also raising the expectation standard for coherence, continuity, identity stability, and failure repair.

Auditory interaction is therefore used as a detailed case of the 5A framework. Speech-to-text and text-to-speech define the input and output boundaries of interaction. ChitT5, MultiSessionT5, and the final local Ollama-based system show increasing forms of apparent conversational ability. Local processing, prompt boundaries, memory control, retrieval grounding, stable robot identity, and reliable lip-synchronised speech then define how alignment and awareness were implemented in the auditory pipeline.

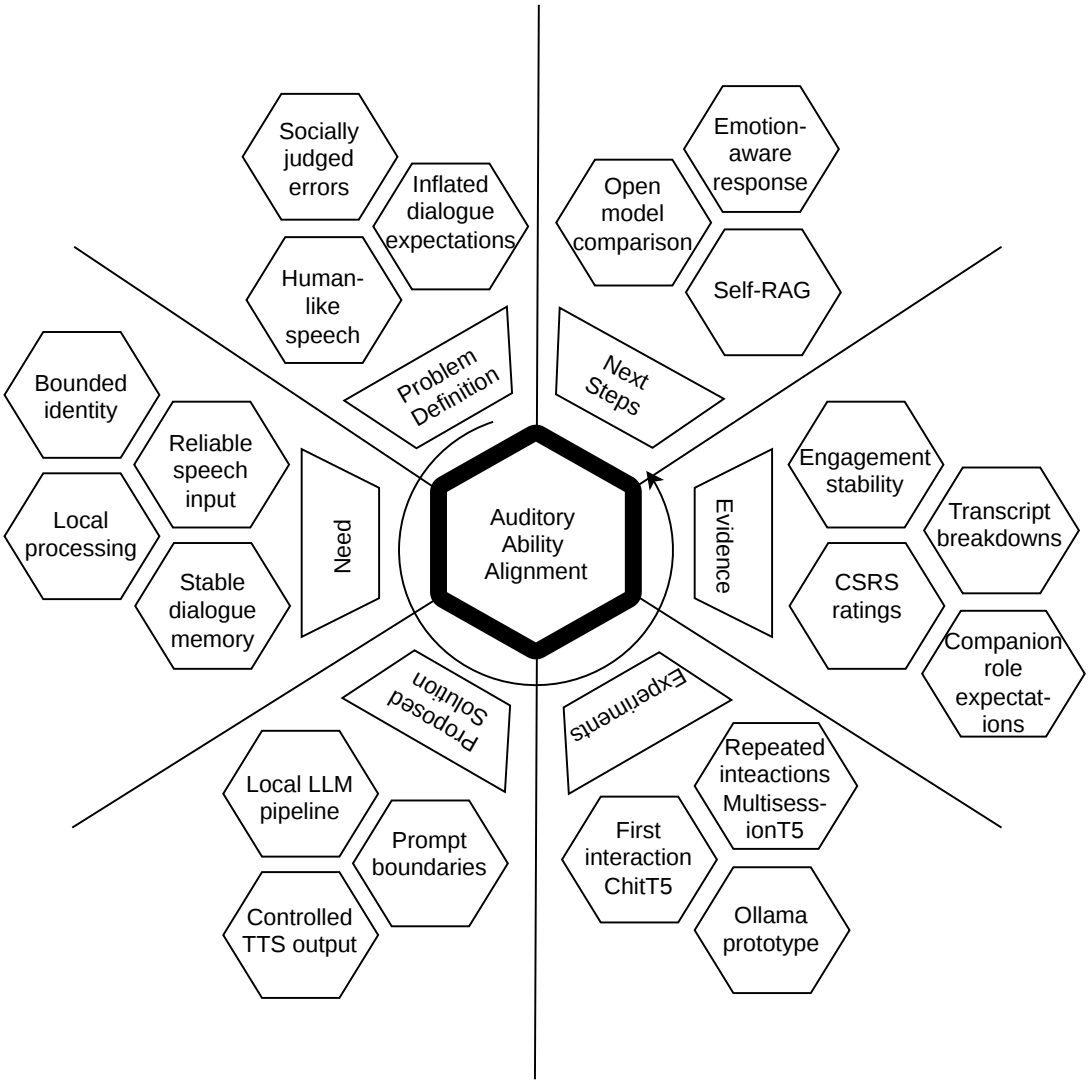


Figure 5.1: Overview of the auditory expectation-management pipeline.

5.1 Overview: Auditory Interaction as Ability Alignment

Ameca's auditory pipeline begins from an embodied expectation rather than a neutral input-output relation. The robot's humanlike face, gaze, voice, moving lips, and physical presence create expectations of conversational and perceptual competence [Li 23b, Hoorn 24]. Default face tracking strengthens this expectation when Ameca appears to look at an interlocutor before answering. Users may therefore expect the robot to listen, attend, understand, and follow basic social rules before the dialogue architecture has shown what it can support. Gaze mechanisms are discussed in Chapter 6; in this chapter, gaze raises the standard against which speech is judged.

Auditory interaction links anthropomorphism to apparent ability through speech-to-text, dialogue generation, memory, retrieval, emotion labels, and text-to-speech. Fluent answers can imply understanding, remembered details can imply continuity, and factual answers can imply knowledge. Mishearing, delay, topic drift, contradiction, or mismatched expression

therefore become embodied failures, since the error is produced through a humanlike face and voice.

The first auditory system used in this work was a bounded rule-based architecture [Paplu 23a]. It combined speech recognition, keyword-based intent handling, manual response templates, topic control, factual response retrieval, user profiling, and interaction summarisation. Detailed implementation notes are provided in Appendix A.9. The system supported alignment and awareness through explicit rules and topic boundaries, while its limited handling of intention, context, nuance, and adaptive grounding conflicted with the expectation created by Ameca's body [Kim 24b]. LLM-based dialogue addressed this limitation through fluency and adaptivity, while increasing the need for coherence, memory control, identity stability, and repair.

5.2 Speech-to-Text

Speech-to-text defines the first technical boundary in Ameca's auditory interaction. Misrecognition changes the interaction state before dialogue generation begins, so later responses can appear incoherent even when the language model follows the input it received. In a highly anthropomorphic robot, users may experience this error as misunderstanding, ignoring, or social failure instead of ASR limitation.

Two STT front-ends were used across the systems in this chapter. ChitT5 and MultiSessionT5 relied on a lightweight Google SpeechRecognition interface. The final Ollama-based system replaced this with a local WhisperX transcriber. Table 5.1 compares these two input boundaries in terms of deployment, robustness, interaction risk, and their 5A role. Detailed implementation notes and pseudocode are provided in Appendix A.10.

For ChitT5 and MultiSessionT5, Google STT was practical because it was easy to integrate and sufficient for short, low-stakes dialogue. Its limitations became clearer once interactions became longer. Accent sensitivity, empty transcriptions, and unstable latency could derail the dialogue before the LLM generated a response. This created an *Ability* mismatch, since Ameca appeared able to listen while the system could not always capture the user's utterance reliably.

WhisperX was introduced in the final Ollama-based system to make transcription more local, predictable, and controllable. This supports *Awareness* because the audio no longer depends on a cloud recognition step, and it supports *Alignment* because the dialogue system receives more stable input. WhisperX still has visible limits in this implementation, including non-streaming transcription, GPU memory requirements, and failure under poor noise or pause-threshold settings. STT therefore improves the boundary of auditory interaction, but it does not remove the expectation gap. It only makes the first step of the pipeline easier to control.

5.3 Language Model Architecture

The language model architecture was developed as an alignment and awareness problem. Ameca's humanlike embodiment makes fluent speech socially meaningful, while black-box response generation can make the system boundary harder to read when users cannot distinguish generated, remembered, inferred, retrieved, or hallucinated content.

Table 5.1: Speech-to-text solutions used across the EMAH system auditory systems.

Criterion	Google-STT	WhisperX	5A
Deployment	Cloud-based recognition through `recognize_google`	Local GPU-based transcription	Awareness improves when speech processing becomes local and controllable.
Robustness	Easy to integrate, but weaker for accents, empty outputs, and latency variation	More stable for spontaneous speech, but still noise- and threshold-sensitive	Ability depends on whether Ameca can reliably capture the user's utterance.
Interaction risk	Misrecognition can derail ChitT5 and MultiSessionT5 before response generation	Blocking transcription can still delay turn-taking	Anxiety can appear when users experience ASR errors as misunderstanding or incoherence.
System role	Used in ChitT5 and MultiSessionT5	Used in the final Ollama-based system	Alignment improves when the input boundary better matches the dialogue system's needs.

Local and controllable LLM pipelines were therefore prioritised over proprietary cloud APIs. The systems had to run on lab infrastructure for reproducibility, privacy, and stable access during user studies. The architecture also needed fine-grained control over conversational style, memory handling, factual grounding, safety, and integration with the EMAH system control stack, including speech recognition, text-to-speech, facial expression, gesture handling, and Finroc-based robot control.

Table 5.2 situates this design choice within recent work on LLM-powered social robots. Existing systems show that LLMs can make social robots appear more fluent and adaptive, while fluency increases expectations for robustness, privacy, memory, role stability, and repair. The related works therefore frame dialogue ability as expectation management in embodied HRI.

5.3.1 ChitT5: Single-session Dialogue Ability

ChitT5 was the first LLM-based dialogue system developed for Ameca [Ashok 24a]. It supported a single-session icebreaker interaction where Ameca had to sustain short, low-stakes conversation with a first-time interlocutor [Mauliana 23]. A short demonstration of the single-session system is available on YouTube[1].

[1] https://www.youtube.com/watch?v=fXFIrRVQjBM

Table 5.2: Comparison of relevant works on LLM-powered Social Robots

Reference	Model	Social Robot	Dialogue Type	Use-case	Sample
[Billing 23]	OpenAI GPT-3	Pepper & Nao	Open domain	Transform text-based chat to verbal HRI	Demo at HRI'23 (N/A)
[Paplu 23a]	OpenAI GPT-3.5-turbo	Ameca	Open domain	Integrate LLM into Finroc framework	PhD final experiment (N/A)
[Khoo 23]	OpenAI GPT-3 (fine-tuned)	QTrobot	Open domain	Autonomous HRI for wellbeing insights	Older adults (N=12)
[Kim 24a]	OpenAI GPT-3.5	Pepper	Task oriented	Evaluate LLM capability on four tasks	University students (N=32)
[Verhelst 24]	OpenAI GPT-3.5	Furhat	Task oriented	Customized educational content for Spanish learners	School students (N=21)
[Addlesee 24]	Vicuna-13b-v1.5	ARI	Open domain	Multi-party conversation	Hospital memory clinic (N/A)
[Kim 24b]	OpenAI GPT-4o-mini (fine-tuned)	piBo	Open domain	Child-centric intention-aware dialogue	Adult human judgement (N=24)
[Irfan 25]	OpenAI GPT-3.5-turbo	Furhat	Open domain	Identify privacy risk awareness	Adults (N=39)
[Pinto-Bernal 25]	OpenAI ChatGPT-3.5	Pepper	Open domain	Memory-aware HRI with elderly	University students (N=43) Older adults (N=7)
[Kang 25b]	OpenAI GPT-4o-mini	Nadine	Open domain	Mitigating uncanny perceptions via LLM	University Students (N=68)



How is an LLM-powered social humanoid robot rated on its communication and conversational abilities?

ChitT5 marks the first clear ability step in the auditory pipeline. Ameca's body already created the anthropomorphic expectation through face, gaze, voice, and moving lips; ChitT5 added open-domain small talk to that body. The system made Ameca appear conversationally capable, while raising expectations for topic continuity, self-consistency, and socially fitting responses.

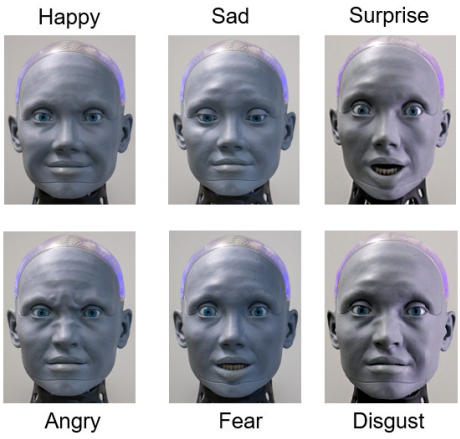
Table 5.3: Two analysis components of the ChitT5 study [Ashok 24a]: human evaluation of single-session dialogue ability and transcript-level inspection of interaction breakdowns.

Goal	**System**	**Sample**	**Finding**	**5A**
Evaluate Ameca as a first-session conversationalist	Fine-tuned Flan-T5-Base with STT, response filtering, and emotion-triggered facial expressions	N=22; six-minute icebreaker interaction	Global communication ratings stayed at the scale midpoint; topic maintenance and follow-up were weakest.	*Ability*: LLM dialogue made Ameca appear conversational, but coherence remained limited.
Inspect interaction breakdowns	Transcript-level analysis of dialogue flow, alignment, ASR errors, and social violations	Same interaction logs	Breakdowns appeared mainly as interaction drift, semantic mismatch, identity confusion, STT errors, and premature endings.	*3A issue*: apparent ability created uncertainty about what Ameca understood, remembered, or meant.

ChitT5 used a fine-tuned Flan-T5-Base model for turn-based response generation [Raffel 20, Chung 24]. The model was trained on DailyDialog, CoQA, and a handcrafted small-talk dataset, then integrated into EMAH system with Google SpeechRecognition, response filtering, and robot expression control [Li 17, Reddy 19]. A DistilRoBERTa emotion classifier mapped each generated response to one of Ekman's six emotion labels, which triggered Ameca's facial expressions through the pose handler [Ekman 87]. Figure 5.2 shows the emotion-expression mapping and HRI setup. The mapping made ChitT5 visibly affective, while introducing mismatch risk when displayed expression did not fit spoken content. Appendix A.11 gives the technical and extended results.

The user study included a pilot test and a main robot study. In the pilot, five participants rated text-only conversations with T5-Base and Flan-T5-Base. T5-Base received $M = 3.0$, $SD = 0.71$, while Flan-T5-Base received $M = 3.8$, $SD = 0.45$. Flan-T5-Base was therefore selected for the robot study. In the main study, 22 participants interacted with Ameca for an icebreaker game lasting approximately six minutes.

Figure 5.3 shows the CSRS item ratings and clarifies what ChitT5 added to Ameca's apparent dialogue ability. The system asked questions, spoke about the user, spoke about itself, agreed, and expressed simple opinions during a first encounter [Ashok 24a]. These behaviours correspond to ability in the 5A framework because Ameca no longer appeared limited to scripted turns. The weaker ratings for topic initiation, topic flow, and follow-up comments show that ChitT5 could produce local conversational turns, while shared interactional context remained unstable. Ameca's body, voice, gaze, and moving lips made



(a) Facial expressions used for emotion-triggered output, validated in prior Ameca work [Berns 24].



(b) HRI setup for the ChitT5 human evaluation.

Figure 5.2: ChitT5 expression mapping and human-evaluation setup.

users expect dialogue that was socially coherent, not only grammatically possible. ChitT5 increased responsiveness, while the system still lacked stable alignment between apparent conversational ability and reliable conversational competence.

The Godspeed results help place this mismatch. Participants evaluated Ameca as likeable and moderately intelligent, anthropomorphic, and animated, yet most still classified the robot as a stranger instead of a friend. ChitT5 increased perceived ability, while Ameca's humanlike embodiment, dialogue behaviour, and social role did not fit strongly enough to make the robot a settled social partner. Detailed Godspeed values and extended CSRS results are reported in Appendix A.11. Since the ChitT5 study used post-interaction data only, the results show post-interaction evaluation rather than change in perceived anthropomorphism, ability, or uncertainty.

The first mismatch was conversational continuity. ChitT5 made Ameca appear able to open a social exchange, ask questions, and respond in turn, while interaction logs showed drift when the robot failed to answer the question, shifted topic abruptly, or made the exchange feel one-sided. Ameca looked and sounded like a partner in a shared conversation, while the system could not always sustain the context it had created.

The second mismatch was identity stability. Conflicting self-references, inconsistent names, and unclear gender or role cues weakened social coherence. Ameca's body created the expectation that voice, name, identity, and role belonged to one agent; the dialogue system then failed to maintain that identity consistently.

The third mismatch was auditory and affective grounding. STT distortions could derail the dialogue before response generation, and emotion-label hallucinations could make Ameca display an angry facial expression during otherwise neutral dialogue. The failure became embodied because Ameca speaks through a face, turning mismatched words, hearing errors, and expressions into cues for how users read the robot.

ChitT5 showed that first-session responsiveness was insufficient for Ameca. The system made the robot more conversational, while post-interaction evidence still showed limits

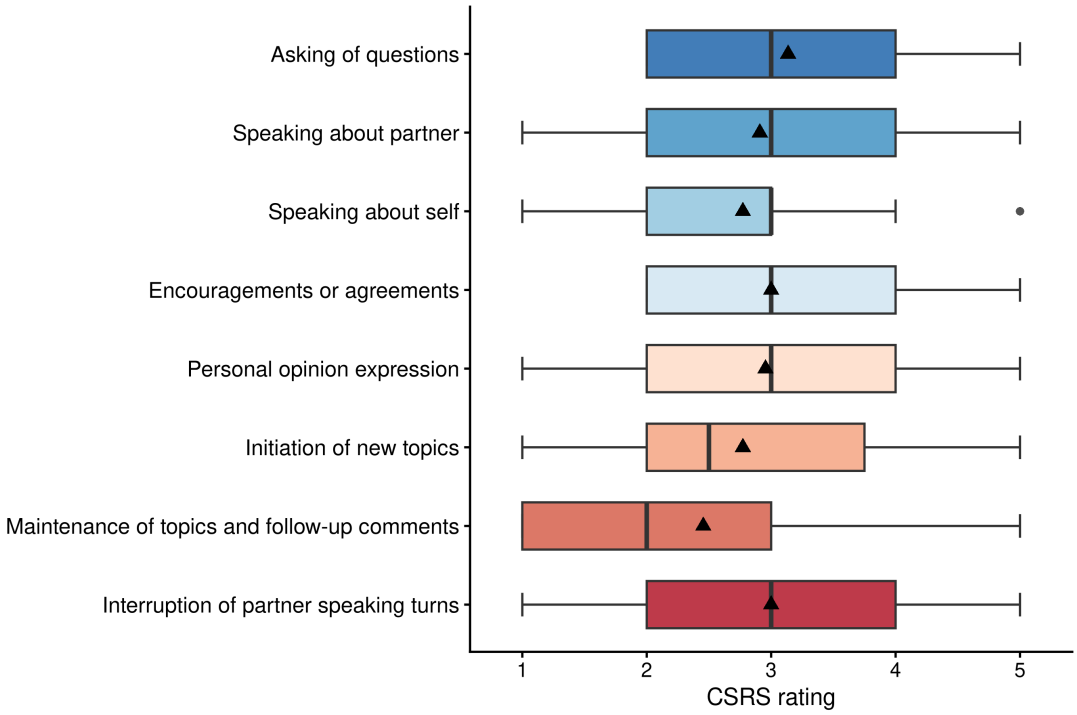


Figure 5.3: Conversational Skill Rating Scale item ratings for Ameca in the ChitT5 study.

in continuity, identity stability, factual and self-related grounding, and expression timing. The next step therefore required context-aware interaction, multi-session memory, and factual retrieval.

5.3.2 MultiSessionT5: Repeated Interaction, Memory, and Role Stability

MultiSessionT5 was the second LLM-based dialogue system developed for Ameca

[Mauliana 25]. It extended ChitT5 from a single icebreaker encounter to repeated interaction across four sessions [Mauliana 24]. The expectation problem changed from sounding responsive in one short conversation to appearing familiar, emotionally responsive, and able to remember the user across time.


How does interaction with an LLM-powered social humanoid robot across multiple sessions affect its role perception as a companion?

MultiSessionT5 marks the second ability step in the auditory pipeline. The system added emotion-aware response generation, user-specific memory, and factual retrieval. These functions made Ameca appear less like a one-shot demonstrator and more like a continuing conversational agent. Memory then raised a stronger expectation than single-session fluency, since remembered details had to be correct, affect had to fit the turn, and identity had to remain stable across sessions.

Table 5.4: MultiSessionT5 as repeated dialogue ability in Ameca.

Goal	System	Sample	Finding	5A
Evaluate Ameca as a repeated-interaction companion [Mauliana 25]	Flan-T5-Large with emotion labels, user memory, session summaries, and factual retrieval	N=15 recruited; N=13 completed four sessions	Engagement stayed stable with a small descriptive increase; participants rated Ameca higher for light conversation than for emotional closeness or serious personal disclosure.	*Ability*: memory and emotion made Ameca appear more companion-like, but the strongest support remained limited to low-stakes conversation.
Inspect repeated-interaction breakdowns [Mauliana 25]	Transcript-level analysis of memory use, role consistency, topic flow, STT errors, and endings	Four-session interaction logs	Breakdowns appeared mainly as memory lapses, role confusion, repetition, weak repair, abrupt disengagement, and topic drift.	*3A issue*: repeated ability raised expectations of continuity, but unstable memory and identity made the robot harder to read.

MultiSessionT5 shifted expectation management from single-turn responsiveness to repeated-session continuity. The system used Flan-T5-Large as a multi-task model for response generation, emotion classification, and background-information detection [Chung 24, Crawshaw 20]. It was trained with DailyDialog, EmpatheticDialogue, Topical-Chat, and Blended Skill Talk [Li 17, Rashkin 19, Gopalakrishnan 23, Smith 20]. The architecture included retrieval-augmented generation for factual questions about the robotics lab, user-specific JSON memory, session summarisation, and gesture output through the EMAH system control stack [Lewis 20, Gliwa 19]. Detailed training, datasets, prompts, automatic metrics, inference settings, and routing rules are reported in Appendix A.12. Figure 5.4 shows where memory, retrieval, emotion labels, and embodied output entered the dialogue pipeline.

The human evaluation used a controlled four-session HRI study with N = 13 participants completing all sessions. Before each session, participants rated their current state and baseline perception of the robot. After each session, they rated engagement, usability, safety, relationship perception, and role-related items. The final session included additional expectation and reflection measures. The full procedure, questionnaire structure, demographics, and statistical tables are reported in Appendix A.12.

Pre-session measures showed that participants did not approach Ameca as a neutral machine. Ameca was already evaluated through social categories before repeated interaction began. Sociability was high before interaction and remained high after the fourth session. Disturbance showed a small descriptive decrease from pre-interaction to post-interaction, without statistical significance. The result remains descriptive and shows that repeated interaction with this system did not increase reported disturbance in this small sample.

Engagement remained stable across the four sessions, with a small descriptive increase

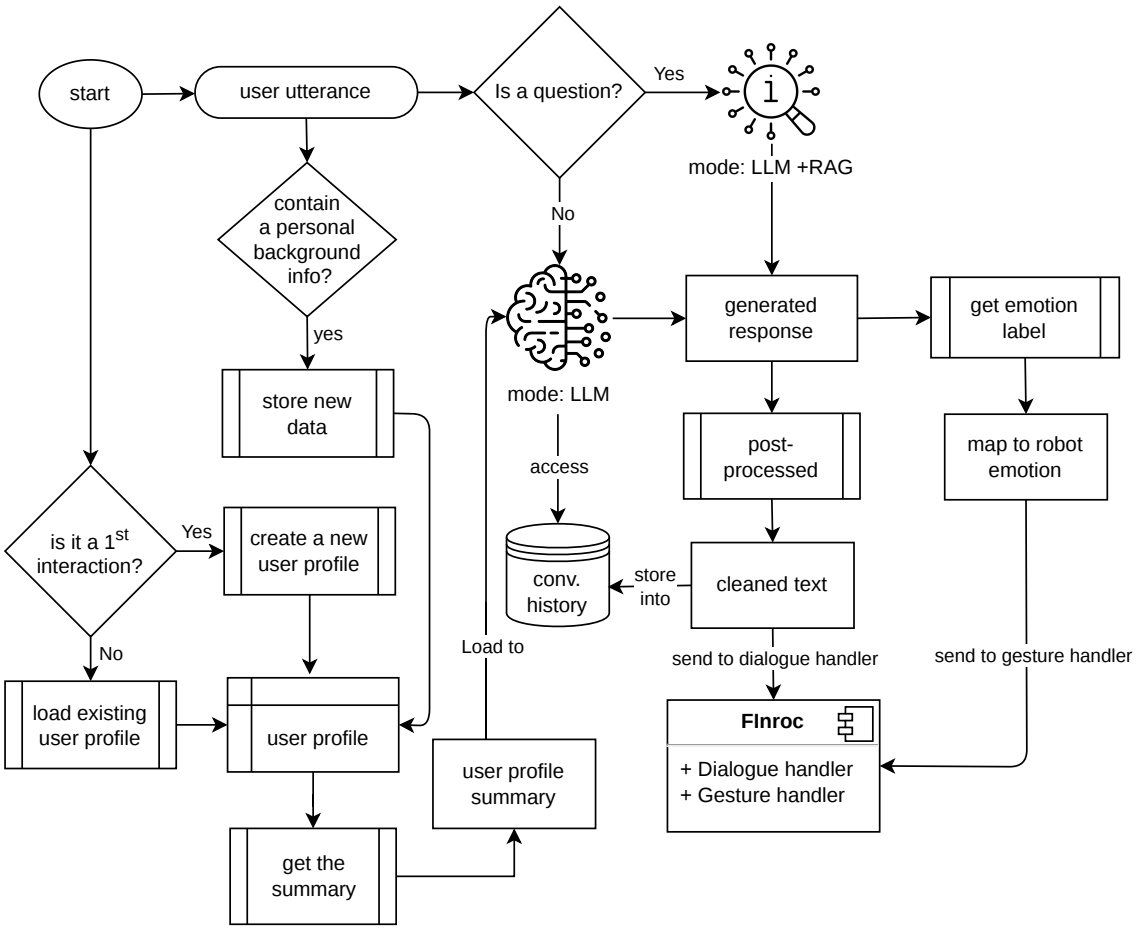


Figure 5.4: MultiSessionT5 workflow for repeated dialogue, memory, retrieval, and gesture output.

from Session 1 to Session 3 and a similar level in Session 4. Figure 5.5 shows engagement ratings across the four sessions. The Friedman test did not indicate a significant change across sessions, $\chi^2(3) = 3.61$, $p = .307$. The result does not show increasing engagement, but it indicates that repeated interaction was tolerated and that the robot remained usable across sessions without a visible engagement drop.

Expectation ratings clarified the companion role that Ameca could plausibly support. Participants rated Ameca higher for light social conversation, such as talking about hobbies and being a conversation partner, than for advice, emotional connection, or serious personal matters. Figure 5.6 shows the expectation ratings related to Ameca as a conversation partner. Repeated LLM-based interaction supported a bounded low-stakes companion role, while the data did not support deeper expectations of emotional closeness or personal disclosure.

Role-bounded companionship is central for alignment. MultiSessionT5 gave Ameca apparent memory and emotional responsiveness, while the safest social role remained light conversation with a familiar robot. Framing the system as emotionally close without stable memory, repair, privacy clarity, and identity consistency would reduce readability and increase misleading capability impressions.

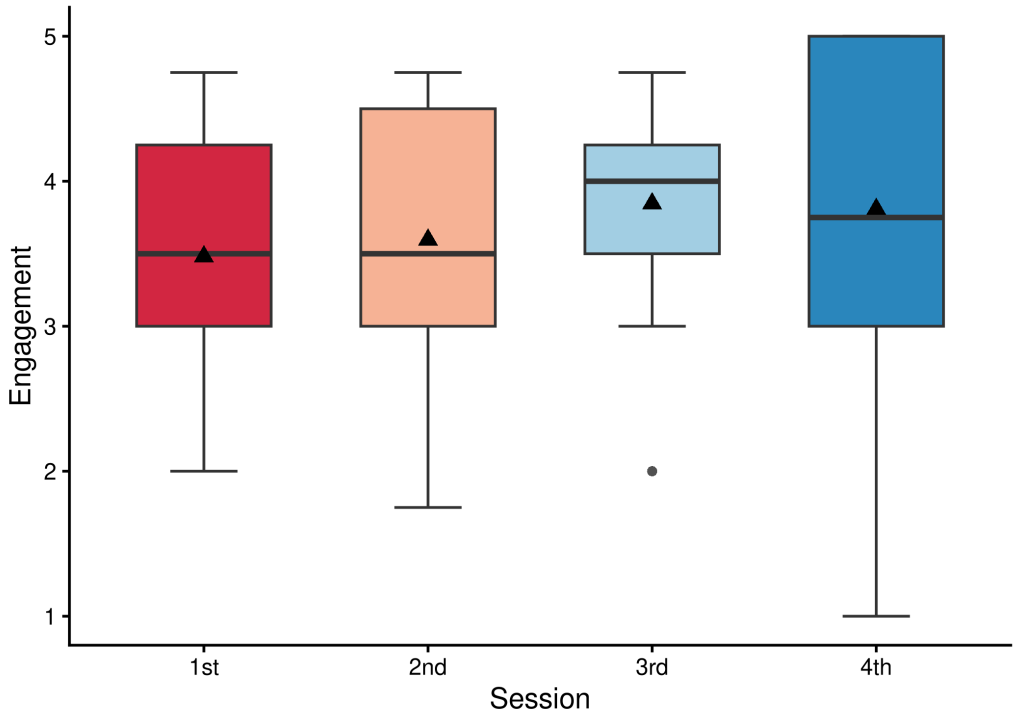


Figure 5.5: Engagement ratings across four MultiSessionT5 interactions.

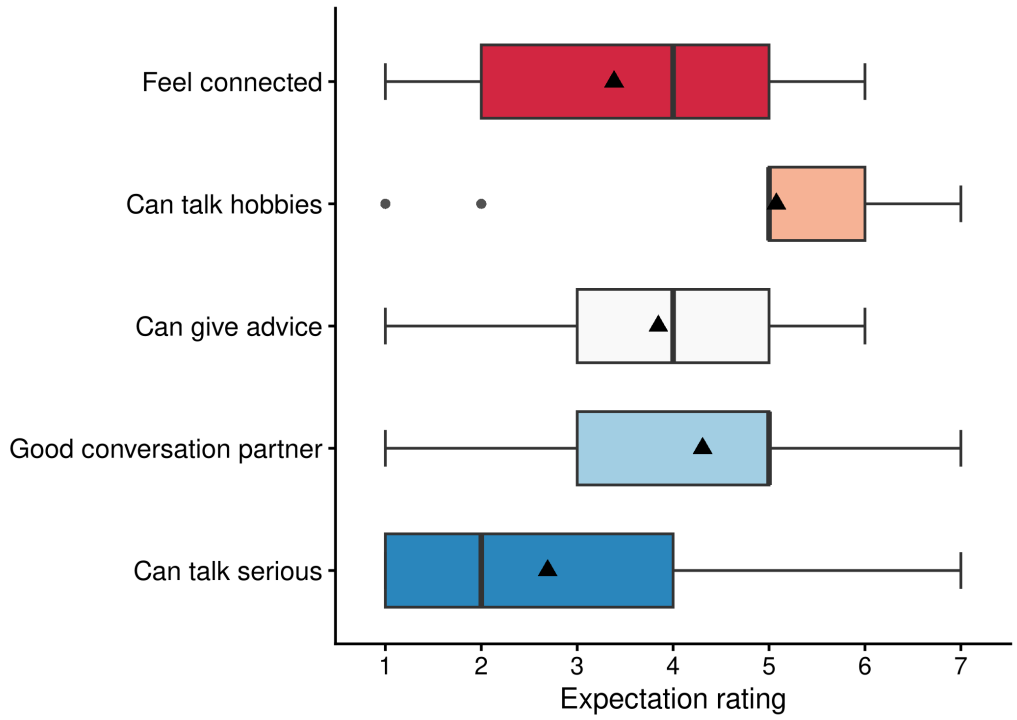


Figure 5.6: Conversation-partner expectation ratings after repeated interaction with Ameca.

Transcript analysis showed where the expectation gap remained. Repeated interaction did not solve the limitations seen in ChitT5; it made some more visible. Memory references became risky when vague, wrong, or unfinished. Repetition became more salient as users heard similar phrases across sessions. Identity errors became more disruptive because Ameca was no longer a first-time stranger. Alternation between robot-like and humanlike self-descriptions affected awareness by making it unclear what Ameca knew, remembered, or believed about itself.

MultiSessionT5 shifted the main mismatch from first-session coherence to repeated-session continuity. ChitT5 showed that fluent local turns were insufficient. MultiSessionT5 showed that memory had to be stable, bounded, and role-consistent. Session-level memory and reduced affect mismatch made Ameca appear more familiar, while transcript patterns still showed vague memory references, repetition, unstable self-descriptions, unsupported topic shifts, and abrupt endings. The study therefore remains descriptive, showing how repeated interaction made continuity, role stability, and memory accuracy more central to evaluation.

Added ability increased the need for alignment. Table 5.5 summarizes the technical progression from ChitT5 to MultiSessionT5 and its 5A relevance. Memory, emotion labels, retrieval, and self-presentation supported Ameca only when they made the robot easier to interpret. Instability in these components weakened the fit between humanlike presentation and actual dialogue competence, motivating the final Ollama-based system with local processing, stronger prompt boundaries, retrieval control, and explicit robot identity.

5.3.3 Open-Source LLM Integration with Ollama

The final auditory system consolidated ChitT5 and MultiSessionT5 into a locally hosted open-source LLM pipeline using Ollama[2]. The design changed from model-level response generation to a locally controlled pipeline where prompt design, robot identity, capability boundaries, memory, retrieval, and failure handling were implemented as expectation-management mechanisms [Ashok 26].

How can a local LLM pipeline make Ameca's dialogue behaviour more bounded, transparent, and aligned with its robot identity?

The Ollama system targets alignment and awareness. ChitT5 increased apparent conversational ability, and MultiSessionT5 added apparent memory, emotional responsiveness, and factual grounding. The Ollama system kept these functions within a stronger behavioural boundary. Ameca's language behaviour had to match the robot users were actually interacting with, namely a humanoid research robot with limited perception, imperfect speech recognition, bounded memory, no human feelings, and explicit hardware constraints.

The system begins by loading or creating a user profile through a spoken keyword. It retrieves the stored conversation history and session summary, then passes the current utterance to a locally hosted LLM through a stable system prompt. The prompt defines Ameca as a robot and specifies its physical body, sensing limits, speech-recognition limits,

[2] https://ollama.com/

Table 5.5: Technical shift from ChitT5 to MultiSessionT5 and its 5A relevance.

Category	ChitT5	MultiSessionT5	5A
Dialogue scope	Single-session ice-breaker dialogue	Repeated interaction across four sessions	*Ability*: Ameca appeared less scripted and more socially persistent.
Model design	Fine-tuned T5/Flan-T5 for response generation	Multi-task Flan-T5-Large for response generation, emotion labels, and background detection	*Ability*: added apparent emotional and personal-context handling.
Context handling	Short turn history	Session summaries and user-specific memory	*Alignment*: memory supported familiarity, but raised expectations of continuity.
Grounding	No factual retrieval	Retrieval support for lab-related factual questions	*Awareness*: retrieval made system boundaries more controllable, but did not remove drift.
Embodied affect	Generated emotion labels triggered facial expressions	Emotion labels supported gesture/expression control with reduced affect mismatch risk	*Alignment*: affect output became less disruptive, but still depended on correct interpretation.
Main limitation	Weak topic flow, follow-up, identity stability, and STT sensitivity	Memory lapses, repetition, role confusion, unsupported claims, and abrupt endings	*3A issue*: stronger apparent ability made inconsistencies more visible.

hardware constraints, privacy rules, and failure responses. This directly addresses the identity instability observed in ChitT5 and MultiSessionT5. The full prompt is provided in Appendix A.13.1. The prompt-bounded Ameca identity also connects to artificial identity design, where coherent robot identity depends on recognisability, behavioural consistency, memory, role clarity, artificiality, and transparency across interaction contexts [Bransky 26].

At the time of writing, the Ollama-based system has been implemented and briefly tested on Ameca, while the full ablation study is still in progress. The current implementation is therefore a system-integration contribution rather than a completed user-evaluation result. The codebase provides an initial open-source framework for local LLM control, user profiling, memory handling, emotion-label generation, optional retrieval, and robot gesture output[3]. Ongoing work extends the local Ollama pipeline by replacing blocking WhisperX transcription with lower-latency Faster-Whisper, adding Silero VAD for speech-segment detection, using Llama 3:8B for emotion classification and response generation, storing local JSON user profiles, and developing RAG support [Esiagu 26].

[3] https://github.com/aashok94/llm_powered_social_robot

Table 5.6: Ollama-based local LLM integration as a bounded dialogue system for Ameca.

Goal	**System**	**Design**	**Expected role**	**5A**
Consolidate the prior auditory systems into a local open-source framework	Ollama-served local LLM with memory, retrieval, emotion labels, and gesture output	Prompt-bounded Ameca identity, capability limits, failure protocol, privacy rules, and ethical red lines	Dialogue behaviour becomes more controllable, reproducible, and less dependent on proprietary APIs	*Alignment*: the robot's speech is constrained to match its real role and system limits.
Reduce misleading persona shifts in LLM-based HRI	Structured system prompt based on HRI prompting guidelines [Ashok 26]	Explicit identity, transparency, task, capability boundary, user adaptation, and failure handling	Ameca is instructed not to claim human identity, unsupported perception, or unavailable capabilities	*Awareness*: users receive clearer signals about what the robot can and cannot do.

The implemented system ran on an external robot PC with an NVIDIA GeForce RTX 4090. Local open-source LLM deployment is therefore technically feasible within the EMAH system control stack, while comparative HRI performance evidence is still pending. The contribution for this thesis is the integration of earlier dialogue abilities into a more transparent and locally controlled architecture, where Ameca's identity, capability boundaries, memory use, retrieval access, and failure responses are explicitly constrained.

5.4 Text-to-Speech (TTS)

Text-to-speech defines the final auditory boundary before Ameca's dialogue becomes embodied. Generated text is heard as language while also being produced through a humanlike face with synchronised lips, voice, gaze, and gesture. TTS therefore affects both ability and alignment, since a fluent response can still fail when voice, timing, or mouth movement does not fit the body.

Earlier lab work examined whether vocal cues alone could convey affective states in social robots [Hirth 12a]. Based on work showing that emotions can be expressed through speech disturbances and prosodic variation, including pitch and tempo [Knapp 72], Hirth et al. systematically varied speed, pitch, volume, and timbre in synthetic speech. The system used the AmiASR framework and a Loquendo TTS synthesizer, and tested twelve speech-parameter profiles using the sentence "Das Quadrat liegt richtig" from a tangram-game scenario [Hirth 12a]. In a user study with participants unfamiliar with ROMAN[4], listeners rated the perceived presence of six emotions. Emotion recognition was difficult, but each target emotion had at least one speech profile with more than 50% listener agreement. The study provides an early controlled reference for linking robot vocal parameters to perceived affect.

[4] https://rrlab.cs.rptu.de/en/robots/more-robots/roman

The present work did not extend this line through custom emotional speech synthesis on Ameca, since alignment required reliable facial articulation. Alternative TTS engines could generate audio, but they did not reliably drive Ameca's mouth movement. Speech played from the robot without matching lip movement would add audiovisual mismatch to an already highly anthropomorphic system. Reliable lip-synchronised speech was therefore prioritised over fine-grained emotional prosody.

Ameca uses a licensed Acapela[5] TTS engine through the robot's Tritium[6] REST interface. The Finroc module `mTextToSpeechHandler` receives text from the dialogue subsystem, filters characters for Acapela compatibility, selects a predefined voice, and forwards the request to Ameca's Acapela server. The available presets include UK English voices such as Graham and Lucy and German voices such as Julia and Klaus. These choices affect social interpretation because timbre, gender cues, and accent can change how Ameca's role is read. Implementation details are provided in Appendix A.14, including the TTS pipeline summarized in Algorithm A.3.

The TTS handler converts non-ASCII characters, including German umlauts, into Acapela-compatible sequences before publishing the cleaned text for logging and interface mirroring. Tritium then relays the text to Acapela, which synthesises speech and drives phoneme-aligned mouth movements. These visible mouth shapes, or visemes, provide the visual counterpart to phonemes and allow Ameca's lips to move in synchrony with speech. Synchronisation is central for alignment, since voice and face must appear to belong to the same embodied agent.

Voice cannot be handled as simple playback in Ameca. Chapter 4, Section 50, showed that voice interacts with body, perceived gender, facial expression, timing, and social role [Jafarova 25, Puertas Céspedes 25, Beck 26]. The auditory pipeline therefore used reliable lip synchronisation and bounded voice presets as alignment mechanisms, with voice output constrained by the same expectation-management logic as dialogue generation.

Acapela produced low-latency speech with reliable articulation across the available English and German voices. Its limitation is expressive control. The system does not provide the fine-grained access to prosody or phoneme timing needed for custom emotional speech generation. This limitation was acceptable for the thesis because Ameca already carries strong anthropomorphic expectations through face, gaze, dialogue, and gestures. Unstable affective prosody would increase mismatch risk, so reliable timing, clear articulation, and lip synchronisation were selected as the safer design choice.

Ongoing work examines how voice gender cues affect perceived agency and judgements of Ameca's facial expressions. Emotion-aware custom TTS, including systems such as Emojivoice [Tuttösí 25a], remains a future extension for the auditory pipeline.

5.5 Discussion

Ameca's auditory system created expectation mismatch through the coupling of listening, memory, dialogue generation, voice, facial movement, and response timing. Users did not experience speech-recognition errors, weak memory references, or delayed answers as

[5] https://www.acapela-group.com/
[6] https://engineeredarts.com/software/tritium

isolated module failures. The embodied output made Ameca appear to misunderstand, drift, hesitate, or contradict itself.

The speech pipeline exposed the first technical boundary. Google STT supported short interactions, while accent sensitivity, empty transcriptions, and latency variation could derail dialogue before response generation. WhisperX shifted the boundary toward local and controllable processing, while turn-taking delay and noisy input remained open risks. TTS created the output boundary, where generated text had to become speech through a humanlike face. Reliable lip synchronisation and bounded voice presets were safer for Ameca than experimental emotional prosody, since mismatched voice, mouth movement, and affect would add another source of unreadability.

The dialogue systems showed how added ability increased the need for alignment. ChitT5 made Ameca more fluent in a first encounter, while weak topic continuity, identity instability, STT errors, and affective mismatch limited coherence. MultiSessionT5 added memory, emotion handling, and retrieval, while repeated interaction made continuity harder to maintain. Remembered details could support familiarity, while vague memory, repetition, unsupported claims, or unstable self-description made Ameca harder to place. Fluency, memory, and affective output made users expect coherence that the system could not always sustain.

The Ollama-based system responded through local, prompt-bounded control. Its contribution is an inspectable robot-dialogue architecture with local inference, explicit robot identity, capability boundaries, retrieval control, privacy constraints, and failure protocols. A highly humanlike robot needs limits visible in behaviour, not hidden only in implementation.

Auditory interaction therefore provides the speech and memory layer of expectation management. Speech, memory, and voice make Ameca easier to interact with when they make the robot more readable. The design objective is a dialogue system whose apparent ability, actual limits, and robot identity remain aligned during embodied interaction.

6. Visual Perception and Gaze Management

Chapter 3 defined the anthropomorphic expectation gap as the mismatch between what a highly humanlike robot appears to promise and what the implemented system can reliably deliver. Visual perception and gaze management examine this gap at the level where eyes, camera input, visual recognition, scene answers, speaker selection, and gaze behaviour become one social signal in Ameca.

Ameca's eyes function as expectation-setting interfaces, since they suggest that the robot can see, attend, understand spatial context, and distribute attention socially before the cameras and perception modules have shown what they can reliably support. The eye cameras and chest-mounted RGB-D camera made gesture recognition, object detection, scene awareness, visual question answering, face tracking, and gaze control technically possible, while each added visual function raised the standard for the next behaviour. A detected gesture can imply intention understanding, a scene answer can imply situated perception, a gaze shift toward a speaker can imply social attention, and reduced gaze toward one person can be read as exclusion, disregard, or interactional failure.

This chapter uses visual perception as a case of ability alignment, where Ameca's visual and gaze behaviour had to match the expectations created by its humanlike eyes, face, and body. The chapter first situates the work in relation to social perception, scene awareness, visual question answering, gaze control, active-speaker tracking, joint attention, and behaviour generation. It then moves from visual-percept modules, to VQA, to gaze-management work, and finally to the speech-gaze Hero selection system for focal-interlocutor selection in live multi-party interaction.

6.1 Overview: Visual Perception Pipeline as Ability Alignment

Visual perception for social robots has to connect detection with interaction, since eyes, head orientation, gaze shifts, and scene-directed answers make users infer that the robot

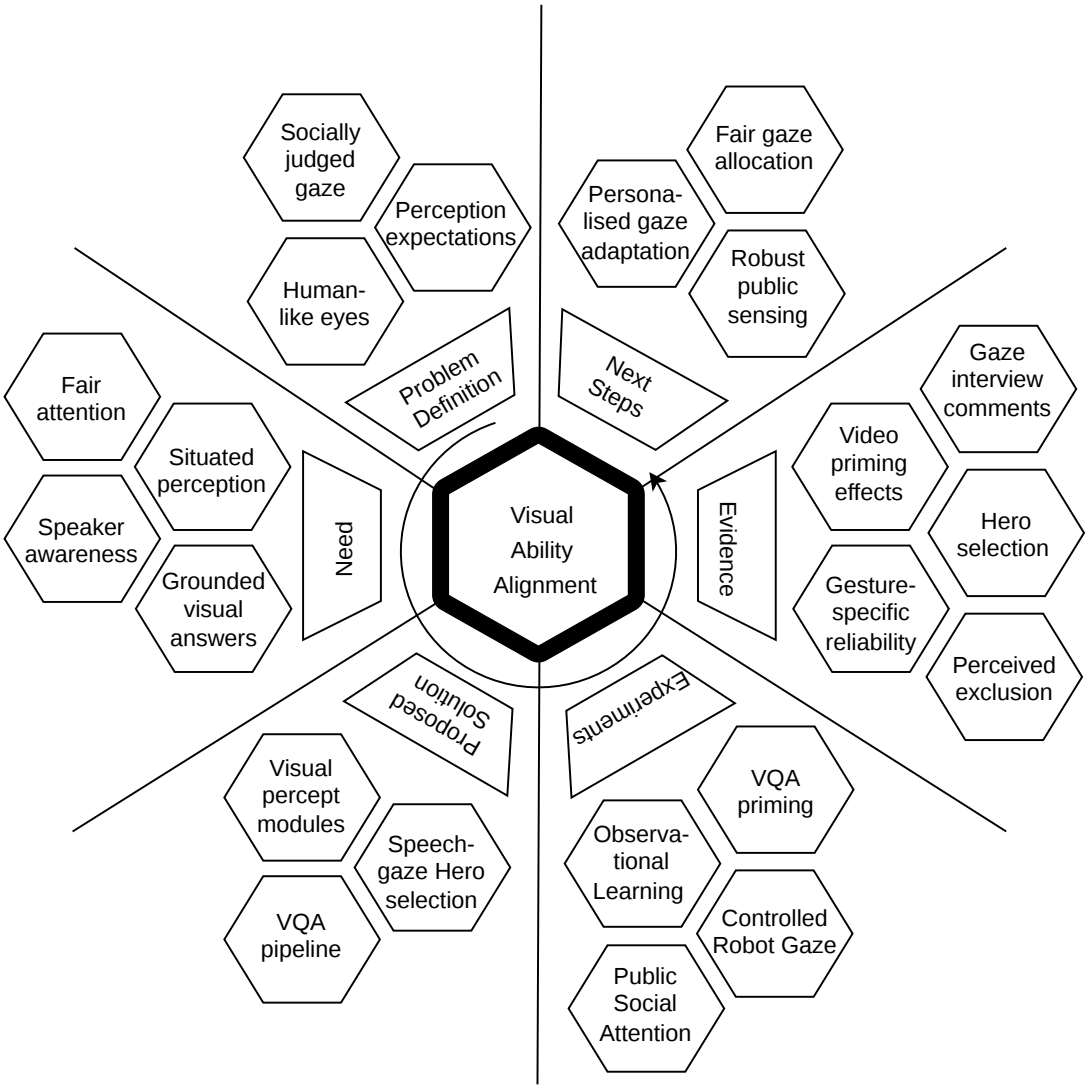


Figure 6.1: Overview of visual perception and gaze management as expectation alignment in Ameca.

can see, attend, and understand the shared environment. Perception pipelines therefore need interactional adaptability in addition to input-level accuracy [Sciutti 18]. Zaraki et al. [Zaraki 16] show this at system level with a Social Perception System for Zeno that detects socially relevant cues such as facial expressions and gestures while also considering real-world context. For Ameca, this issue becomes sharper because its humanlike eyes create an expectation of visual awareness before any perception module is evaluated.

Table 6.1 situates this chapter within work on social perception, scene awareness, gaze management, active-speaker tracking, joint attention, and behaviour generation.

The closest prior system in the Ameca setting is the scene-awareness work by Paplu et al. [Paplu 23b], which fused low-level audio-visual cues, including eye contact, facial expression, interruptions, paralinguistic cues, object detection, and attire detection, to support higher-level social perception and scene-aware dialogue. The system used modules such as an eye-contact detector, ResNet50-based facial-expression recognition, YOLOv5 object detection, and rule-based plus caption-based speech generation. Its evaluation

Table 6.1: Comparison of relevant works on visual perception, gaze, and behaviour generation for social robots

Reference	Model/ Method	Social Robot	Visual Function / Use-case	Sample/ Evaluation
[Zaraki 16]	Social Perception System	Zeno	Visual percepts Social cues	System evaluation
[Paplu 23b]	Eye-contact CNN ResNet50 YOLOv5	Ameca/ Emah	Scene awareness Contextual speech	Video rating study (N=73)
[Okuno 02]	Audio-visual stream association	SIG	Gaze management Talker tracking	Receptionist and companion scenarios
[Qian 22]	Cross-modal attentive fusion	KINOVA platform	Active-speaker tracking	AVRI dataset
[Pan 20]	Heuristic gaze algorithm	Humanoid bust	Gaze management Turn-taking	Multi-party interaction demo
[Gillet 22]	Imitation learning Batch RL	Furhat	Gaze management Turn allocation	User study (N=50)
[García-Martínez 25]	Responsive joint attention system	Mini	Joint attention Distraction response	User study (N=40)
[Wei 26]	VLM gaze reasoning VQ-VAE	Humanoid robot head	Gaze reasoning Target selection	Scenario videos (N=76)
[Roselli 25]	Direct-gaze duration manipulation	iCub	Gaze perception Eye contact	User study (N=41)
[Huang 13]	Robot Behavior Toolkit	Wakamaru/ PR2	Gaze timing Referent grounding	Two lab studies (N=32;24)
[Li 24]	Seq2seq co-speech gesture model	NAO Pepper ARI	Behaviour generation Co-speech gestures	User study (N=23)
[Ko 22]	Kinect LSTM rule-based adaptation	Pepper	Adaptive behaviour Pose response	AIR-Act2Act + user rating (N=25)
[Pekçetin 26]	Standardized robot action database	Pepper	Action perception Robot gestures	HR-ACT database (N=438)

showed that environmental awareness improved perceived realism in video ratings, linking visual perception to grounded and interpretable robot response.

Gaze adds a second layer to the same expectation problem, since a robot that loses the person, looks at the wrong target, or fails to track the current speaker breaks social attention instead of producing only a technical error. Zhong et al. [Zhong 24] addressed this problem for Pepper by improving mobile-target tracking under occlusion through the last bounding box to anticipate likely movement. Okuno et al. [Okuno 02] treated audio-visual attention as robot behaviour by associating auditory and visual streams for talker tracking. Pan et al. [Pan 20] developed a multi-party gaze system for a humanoid Audio-Animatronics bust that sensed people, selected people of interest, and generated gaze behaviours. Stefanov et al. [Stefanov 19] modelled gaze targets in multiparty interaction from speech activity, eye-gaze direction, and head orientation. Recent work extends this direction by linking gaze to active-speaker tracking, joint attention, and VLM-supported gaze reasoning [Qian 22, García-Martínez 25, Wei 26].

For Ameca, this literature defined the technical direction and the evaluation problem, since the anthropomorphic form already made users expect eyes to indicate seeing. Gesture recognition supported visible user input, VQA supported scene-grounded answers, and gaze control supported socially readable attention. The rest of this chapter follows the path from basic visual percepts, to visually grounded answers, to gaze-based attention allocation.

6.2 Visual Percepts for Social Interaction

The first visual systems developed for Ameca addressed the mismatch between a robot that looked able to see and an interaction pipeline that still depended on whether visual cues could be detected, interpreted, and connected to behaviour. These systems were developed as separate attempts to make Ameca's apparent visual ability more reliable through gesture recognition, self-other disambiguation, shared task perception, and coarse human-attribute estimation.

Table 6.2 summarizes these visual-percept threads. The details of datasets, model choices, implementation pipelines, and extended study results are reported in Appendix A.15.

Dynamic hand gesture recognition was the first attempt to make visual input usable as direct interaction input. Earlier RRLab work had already explored dynamic gestures with the Robothespian robot ROBIN [Zafar 16, Zafar 17]. The Ameca/Emah version revisited this direction with video-based deep learning and real-time robot integration [Islam 23]. Temporal features improved classification, and the LSTM outperformed the simpler ANN. The HRI result remained gesture-dependent. Participants rated reaction time positively, while recognition quality varied by gesture. Visually distinct gestures worked better than subtle or circular gestures, showing that camera-based input supports interaction only when the gesture vocabulary fits what the system can reliably distinguish.

The second visual-percept thread reframed "self-awareness" as deployable visual disambiguation [Huang 24]. The system tested whether Emah could distinguish its own RRLab configuration, the standard Ameca robot, nearby people, and everyday objects. This distinction was necessary because a humanoid robot that confuses itself, another robot, or a person is visually unstable before higher social reasoning begins. The detector reached

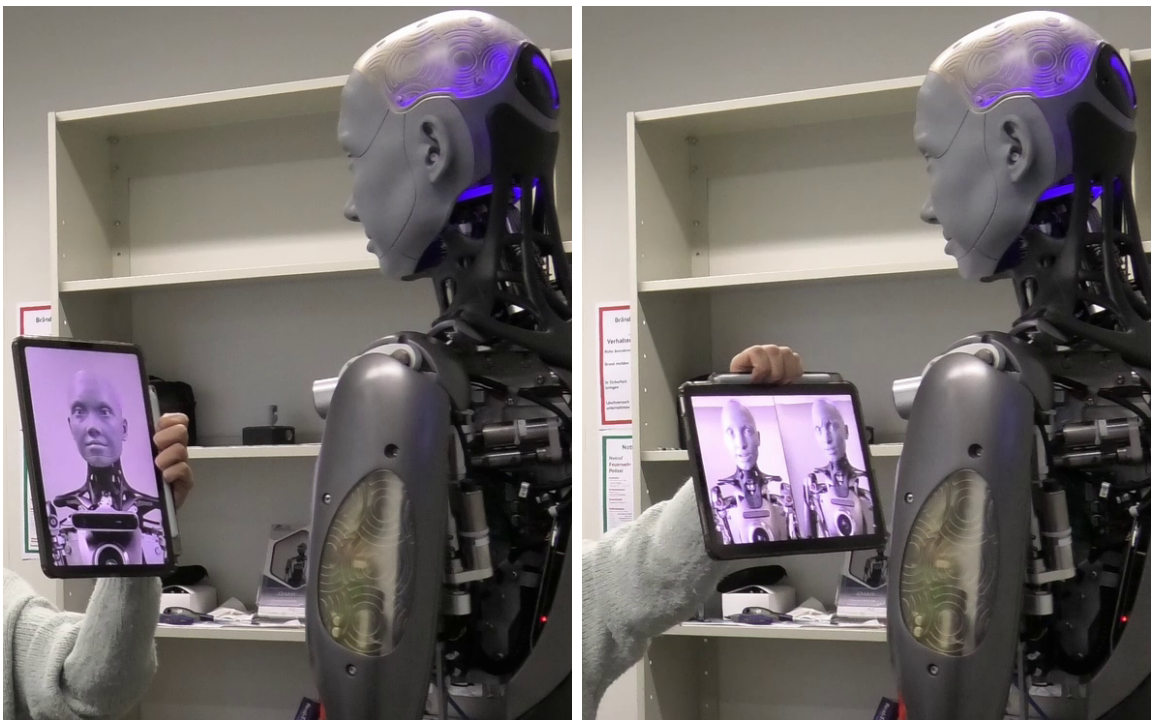
Table 6.2: Visual percepts developed for Ameca/Emah as early visual-ability alignment.

Goal	System	Sample	Finding	5A
Recognise dynamic hand gestures [Islam 23]	ResNet-18 3D feature extraction with ANN and LSTM classifiers	N=11; emergency gesture scenario	LSTM reached 94.7% validation accuracy; live ratings were strongest for visually distinct gestures such as "Victory", "Like", and "Stop".	*Ability*: gestures made visual input actionable, but reliability depended on gesture type.
Separate Emah, Ameca, people, and objects [Huang 24]	YOLOv10 object detector with custom robot classes [Wang 24a]	Custom image dataset	The detector reached 84.5% accuracy; main errors occurred between Emah, Ameca, and person.	*Awareness*: "self-recognition" was treated as practical visual disambiguation, not cognitive self-awareness.
Support Tangram-based shared task perception [Breininger 23, Akinci 24]	OpenCV colour segmentation, ArUco exploration, gaze and feedback logic	N=20; Tangram Game	Godspeed ratings did not show a significant pre-post change; the system showed that perception needed gaze timing and feedback to become socially meaningful.	*Alignment*: visual cues mattered when linked to joint attention and task assistance.
Adapt behaviour from human attributes [Goede 22, Srivastava 22]	Age/gender estimation, DeepFace, OpenCV, Librosa	Scenario-based interaction studies	Age, gender, and facial expression were used to select scripted response styles.	*Ability*: human attributes supported adaptation, but had to remain cautious and correctable.

84.5% accuracy, with main errors occurring between Emah, Ameca, and person. Figure 6.2 shows self-recognition as a practical visual disambiguation task, where Emah had to distinguish its own RRLab configuration from another Ameca-looking robot using robot-specific visual markers, especially the chest-mounted ZED camera.

The Tangram work moved from isolated detection to shared task perception [Breininger 23, Akinci 24]. The first implementation used Ameca's ZED2 camera and OpenCV-based colour segmentation to detect Tangram pieces on a white table. The later Global Workspace Theory-inspired prototype added visual recognition, speech recognition, situational modelling, action selection, and gaze behaviour. The strongest implementation lesson was that visual perception became interactionally meaningful only when it was tied to gaze, feedback, and task assistance. Ameca had to look at the task, notice when the participant looked up, and offer feedback when requested. Figure 6.3 shows the Tangram setup used for this visually grounded joint-attention prototype.

The final visual-percept thread used human attributes for adaptive behaviour [Goede 22, Srivastava 22]. In the self-introduction scenario, Emah estimated age and gender from



(a) RRLAB Ameca (EMAH system) (b) Gen 1 Ameca

Figure 6.2: Robot self-recognition as a practical visual disambiguation task. The system distinguishes the RRLab EMAH system configuration from the standard Ameca appearance by using robot-specific visual markers, especially the chest-mounted ZED camera [Huang 24].

ZED camera input using a Python age/gender detection pipeline[1]. The estimated age range was exposed through Finroc ports and used inside the XML dialogue logic to select either a formal or informal self-introduction, with matching speech and gesture timing. In the conflict scenario, facial emotion recognition was implemented with DeepFace[2], while OpenCV[3] handled image processing and Librosa[4] supported speech-feature extraction. The recognised facial expression selected between two scripted response styles, one apologetic and submissive, and one more assertive and defensive. Later systems still used DeepFace[5] for facial attribute analysis because it provides age, gender, and facial-expression estimates in one module, although its outputs remained unreliable enough to require cautious use.

These four threads show the expectation-gap problem at different levels, since Ameca's body suggests visual competence before the system has earned that competence through reliable links between perception and behaviour. Gesture recognition made user input visible, object detection and self-recognition reduced ambiguity in the visual field, Tangram perception connected objects, gaze, and feedback in a shared task, and human-attribute recognition connected visual classification to adaptive response selection. These systems supported partial alignment by linking specific visual cues to speech, gaze, feedback, or

[1] https://github.com/smahesh29/Gender-and-Age-Detection
[2] https://github.com/serengil/deepface
[3] https://opencv.org/
[4] https://librosa.org/
[5] https://github.com/serengil/deepface

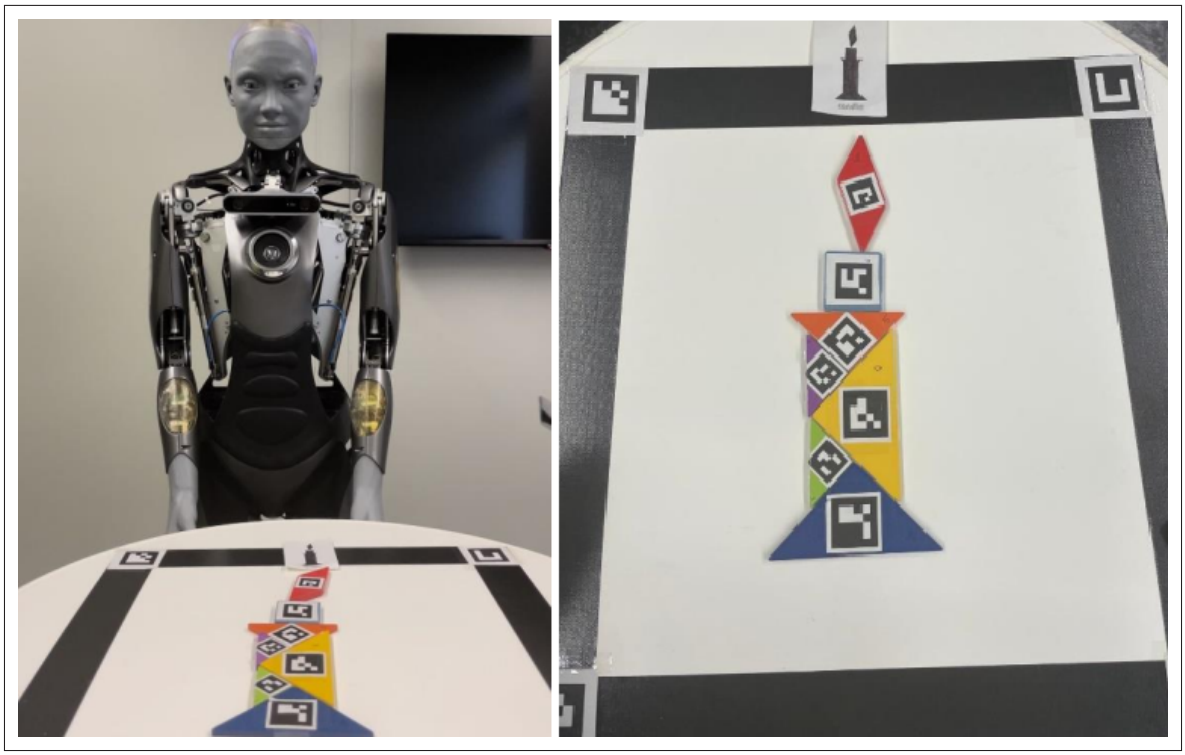


Figure 6.3: Tangram game-playing setup with Ameca [Akinci 24].

behaviour, while the remaining gap showed that a humanlike robot is judged by whether detected cues lead to socially and task-relevant responses.

6.3 Visual Question Answering

Visual question answering tested the expectation gap in its most direct visual form, since Ameca was asked to answer questions about what it could see. A humanoid robot is judged through answer correctness and through the fit between the answer and the perceptual competence suggested by the robot's eyes, face, gaze, and body [Bartneck 04]. VQA therefore brought actual visual ability closer to the anthropomorphic expectation already implied by Ameca's embodiment.

Prior work shows why this expectation is reasonable from the user's perspective. People may spontaneously take a robot's visual perspective, and this perspective-taking can be shaped by robot behaviours such as gaze and reaching [Zhao 15]. A robot camera is therefore not always interpreted as a passive sensor. In face-to-face HRI, users may expect the robot to see from its own position, attend to relevant objects, and answer as a situated interaction partner. Media exposure can add another layer before interaction begins, since public and fictional portrayals of robots can shape expectations of competence, threat, or social exceptionality [Oliveira 24]. The VQA study combined two questions, namely whether Ameca could support visually grounded interaction and whether short positive or negative video exposure changed how users judged that ability before meeting the robot [Rao 24]. A demonstration video of the VQA interaction is available on RRLab's YouTube[6].

[6] https://youtu.be/eY3sdMOWq4s

Table 6.3 summarizes the VQA system and study. The technical comparison between candidate VQA models, example model outputs, priming videos, full measures, descriptive statistics, and question categories are reported in Appendix A.16.

Table 6.3: Two analysis components of the VQA study [Rao 24]: media priming before interaction and live visually grounded interaction with Ameca.

Goal	**System**	**Sample**	**Finding**	**5A**
Test whether media-like exposure shifts expectations before VQA interaction	LLaVA-v1.5-7B integrated with Ameca through Python and Finroc	N=27; three groups; 72 user questions	Positive and negative videos shifted immediate ratings of anthropomorphism and perceived intelligence, but these effects were largely reduced after direct HRI.	*Ability*: VQA made seeing visible, but direct interaction recalibrated the video-based expectation.
Evaluate Ameca as a visually grounded interaction partner	Open-weight VQA model selected for local deployability, latency, memory use, and open-ended question handling	9 participants per condition	Post-HRI differences were limited; likeability change differed across groups, driven by the positive-video condition compared with control.	*Alignment*: final judgement depended less on prior framing and more on the robot's actual behaviour during interaction.

The VQA component was integrated and evaluated as a robot perception module for Ameca, while model training was outside the computational scope of this thesis. Figure 6.4 shows the robot-side VQA loop, where a user question and camera frame are passed to the LLaVA-based VQA module before the image-grounded response is returned through dialogue management and robot behaviour. LLaVA-v1.5-7B[7] was selected for its practical balance between multimodal capability and deployment constraints [Liu 23]. Its architecture connects a CLIP-based vision encoder [Radford 21] with an instruction-tuned language model, allowing visual features to be projected into the language-model space of Vicuna [Chiang 23]. The model was run with 8-bit quantisation to reduce memory use during inference and was connected to Ameca through a Python-based application using Finroc [Reichardt 13].

The implementation also included a guidance mode for location-related questions. During initialization, the system loaded an RPTU campus-map image from the university building and site-plan page[8]. If the user query contained the keyword "guidance", the backend used the map image as the visual input for model inference; otherwise, the current camera frame was used for standard VQA. The detailed pseudocode is provided in Appendix A.16.

The study used three conditions: control, positive video, and negative video [Rao 24]. The positive video showed Ameca performing scene-aware interaction successfully, including

[7] https://github.com/haotian-liu/LLaVA
[8] https://rptu.de/en/campus-maps

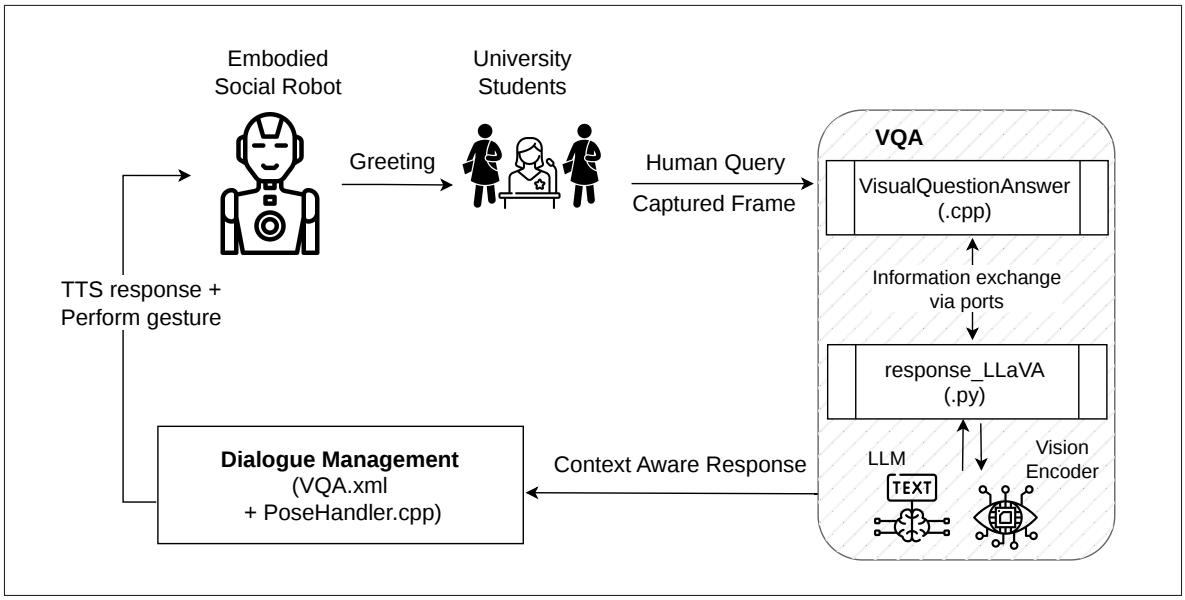


Figure 6.4: Robot-side VQA interaction loop. Ameca receives a user question and camera frame, sends them to the LLaVA-based VQA module, and returns an image-grounded response through the dialogue-management and robot-behaviour pipeline.

responses that connected visible objects to the ongoing dialogue [Paplu 23b]. The negative video showed an unscripted VQA interaction in which Ameca answered incorrectly, vaguely, or humorously. This manipulation followed social learning theory, where observed behaviour can shape later expectations [Bandura 61, Bandura 69, Rosenthal-von der Pütten 13], and expectancy violation theory, where later evaluation depends on whether robot behaviour matches or violates prior expectations [Burgoon 88, Rosén 24a].

The results showed that video exposure shifted immediate perception. Baseline Godspeed ratings did not differ between groups for anthropomorphism, likeability, or perceived intelligence. After the priming video, the positive and negative video conditions differed in pre-to-post-video change for anthropomorphism (positive: $M = -0.62$, $SD = 0.79$; negative: $M = 0.44$, $SD = 0.80$; $W = 72.50$, $p = .005$, $p_{\text{corr}} = .032$) and perceived intelligence (positive: $M = -1.42$, $SD = 0.66$; negative: $M = 0.31$, $SD = 0.43$; $W = 81.00$, $p < .001$, $p_{\text{corr}} = .003$). The manipulation checks for the human interlocutor in the videos did not differ between the positive and negative conditions, suggesting that the effect was linked to how participants interpreted Ameca's visual competence, and not to their evaluation of the person shown in the clip.

After direct interaction with Ameca, the immediate video effects were no longer visible in final Godspeed ratings, system-level ratings, expectancy ratings, or perceived theory of mind. The only significant post-interaction change effect was found for likeability change from pre-rating to post-HRI (control: $M = 0.49$, $SD = 0.44$; positive: $M = 0.40$, $SD = 0.67$; negative: $M = -0.27$, $SD = 0.66$; $H(2) = 6.40$, $p = .041$, $\epsilon^2 = .18$). Pairwise comparisons showed that this effect was driven by the positive-video condition differing from control ($p_{\text{adj}} = .035$). Figure 6.5 shows this pattern across pre-rating, post-video rating, and post-HRI rating.

The VQA study showed that short media exposure can prime expectations of Ameca's

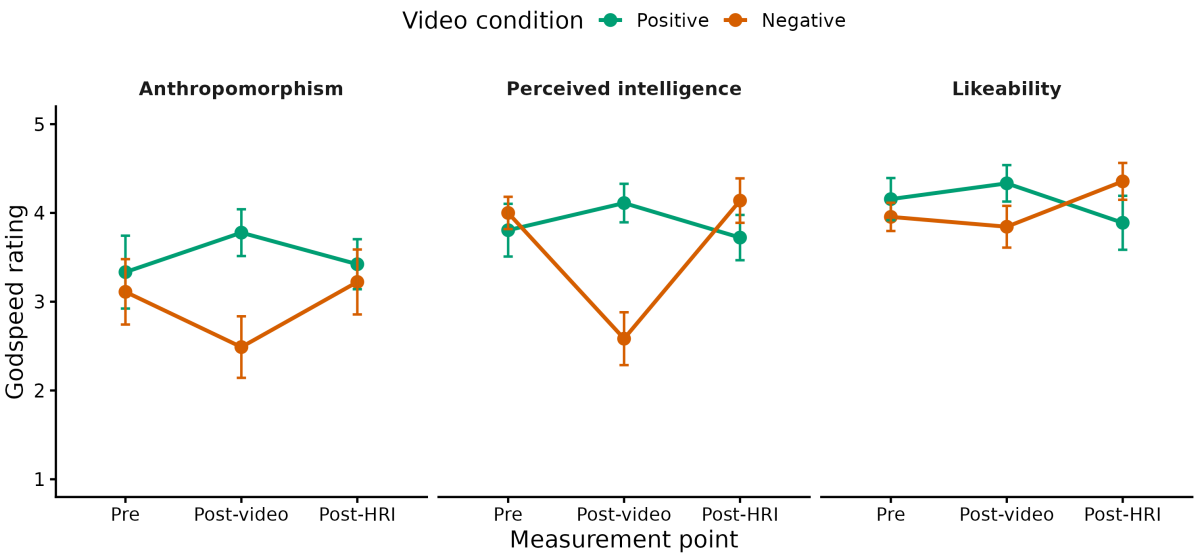


Figure 6.5: Mean Godspeed ratings across pre-rating, post-video rating, and post-HRI rating for the positive and negative video conditions. Error bars show standard errors.

visual competence before direct interaction, while direct interaction can partly override that framing. This relation is especially relevant for highly anthropomorphic robots, since many users first meet such systems through edited or public videos. In this study, the positive video lowered anthropomorphism and perceived intelligence, while the negative video increased both. The live VQA interaction reduced the video-driven differences, suggesting that Ameca's actual answers, timing, and interactional behaviour became more influential for final judgement. VQA made the visual expectation gap measurable, since users expected a robot with eyes to answer as if it could see, while their final judgement depended on whether the live system made that expectation believable.

6.4 Robot Gaze Management

Gaze was the most socially visible visual behaviour in Ameca. Users did not treat eye movements as neutral animation, since unstable fixation, poor timing, and limited emotional readability shaped how they interpreted attention. In group interaction, gaze became more consequential when a faulty left eye-camera made Ameca over-attend to people on its right side, leaving participants on the left with reduced visual attention. The technical problem therefore became social, since the robot's eyes suggested attention while the system did not yet distribute that attention reliably.

Gaze management functioned as ability repair, since Ameca's anthropomorphic face made users expect eye contact, speaker awareness, and fair attention allocation. The gaze systems in this section attempted to align that expectation with the robot's actual perceptual pipeline. Table 6.4 summarizes the development path from default gaze, to single-person gaze control, active-speaker detection, perceived exclusion, and the final Hero-selection system. The detailed model implementations, intermediate algorithms, questionnaires, and extended prototype results are provided in Appendix A.17.

Table 6.4: Development path from default gaze to speech-gaze Hero selection in Ameca.

Goal	System	Sample	Finding	5A
Expose default gaze bias	Default Ameca gaze routine	Prior interaction observations	Default gaze caused unstable attention, irrelevant targets, and side-biased gaze allocation.	*Anthropomorphism*: eyes suggested social attention before the system could deliver it.
Improve single-person gaze control	Eye-contact CNN with face-contingent gaze gestures [Mandal 23]	Pilot study, N=6	Gaze following, blinking, and engagement were rated positively, but lip-sync, delay, and gaze height remained problematic.	*Ability*: face-contingent gaze improved control, but multimodal coordination remained fragile.
Detect the active speaker in groups	TalkNet active-speaker detection [Sangappa 24]	Who-is-speaking game, N=12	Comfort and willingness to participate were positive; perceived recognition accuracy was weak ($M = 2.58$).	*Alignment*: group gaze required speaker detection, not only face tracking.
Test gaze as social inclusion/exclusion	Default gaze in group interviews	Group-interview study, N=35	Participants interpreted reduced robot gaze as exclusion, often attributing it to standing position or their own German-speaking insecurity.	*Anxiety*: technical gaze bias became a social interpretation problem.
Select a focal interlocutor in public HRI	Speech-gaze Hero selection	In-the-wild deployment; 47 utterances, 24 visible people	Best full system reached 0.913 accuracy and 0.882 macro F1 at a 3 s window.	*Alignment*: speech and eye contact made robot attention more socially grounded.

6.4.1 From Gaze Bias to Speaker Detection

The default gaze routine was the starting point because it made the mismatch visible. In single-person interaction, Ameca could produce jittery saccades, look toward irrelevant faces or windows, and redirect participant attention away from the conversation. In group interaction, the faulty left eye-camera created a systematic right-side bias. Algorithm 6.1 summarizes this default routine. The code is relevant through its social effect, since gaze appeared intentional even when unstable target selection drove the behaviour.

The first ability-repair attempt addressed gaze in a single-person setting [Mandal 23]. The system used a pre-trained eye-contact detection model for egocentric video [Chong 20][9], connected to Ameca through a Python/C++ Finroc interface [Reichardt 12, Ashok 24b].

[9] https://github.com/rehg-lab/eye-contact-cnn

Algorithm 6.1: Default gaze behaviour routine

```
Initialise gaze variables and candidate target positions;
setCameraBrightness();
registerGazeListeners();
foreach time step do
    updateGazeTarget();
    if valid target exists then
        (x, y) ← target position;
        (θ_x, θ_y) ← computeEyeAngles(x, y);
        updateEyeGaze(θ_x, θ_y);
    else
        randomGaze();
    end
end
removeGazeListeners();
```

When the chest-mounted ZED RGB-D camera detected a face, the model estimated eye-contact-related coordinates and mapped them to one of 90 predefined eye-gaze gestures, such as `eyeFront(0,0)` and `eyeLeft(20,0)`. Periodic blinking every 3 to 5 seconds was added to make gaze less static. The implementation improved the robot's ability to look toward a visible interlocutor, while it still depended on fragile face tracking, deprecated dlib 19.13.0 dependencies, and careful coordination with other face behaviours.

A pilot study with six participants tested the system while participants moved around Ameca and then interacted with it. Ratings were generally positive but descriptive only. The strongest ratings were for gaze following (Q1: $M = 4.50$, $SD = 0.55$), blinking enhancing the interaction (Q4: $M = 4.33$, $SD = 0.52$), and engagement (Q7: $M = 4.33$, $SD = 0.82$). Head-eye coordination and overall satisfaction were more moderate (Q8 and Q9: $M = 3.83$, $SD = 0.75$). The item "Ameca kept staring at me too much" was rated near neutral (Q2: $M = 3.17$, $SD = 0.41$). Open-ended comments still pointed to the remaining mismatch: "Perhaps a little flexibility issue when turning", "No lip sync, little bit of delay in eye tracking", and "Eye gaze was slightly lower than the face, lip sync was missing". The system made gaze more controllable while gaze, head movement, lip synchronisation, and dialogue still lacked the timing needed for one coordinated behaviour.

The second ability-repair attempt moved from face-contingent gaze to active-speaker detection [Sangappa 24]. In multi-party interaction, Ameca needed speaker awareness in addition to face presence. Several open-source speaker and lip-movement models were explored, including LipNet-based and lightweight lip-reading approaches [Assael 16, Liao 23]. These models did not transfer cleanly to the live robot setting due to deprecated dependencies and noisy real-time HRI conditions. TalkNet was therefore selected for active-speaker detection [Tao 21][10]. In the robot implementation, faces were detected with S3FD, tracked across frames, paired with the audio stream, and passed to TalkNet. If an active speaker was detected, the speaker's face location was converted into gaze target coordinates.

[10] https://github.com/TaoRuijie/TalkNet_ASD

The TalkNet prototype was evaluated with a "Who is Speaking?" game. Twelve participants, including 11 male and 1 female participant aged 24 to 32, interacted in groups of three around Ameca. After each speaking turn, Ameca was expected to look toward the active speaker and ask whether they were the one talking. Participants rated the interaction format positively, especially comfort during the experiment ($M = 4.09$) and willingness to participate again ($M = 4.10$). Understanding of the robot's speaker cues was also moderately positive ($M = 3.90$). Perceived active-speaker recognition accuracy was weaker ($M = 2.58$), and overall satisfaction was mixed ($M = 3.50$). The game was acceptable, while speaker-detection behaviour remained insufficiently robust for general live HRI.

A later group-interview study showed the social consequence of this limitation. Ameca's default gaze behaviour was tested with 35 participants, including 17 male, 17 female, and 1 other participant aged 22 to 33. Five participants per session stood in a semicircle around the robot, with 165 cm distance to Ameca and 60 cm between participants. Ameca asked group-interview questions while directing attention to selected participants. Participants who felt excluded often attributed the exclusion to their standing position or to their own insecurity in German-speaking situations. A sensor or gaze-control bias was therefore interpreted as a social signal of inclusion or exclusion. The finding also connects to fairness in robot learning, since a robot deployed in human groups can reproduce or amplify unequal treatment when sensing, data, or implementation biases affect who receives attention [Londoño 24]. This turns a technical ability limit into anxiety through uncertainty, since users may not know whether they were ignored due to position, language insecurity, or the robot's sensing limits. In social robotics, gaze allocation is also a transparency and dignity issue, because users should be able to understand why the robot attends to some people and not others [Torras 24]. Figure 6.6 shows a simulated group-interview setup in which gaze allocation becomes socially consequential.

Across these prototypes, gaze emerged as a bottleneck for Ameca-based HRI. Default gaze created unstable and biased attention. Single-person tracking improved control, while group interaction still required speaker awareness. Open-source active-speaker models made the right problem visible, while noisy live HRI still challenged reliable deployment. These limitations motivated a tailored speech-gaze attention system for Ameca.

6.4.2 Human "Hero" Detection for Active Speaker Selection

Robot attention systems for social robots have been explored for several years [Lanillos 15], and earlier systems already treated robot attention as an audio-visual selection problem, from real-time talker tracking [Okuno 02] to multi-party gaze target selection [Pan 20] and audio-visual speaker tracking under noise [Qian 22]. Ameca needed focal-interlocutor selection during live interaction, since groups formed and dissolved in front of the robot and gaze had to remain linked to speech, engagement, and response direction. The system therefore combined active-speaker detection with eye-contact estimation so that Ameca could select one focal interlocutor, or "Hero", and align gaze, head pose, and spoken response toward that person. A demonstration video of the speech-gaze Hero selection system is available on RRLab's YouTube[11].

[11] https://youtu.be/AQwq9a1wR_c



Figure 6.6: Simulated group-interview setup illustrating the robot-attention and perceived-exclusion scenario. The image recreates participant positions with colleagues because photographs from the original study could not be shown due to ethical constraints.

The final model was developed as a multimodal attention pipeline for eye-contact estimation and active-speaker detection [Huang 25]. For eye contact, NITEC provided binary eye-contact labels [Hempel 24]. Columbia Gaze did not provide binary labels directly, so labels were derived from its controlled setup. Images with zero head yaw and gaze directed at the central target were labelled as gazing, while the remaining head-pose and gaze-target combinations were labelled as non-gazing [Smith 13]. Faces were cropped to 224×224, normalised with ImageNet statistics, and paired with a 62-point eye-region landmark vector extracted using MediaPipe Face Mesh [Kartynnik 19]. Eye-contact detection used a late-fusion architecture adapted from LOOK [Belkada 21], with a ResNet-50 visual stream and a landmark stream. Active-speaker detection used LASER [Nguyen 26], an extension of LoCoNet [Wang 24b], fine-tuned on WASD [Roxo 24]. LASER used face appearance, audio, and lip motion together. For deployment, the temporal window was reduced from 8 s to 3 s to reduce inference latency. The two modalities served different social-attention functions, with active-speaker detection identifying who was producing speech and eye-contact estimation identifying who was directly engaging the robot. Their combination produced the Hero candidate.

Figure 6.7 shows the real-time social-attention system. The system used Ameca's chest-mounted ZED stereo RGB camera for visual input and a single-channel microphone for audio. Video was processed at 15 FPS and audio was recorded at 16 kHz into a rolling ring buffer. These settings preserved enough temporal detail for speech-gaze alignment while allowing real-time execution on the external robot PC. Every 0.5 s, the system extracted the recent audio window, aligned face crops, and lip landmarks for each tracked identity. This update step kept gaze changes responsive without running the full pipeline on every

frame. Faces were detected with S3FD[12] and tracked with SORT[13] due to their lightweight configuration for live deployment. Short-term tracks were stabilised using MediaPipe Face Mesh, while persistent identities were maintained using InsightFace[14] embeddings with an HSV histogram fallback. This separation was needed because the public setting involved moving visitors, partial occlusions, changing group formations, and inconsistent lighting.

Hero selection combined smoothed speaker and gaze probabilities. LASER produced frame-level speaking probabilities, which were averaged over the current 0.5 s hop and smoothed with an exponential moving average. The eye-contact model produced per-frame gaze probabilities, aggregated with Gaussian time weighting. Separate ON/OFF thresholds prevented identities from repeatedly switching in and out of speaker or gaze states when confidence fluctuated. If the strongest speaker and strongest eye-contact candidate matched, that person was selected as Hero. If the modalities disagreed, a weighted fusion score selected the best candidate. The selected Hero's face-box centre was mapped to robot eye yaw and pitch through a fixed camera-to-eye mapping. Commands were streamed to Ameca's eye controller at 20 to 30 Hz through Tritium WebSocket, and Hero metadata was forwarded to the dialogue module through local IPC.

Algorithm 6.2 summarizes the speech-gaze Hero selection routine. The algorithm stays in the main text because it shows the main technical contribution of the gaze chapter, namely the progression from reactive face tracking to socially grounded attention selection.

The default Ameca person-tracking from Engineered Arts was disabled during deployment and replaced by the Hero-selection gaze controller. The deployed system also supported interaction via a local conversational interface implemented in Python using Ollama. In standard speech-chat mode, user speech was transcribed with WhisperX large-v3 and passed to a local `gemma3:12b`[15] model. For scene-related questions, the system switched to a vision-grounded mode in which the most recent RGB frame from the ZED camera was passed with the user query to `llava:latest`[16]. Jeong and Ghosh [Jeong 25] implemented a separate mimic mode that could be triggered through the verbal command "Mimic me", while attention analysis used only the speech-gaze Hero pipeline. To reduce robot self-capture, the system monitored TTS activity, applied a 0.45 s post-speech cooldown, and discarded inputs with high lexical overlap with the robot's previous utterance. During deployment, audio, video, per-track speaker scores, gaze scores, binary states, and Hero transitions were logged for later reconstruction of attention dynamics.

The in-the-wild deployment was conducted during the Senior Citizens' Fair in Kaiserslautern within a broader comparison of Ameca and the avatar Ann-Sophie [Jahn 26]. The broader study included 25 participants, including 17 female and 8 male participants aged 24 to 90, of whom four interacted only with the avatar, five only with Ameca, and sixteen with both systems. For the Hero analysis, only Ameca interaction segments with a visible interlocutor, usable speech transcription, stable tracking, and a valid Hero prediction were retained. This resulted in 47 utterances across 7 clips and 24 visible people, including 15 female and 9 male people, with clip durations ranging from 1 min 09 s to 5 min 34 s and

[12] https://github.com/yxlijun/S3FD.pytorch
[13] https://github.com/abewley/sort
[14] https://github.com/deepinsight/insightface
[15] https://ollama.com/library/gemma3
[16] https://ollama.com/library/llava

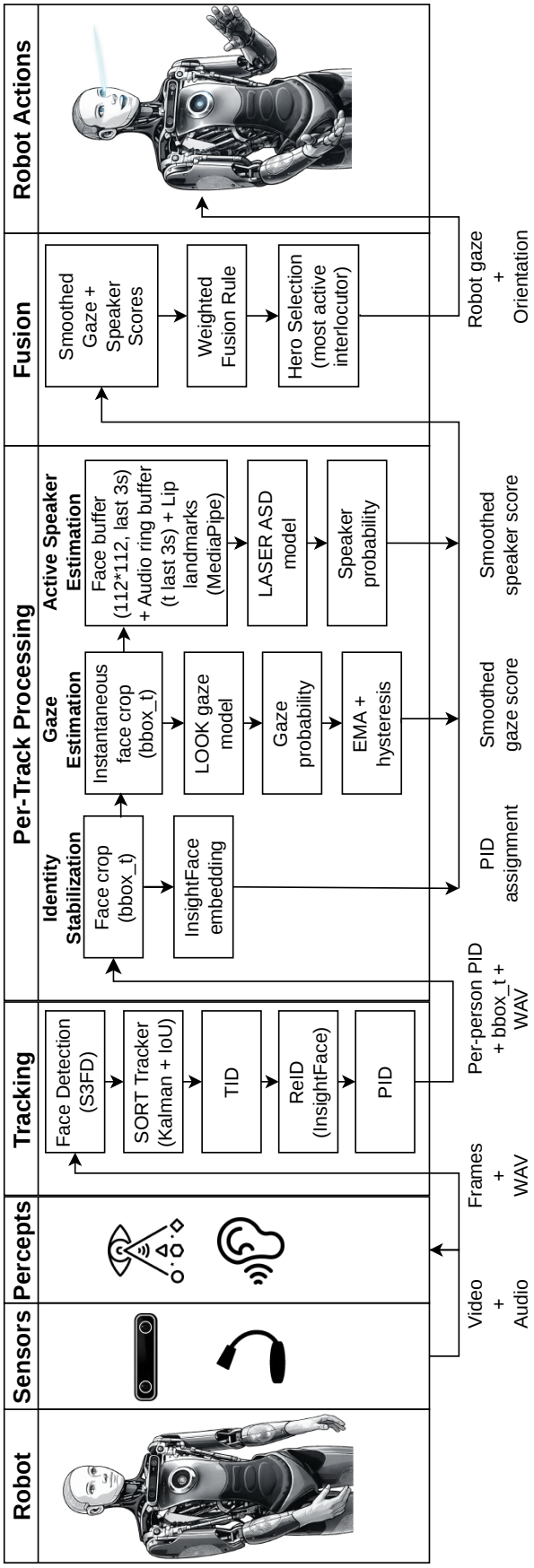


Figure 6.7: Overview of the real-time social-attention system.

Algorithm 6.2: Speech-gaze Hero selection for embodied attention

Initialise identity bank and modality thresholds;
foreach *time step* **do**
  acquire video frame and audio buffer;
  $faces \leftarrow$ `detectFaces(`*video frame*`)`;
  $tracks \leftarrow$ `trackFaces(`$faces$`)`;
  foreach *track* T_i **do**
    $s_i^{spk} \leftarrow$ `speakerScore(`T_i, *audio*`)`;
    $s_i^{gaze} \leftarrow$ `gazeScore(`T_i`)`;
    $\hat{s}_i^{spk} \leftarrow$ `smooth(`s_i^{spk}`)`;
    $\hat{s}_i^{gaze} \leftarrow$ `smooth(`s_i^{gaze}`)`;
  end
  if *speaker and gaze select the same identity* **then**
    $Hero \leftarrow \arg\max_i \hat{s}_i^{spk}$;
  else
    $f_i \leftarrow$ `fuseScores(`$\hat{s}_i^{spk}, \hat{s}_i^{gaze}$`)`;
    $Hero \leftarrow \arg\max_i f_i$;
  end
  if $Hero$ *satisfies minimum confidence* **then**
    $(\theta_x, \theta_y) \leftarrow$ `mapToEyes(`$Hero$`)`;
    `updateGaze(`θ_x, θ_y`)`;
    `sendHero(`$Hero$`)`;
  end
end

sentence lengths from 2 to 25 words. Two independent annotators produced sentence-level speaker labels using synchronised audio-video playback, and disagreements were resolved through discussion before metric computation. Figure 6.8 shows the public deployment setup.

The qualitative interviews from the public deployment showed why gaze control remained socially risky. Ameca's eyes carried anthropomorphism and ability at the same time, making the robot look humanlike and perceptually attentive before users knew how the attention system worked. Some participants interpreted this positively. One participant described the robot as having a "nice face when HE smiles" (ID10, 75 years old, female), while another linked its humanlike impression directly to gaze: "First impression very human, especially her eyes, as if SHE could look at me because SHE moves her eyes" (ID12, 75 years old, female). The same behaviour produced anxiety for others. One participant said, "I find that frightening. HE always stares like that. HE always seeks eye contact. That's very uncomfortable" (ID16, 50 years old, female). Another described the experience as "a bit creepy, exciting, like the eyes and body following you" (ID19, 27 years old, female), and another wrote that "IT looks right into your eyes" (ID25, 33 years old, female). These comments show that gaze control shaped whether Ameca was read as attentive, humanlike, fixed, or intrusive.

Experimenter observations during the public deployment also showed that Ameca attracted heterogeneous bystander orientations beyond the evaluated interaction segments. Some

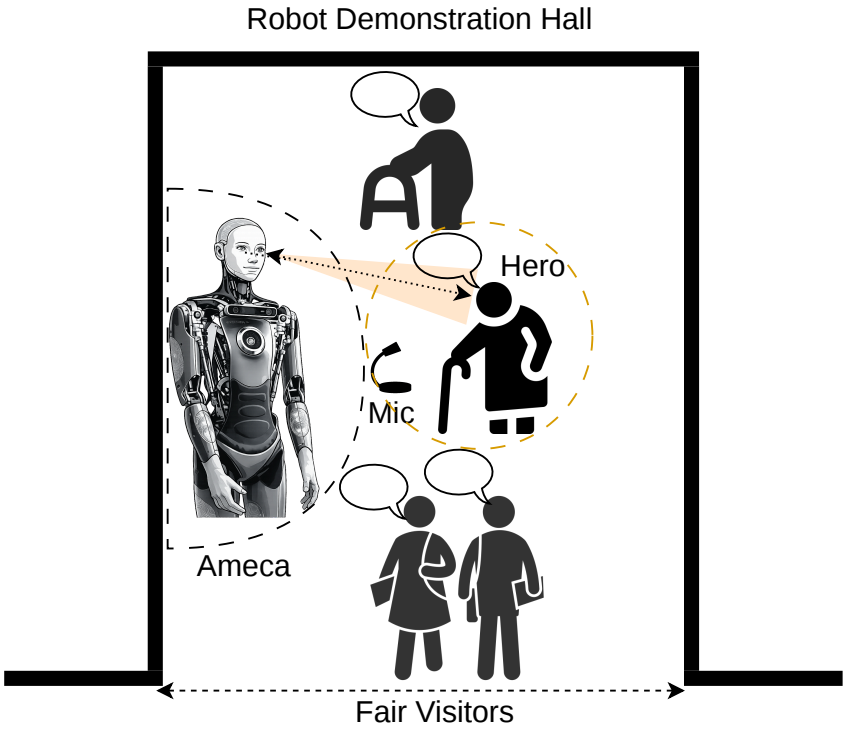


Figure 6.8: Public deployment setup and gaze-speech attention loop.

visitors observed from a distance, some read the accompanying material and left, some discussed the robot in groups, some focused on taking photos or videos, and some attempted to touch the robot or repeatedly controlled the microphone button. These behaviours extend public-space HRI observations of atypical user orientations [Blair 25] and show why gaze and attention systems must be evaluated together with crowd behaviour, interaction guidance, and platform protection. An illustrative public-deployment image is provided in Appendix Figure A.4.

A related proxemics study with Ameca showed that comfort during small-talk interaction could be estimated from mobile eye-tracking features, with interpretable gaze- and pupil-based signals supporting the chapter's argument that discomfort around gaze and distance can be measured in highly anthropomorphic HRI [Kushina 25].

The Hero model was evaluated at sentence level using 47 retained utterances. Hop-level predictions were aggregated by majority vote over each utterance interval, and performance was reported with accuracy and macro F1. Speech was the stronger cue for speaker attribution, gaze helped as a complementary attention signal, and the best trade-off came from combining both. The full speech-gaze system with EMA smoothing performed best at the 3 s window, reaching 0.913 accuracy and 0.882 macro F1. Longer windows reduced responsiveness, which is problematic in public interaction where turns are short and speakers change quickly. The result supports alignment, since the robot's visible gaze became more tightly linked to the person speaking and engaging with it. The full ablation and runtime comparison are reported in Appendix A.17.

A controlled real-time demonstration with three participants confirmed the intended behaviour qualitatively. Participants addressed the robot in turn while remaining in the camera field of view. The system updated the Hero prediction every 0.5 s and

redirected Ameca's gaze as speaker turns changed. Figure 6.9 shows a controlled real-time demonstration of this behaviour.



Figure 6.9: Controlled active-speaker demonstration with three participants. The overlay shows the detected face tracks, speaker score, gaze score, and selected Hero ID, while Ameca redirects gaze toward the current speaker.

The Hero system moved gaze management from reactive face tracking to socially grounded attention selection. It supported alignment by linking Ameca's gaze more closely to who was speaking and who was engaging the robot. Public deployment still showed that gaze can be read as warm, attentive, fixed, or frightening depending on sensing stability, timing, and distribution across people. A humanlike robot is judged by whether its attention feels fair, responsive, and socially readable, not by face selection accuracy alone.

6.5 Discussion

Ameca's visual system created expectation mismatch through the coupling of gesture recognition, object detection, VQA, gaze control, and speaker selection. Users did not experience a wrong VQA answer, unstable face track, missed speaker, or biased gaze target as isolated module failures. Embodied visual behaviour made Ameca appear to not see, not notice, ignore, stare, or attend unfairly.

The early visual-percept systems exposed the first boundary. Dynamic gesture recognition made user input visible, while reliability depended on gesture distinctiveness. Object detection and robot self-recognition reduced ambiguity in the visual field, while the hardest errors appeared where embodiment mattered most, namely between Emah, Ameca, and person. Tangram perception showed that object detection alone was insufficient, since the robot also needed gaze timing and feedback to appear as a shared-task partner. Human

attribute recognition supported adaptive responses, while age, gender, and expression had to remain cautious starting cues instead of fixed truths about the interlocutor.

The VQA study made the expectation gap measurable. LLaVA was integrated as a candidate perception module for embodied interaction. The contribution was the test of whether a locally deployable VQA system could make Ameca's answers fit the visual competence suggested by its eyes and face. Short videos shifted immediate ratings of anthropomorphism and perceived intelligence, while direct interaction reduced these differences. Media shaped the starting expectation; Ameca's live answers, timing, and behaviour shaped final judgement.

The gaze studies made the visual expectation gap socially sharper. Default gaze was not neutral. In group settings, sensor bias could become social bias, since participants who received less gaze could interpret this as exclusion. Single-person gaze control and active-speaker detection improved sections of the pipeline, while fragile tracking, latency, robot noise, and coordination with dialogue and lip-synchronisation remained visible. The Hero system provided the strongest alignment step by connecting speech activity and eye contact to one focal interlocutor.

Visual perception therefore provides the seeing and attention layer of expectation management. The design objective was not humanlike vision, but reliable, bounded, and socially readable links between what Ameca can detect, what it can answer, and whom it attends to.

7. Robot Behaviour Generation and Adaptation

Chapter 3 defined the anthropomorphic expectation gap as a mismatch between what a highly humanlike robot appears to promise and what the system can reliably deliver. This chapter examines that gap through behaviour generation, where Ameca's dialogue, perception, and task responses become visible through facial expressions, gaze shifts, head motion, arm gestures, hand configurations, mimicry, and pointing. Ameca's body already makes movement socially interpretable, so a smile can be read as emotion, a gaze shift as attention, a copied gesture as learning, and a pointing movement as task understanding, even when each behaviour is produced through platform-specific control values, authored sequences, runtime demands, perception triggers, and timing constraints.

Ameca's public reception shows how this morphology shapes behaviour expectations before a task begins. In a local report on an elderly-care demonstration in Germany, Ameca was described as having a "black-grey, metal-like exterior" that "reminds one of an alien", while its artificial voice and steady gaze were foregrounded as features of the encounter [Schwitalla 25]. Similar comparisons appeared during demonstrations when visitors asked whether Ameca was related to the robot figure from *Ex Machina*[1]. These reactions show why behaviour generation on Ameca required legible, timed, and restrained movement: the robot had to move enough to be read as socially responsive, while its visible mechanics and fixed base had to prevent the generated behaviour from over-promising human ability.

The behaviour-generation work addressed the practical alignment problem of making Ameca's visible actions readable while limiting over-promises of human emotion, intention, understanding, or motor competence. Early authored expressions, gaze sequences, and manually timed gesture files made facial configurations, gaze targets, pose files, and dialogue timing inspectable on the physical platform, while later runtime control, mimicry, and perception-driven pointing increased interaction grounding but also exposed new risks of delay, weak motion, unstable mapping, and morphology-constrained behaviour.

[1] https://en.wikipedia.org/wiki/Ex_Machina_(film)

This chapter examines behaviour generation as an *Ability–Alignment* layer of the 5A framework by defining Ameca as a behaviour-generation platform and then analysing facial behaviour, eye behaviour, gesture mimicry, and task-oriented pointing as generated robot behaviours that must remain restrained, timed, morphology-aware, and task-connected to avoid being read as evidence of human emotion, understanding, or intention.

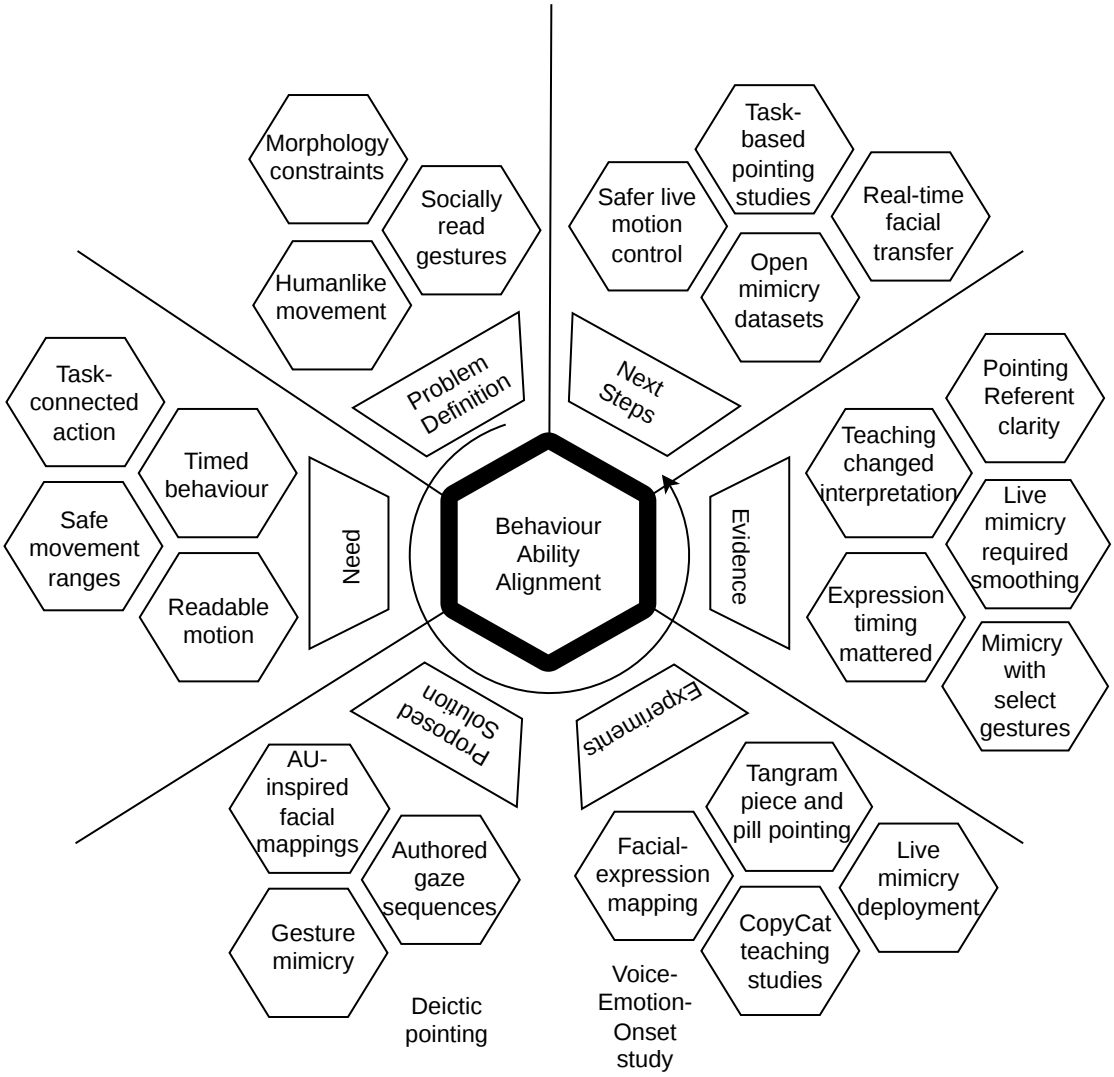


Figure 7.1: Overview of behaviour generation and adaptation as ability alignment in Ameca.

7.1 Overview: Behaviour Generation as Ability Alignment

Behaviour generation on humanoid robots has been treated as a functional interaction mechanism because facial expression, gaze, posture, gesture, timing, and task-oriented movement regulate how interlocutors interpret social presence, attention, and action. Kismet used expression, gaze, posture, and voice to regulate engagement [Breazeal 03], behaviour repertoires organised social cues around interaction goals [Huang 13], and humanoid motion work showed that naturalness depends on timing, smoothness, and

the coordination of voluntary and involuntary movement rather than on target poses alone [Miyashita 04, Xiao 14].

Table 7.1 situates this chapter within work on expressive behaviour, behaviour repertoires, gesture-based HRI, facial mimicry, and real-time affective response generation. The one-to-one mimicking work in this thesis was shaped by discussions with Marcel Heisler on learned facial mimicry [Heisler 24, Heisler 25] and by knowledge transfer from Peizhen Li on practical Ameca control and real-time embodied deployment [Li 25b, Li 25a].

Table 7.1: Relevant works on robot behaviour generation and mimicry.

Reference	Robot	Channel	Generation method	Goal
[Breazeal 03]	Kismet	Face, gaze, posture, voice	Affective architecture and expressive motor system	Social interaction regulation
[Miyashita 04]	Robovie-III	Whole-body motion	Hybrid voluntary and involuntary motion control	Natural humanoid motion
[Huang 13]	Wakamaru	Social cues, mainly gaze	Social cue repository and activity model	Task and communicative goals
[Xiao 14]	Nadine	Upper-body gesture, face, speech	Kinect and CyberGlove recognition with robot response generation	Gesture-based HRI
[Koschate 16]	Humanlike robot head	Facial emotion display	Basic emotion display compared with neutral blinking	Reduce uncanniness
[Mishra 23]	Furhat	Facial expressions	LLM-based real-time emotion generation	Context-appropriate affect
[Heisler 24, Heisler 25]	Android robot head KIM	Facial mimicry	Learned mapping from facial features to actuator values	Human-to-robot facial transfer
[Kirjavainen 25]	Ameca Gen 2	Facial expressions	Authored appropriate and inappropriate expression sequences	Emotional alignment in service HRI
[Li 25a]	Ameca Desktop	Facial expression control	Image-control value dataset for Ameca expressions	Nuanced facial imitation
[Li 25b]	Ameca Desktop	Facial emotion response	Vision-language emotion recognition with predefined expressions	Real-time affective HRI
[Zhu 26]	Rena	Facial expression	End-to-end facial imitation learning with robot-specific dataset	Autonomous facial expression generation

7.2 Platform and Behaviour Generation Constraints

Ameca as Behaviour Platform

Ameca served as the physical platform for the behaviour-generation work in this thesis as shown in Figure 7.2. The RPTU Robotics Research Lab system[2] is a first-generation Ameca platform corresponding to the Gen 1 system described in the Engineered Arts documentation[3]. The robot specification lists a height of 187 cm, 49 kg humanoid platform with 51 degrees of freedom, binocular eye-mounted cameras, binaural ear-mounted microphones, a chest speaker, and an externally mounted ZED2 camera on the torso [Robotics Research Lab, RPTU 22], with 52 motors and an arm span of 180 cm [Hold, Markus 23d]. These properties made Ameca a full-body social robot platform, but the behaviour-generation work in this thesis concerns the expressive upper body because the legs complete the standing humanoid silhouette without supporting locomotion.

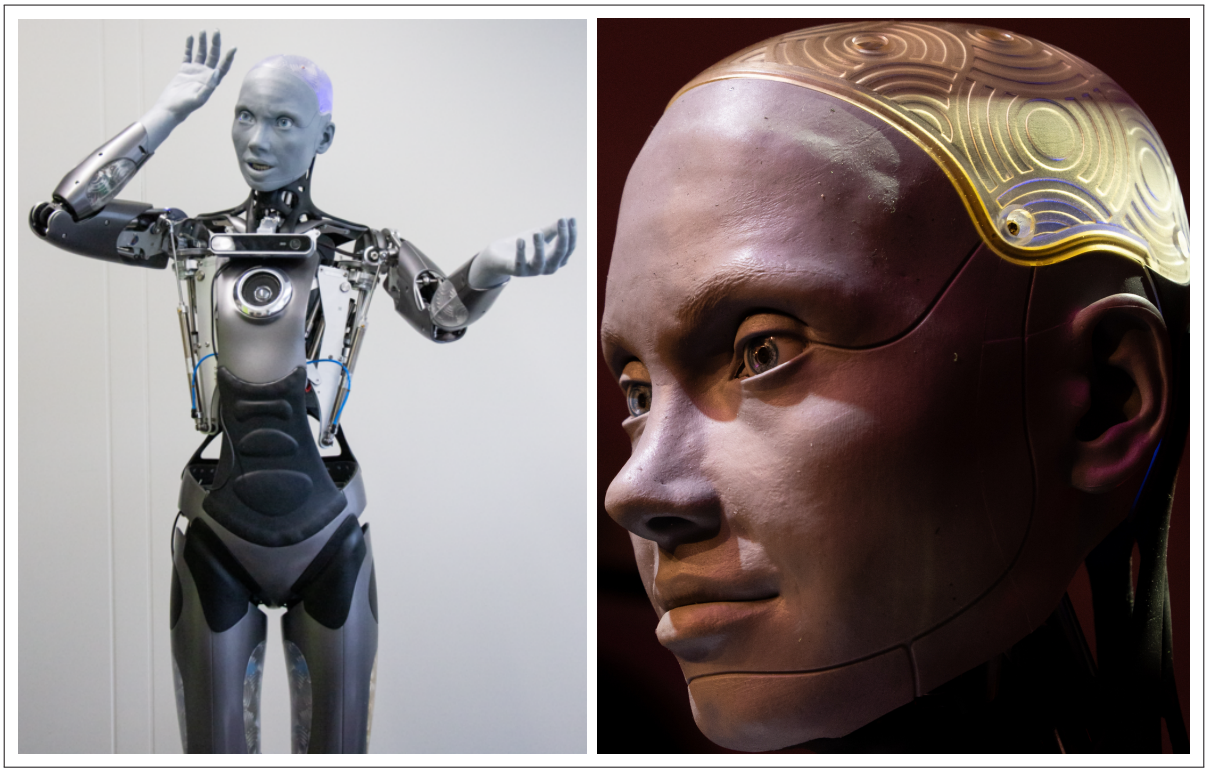


Figure 7.2: Left: first-generation Ameca during an Engineered Arts pre-scripted singing performance. Right: close-up of Ameca's facial surface, articulated eyes, and silicone skin. Photo credit: Abdalla Mohamed

Ameca's visible morphology defines which movements can become socially meaningful on the platform. The robot is built as a modular fixed-base humanoid with legs, feet shells, torso shells, arm shells, silicone covers, a head, and an exposed neck region, so behaviour generation addresses the face, eyes, head, neck, torso, arms, wrists, and hands. The smooth grey silicone face, blue eyes, articulated eyelids, movable jaw, silicone-covered

[2] https://rrlab.cs.rptu.de/en/robots/ameca
[3] https://wiki.engineeredarts.co.uk/Ameca

hands, and humanlike upper-body proportions make the robot appear strongly humanlike [Hold, Markus 23c], while the exposed neck, loudspeaker, torso camera hardware, shoulder mechanisms, cable routing, actuator housings, transparent covers, air lines, and tendon-like elements keep the mechanical structure visible. This mixed human–mechanical morphology prevents a complete human illusion, but it still makes eyelid movement, mouth motion, gaze direction, head pose, arm motion, and stillness readable as attention, uncertainty, amusement, distance, hesitation, or communicative intent.

Ameca's facial morphology also links external appearance to mechanical reconstruction. The manufacturer describes a production process in which a human face is captured through photogrammetry with 36 high-resolution cameras, key facial poses are collected for mould, mechanical, and animation design, the scan is cleaned in 3D modelling software, a stereolithography mould is printed, and silicone is injected, de-seamed, detailed by hand, and mounted over the robotic head [Hold, Markus 23a]. The face is therefore a human-coded surface placed over motors, linkages, air lines, and animation parameters, so facial behaviour on Ameca has to translate from human facial structure to moulded silicone, mechanical actuation, and visible expression. The same translation applies to the body, where humanlike arms, hands, gaze, and posture invite movement categories that the actuator-driven system can only approximate.

Mechanical Degrees of Freedom and Actuation

Ameca's behaviour-control space is defined by the movements that can be addressed through the actuated face, eyes, head, neck, torso, arms, wrists, and hands. Engineered Arts describes the platform in terms of articulated movement and motorised control [Hold, Markus 23b], but the behaviour work in this thesis relied on the controllable parameters exposed on the local RPTU Ameca system: brows, eyelids, eyes, head, neck, jaw, lip corners, nose, torso, shoulders, elbows, wrists, and fingers. Table 7.2 summarizes these practical control groups, while the full exposed parameter list is provided in Appendix A.18. The exposed control values shaped every behaviour-generation decision because facial expressions, gaze shifts, arm gestures, wrist poses, and hand configurations could only be authored or adapted within the ranges and value formats provided by the robot interface.

Table 7.2: Practical behaviour-control groups exposed on the RPTU Ameca system.

Control group	Exposed behaviour parameters
Face	Brows, eyelids, jaw, lip corners, nose
Eyes and gaze	Eye pitch, eye yaw, eyelid opening, head and neck orientation
Head and neck	Head pitch, roll, yaw; neck pitch and roll
Torso	Torso pitch, roll, and yaw
Arms and shoulders	Clavicle, shoulder, elbow, wrist pitch, wrist roll, wrist yaw
Hands	Index, middle, ring, and little finger control for each hand

Software Interfaces for Behaviour Control

Ameca Gen 1 runs on Engineered Arts' Tritium 2 robot software framework[4], which provides the backend environment for sequence playback and runtime control demands. Two behaviour-control routes were relevant in this thesis: authored sequence files for stable and inspectable movements, and runtime control demands for behaviours that had to respond to interaction context. Sequence files can be authored and previewed in Virtual Robot[5], a simulator for creating robot performances before transferring them to the physical platform, and the resulting behaviours can later be called from Python or through the RESTful API as reusable movement sequences (see Virtual Robot simulation: [Kirjavainen 25][Figure 2, Pg 110]). Sequence authoring follows an inspectable keyframe workflow in which a control group is selected, keyframes are placed on a timeline, the motion is previewed on the simulated robot, and the synchronized sequence is transferred to Ameca; the manufacturer tutorial illustrates this with a smile sequence in which the face starts from neutral, the left and right smile controls are moved to target values, the expression is stored in a second keyframe, and a final keyframe returns the face to neutral [Jackson 22]. This route made smiles, nods, waves, gaze targets, and scripted gestures reusable on the physical robot as sequence playback. The runtime control demands used the exposed control values through the available API when gaze direction, head pose, live mimicry, or task-oriented movement had to update the eyes, head, torso, arms, or hands according to perception output or dialogue state [Li 25a]. Since the implementation depended on manufacturer-provided interfaces rather than direct low-level actuation access, adaptive behaviour required valid control ranges, smoothing, timing coordination, and safeguards against abrupt or mechanically unreadable motion, which becomes visible in the earlier dialogue-timing approach.

Resulting Behaviour Generation Constraints

Behaviour timing was the first practical constraint after the available control values and software-control routes were identified. Earlier dialogue systems at RRLab coordinated speech and gesture through explicit XML-style dialogue descriptions in which `<mark>` defined the dialogue state, `in_pose_file_name` selected the reusable movement sequence, `<prompt>` defined the text spoken by the robot, `<wait>` inserted a manually chosen delay, and `<goto>` moved the interaction to the next dialogue state. Listing 7.1 shows this implementation pattern: the `greetings` state triggers the `Smile` sequence before Ameca speaks the greeting, the `company` state triggers the `nod_double` sequence before the next spoken sentence, and the `goodbye` state triggers the `wave` sequence before the closing utterance. This made speech–gesture coordination readable and reusable, but synchronisation remained fragile because movement timing depended on fixed wait values rather than on the actual duration of speech, perception, or interlocutor response.

Listing 7.1: Speech–gesture timing in earlier dialogue control

```
<mark name="greetings"/>
<set_port name="in_pose_file_name" value="Smile"/>
<prompt>
  Hello! I'm Ameca, an advanced humanoid robot from the Robot Research Lab.
```

[4] https://wiki.engineeredarts.co.uk/Ameca
[5] https://wiki.engineeredarts.co.uk/Virtual_Robot

```
</prompt>
<wait time="7000000"/>
<goto mark="company"/>

<mark name="company"/>
<set_port name="in_pose_file_name" value="nod_double"/>
<prompt>
I was developed at the company, Engineered Arts, in London.
</prompt>
<wait time="4000000"/>
<goto mark="goodbye"/>

<mark name="goodbye"/>
<set_port name="in_pose_file_name" value="wave"/>
<prompt> Thanks for listening! Have a great day! </prompt>
```

Behaviour generation on Ameca therefore required timing, legibility, and restraint to be solved together: movements had to align with speech, remain readable within mechanical range, and avoid overplaying humanlike interpretation on a body that already invites it. Later API-level demands improved triggering and adaptation, but the same constraint remained across facial behaviour, gaze generation, gesture mimicry, and task-oriented pointing: visible motion had to be coordinated with speech, perception latency, actuator smoothness, and social interpretation.

The platform constraints define what Ameca can move, while the behaviour systems define how those movements were used to make ability visible during interaction. Table 7.3 summarizes the behaviour-generation systems.

7.3 Facial Behaviour Generation

AU-Inspired Facial Expression Mapping

Facial behaviour generation began with the problem that Ameca's face is human-coded in appearance but robot-specific in control. The FACS describes visible human facial movements through action units produced by facial muscle activity and lists 46 single human AUs [Paul 02, p. 6], with these movements commonly organised into upper- and lower-face regions [Zhi 20]. Figure A.5 shows a replicated reference chart based on the upper- and lower-face AU structure presented by Zhi et al. [Zhi 20][p. 1069].

For Ameca, FACS was used as a reference vocabulary for expression design and not as an anatomical model, since the platform exposes brows, eyelids, eyes, jaw, lip corners, and nose controls instead of human facial muscle groups. The first facial-behaviour task was therefore a mapping problem in which human-readable movements had to be approximated through the exposed Ameca controls before complete robot-specific facial configurations were authored.

The AU-to-control mapping defined the usable facial-expression space for Ameca. Brow raising, brow lowering, eyelid opening, lid tightening, nose wrinkling, lip-corner pulling, mouth opening, jaw dropping, blinking, and winking could be approximated, while several cheek, lip, chin, and nostril actions had no direct control equivalent on the Gen-1 platform. Figure 7.3 shows the AU-inspired movement reference used during Ameca facial-control

Table 7.3: Behaviour-generation systems developed for Ameca as visible ability alignment.

Goal	**System**	**Evaluation**	**Finding**	**5A**
Generate readable facial expressions [Shakeebae 24, Terdale 23, Puertas Céspedes 25]	AU-inspired facial configurations and onset manipulation	FACS mapping, expression repertoire, Voice-Emotion-Onset study	FACS provided a reference vocabulary, but Ameca expressions depended on complete robot-specific brow, eyelid, mouth, jaw, and nose configurations and on their timing during speech.	*Ability*: facial behaviour made response state visible, but needed *Awareness* as robot-specific expression rather than human emotion.
Author repeatable gaze behaviour [Mandal 23, Huang 25]	Pre-authored gaze and blink sequences	47 gaze targets and 45 blink-integrated variants	Sequence files gave stable and reusable gaze behaviours, but limited adaptation to live interlocutor position, speaker status, and dialogue state.	*Alignment*: authored gaze supported control, while runtime gaze was needed for interaction grounding.
Teach and reproduce gestures [Tolj 23, Ashok 24b, Sevinc 24, Majumder 25b]	CopyCat, supervised gesture recognition, and storytelling gestures	Rule-based gestures, 15-person landmark dataset, 10-person pilot, 13 story segments	Gesture mimicry worked when the gesture set was reduced or authored around what Ameca could detect and physically reproduce.	*Ability*: mimicry made learning visible, but reliability depended on detection, correction, and morphology-aware reproduction.
Transfer live human motion [Jeong 25, Jahn 26]	One-to-one mimicry pipeline	Public deployment with active-person selection	Live landmark-to-control mapping required target selection, smoothing, clamping, safe ranges, and separate handling of fast body motion and facial updates.	*Alignment*: live mimicry increased apparent ability, but needed constraints to remain readable and safe.
Point to task objects [Akinci 24, Jeong 26]	Tangram and pill-selection pointing	Task-oriented tabletop prototypes	Object-location information was translated into deictic gestures, but pointing remained constrained by perception, calibration, reachability, timing, and referent ambiguity.	*Alignment*: pointing made perception visible only when detection, gaze, arm motion, and interlocutor interpretation matched the same referent.

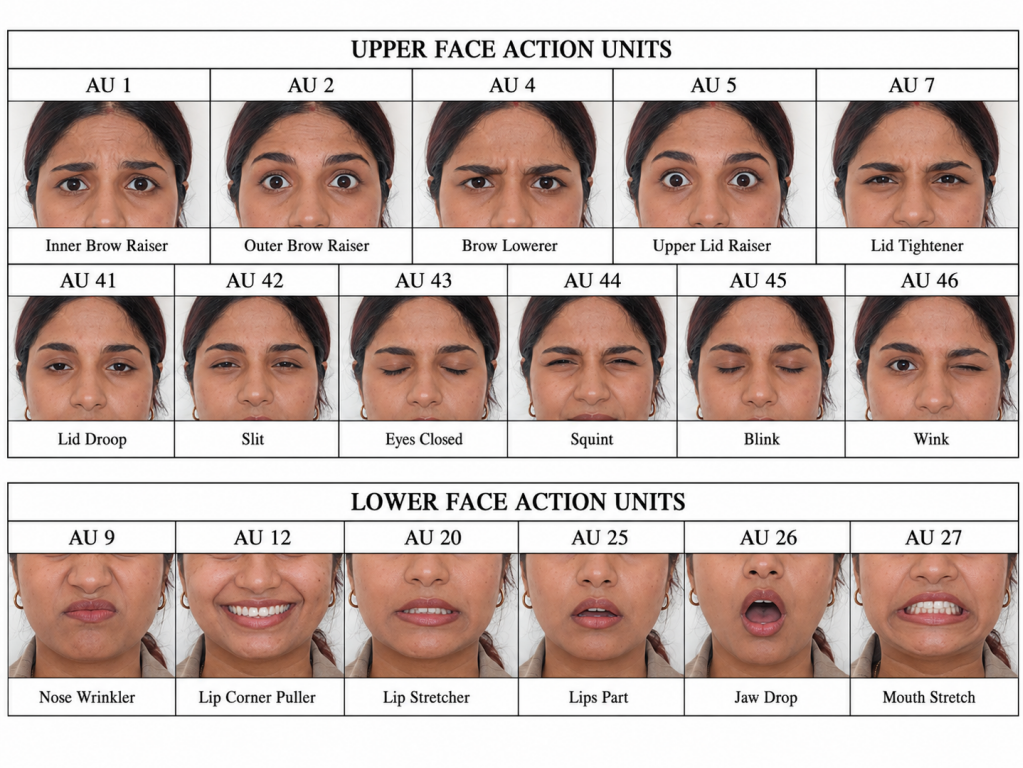


Figure 7.3: AU-inspired facial movement reference for the Ameca facial controls used in this thesis. Image generated using ChatGPT Images 2.0 by OpenAI, 2026 (https://openai.com/index/introducing-chatgpt-images-2-0/), based on an author-provided photograph of the author and AU descriptions. [6]

design. Table 7.4 reports the AU-inspired mappings retained in the main chapter, while the full comparison between human FACS action units and Ameca's available controls is provided in Appendix A.19.

Facial Expressions and Affective Organisation

Ameca's first facial-expressions were created by manually combining the exposed facial controls into reusable robot-specific configurations as sequences [Shakeebae 24]. The combinations were guided by FACS references and checked against correctly scored reference images from the *Facial Action Coding System: Manual and Investigator's Guide* [Ekman 02] and additional visual examples of emotions and action units [Farnsworth 22], but their labels, such as *happy*, *sad*, *confused*, or *frustrated*, were used as interaction-design labels for readable HRI behaviour [Hieida 22]. The full exploratory configuration list is available in Appendix A.19.

The valence–arousal work organised Ameca's facial-expression repertoire beyond discrete labels such as *happy* or *sad* [Terdale 23]. Selected facial controls were linked to their closest AU-inspired basis and grouped by tentative positive/negative valence and low/high arousal, following Kismet's use of affective dimensions to structure complete expressive postures instead of isolated facial components [Breazeal 03]. This step supported behaviour design on the robot face: Ameca produced visible configurations through brow, eyelid, mouth,

Table 7.4: Available AU-inspired mappings for Ameca facial controls.

AU	FACS action unit	Closest Ameca control
AU1	Inner Brow Raiser	Brows L/R
AU2	Outer Brow Raiser	Outer Brows L/R
AU4	Brow Lowerer	Brows L/R
AU5	Upper Lid Raiser	Top Eyelids L/R
AU7	Lid Tightener	Bottom Eyelids L/R
AU9	Nose Wrinkler	Nose Wrinkle
AU12	Lip Corner Puller	Smile L/R
AU20	Lip Stretcher	Smile L/R
AU25	Lips Part	Mouth Open
AU26	Jaw Drop	Mouth Open / Jaw Slew
AU27	Mouth Stretch	Mouth Open
AU41	Lid Droop	Top Eyelids L/R
AU42	Slit	Top Eyelids L/R; Bottom Eyelids L/R
AU43	Eyes Closed	Top Eyelids L/R; Bottom Eyelids L/R
AU44	Squint	Top Eyelids L/R; Bottom Eyelids L/R
AU45	Blink	Top Eyelids L/R; Bottom Eyelids L/R
AU46	Wink	Top Eyelid L/R; Bottom Eyelid L/R

Note. The mapping is AU-inspired and robot-specific; it should not be interpreted as anatomical FACS equivalence.

jaw, nose, and gaze controls, and these configurations were treated as readable expression patterns without equating them with human affective states.

Onset of Facial Expressions

Facial-expression timing created the second alignment problem because a readable configuration can still appear mismatched when it occurs at the wrong moment in speech. The Voice-Emotion-Onset study tested the timing of Ameca's facial expressions by pairing it with the semantically neutral sentence "Our body is made of water" and inserting the expression onset at the beginning, middle, or end of the utterance [Ben-David 11, Puertas Céspedes 25]. The expression set was refined around Ekman's universal[7] emotion categories, including anger, surprise, disgust, enjoyment, fear, sadness, and contempt, as robot-specific design targets and not as claims about human emotion on the robot [Paul Ekman Group 26]. Table 7.5 reports the final control combinations used for seven Ameca facial expressions plus neutral in the study.

Facial behaviour therefore aligned visible Ability only when AU-inspired mappings, complete robot-face configurations, and expression onset were treated as one design problem. The mapping work defined which facial movements Ameca could approximate, the configuration work turned those movements into reusable expression patterns, and the onset study showed that the same expression could become unclear when speech and face were not timed together. The next section shifts from affective readability to attention readability, where Ameca's eyes, head, and neck make gaze visible through controlled movement.

[7] https://www.paulekman.com/universal-emotions/

7.4 Eye Behaviour and Gaze Generation

Eye behaviour in this section refers to the generated motion of Ameca's eyes, eyelids, head, and neck, while Chapter 6 covers the perception pipeline that selects attention targets. In Ameca, gaze becomes socially readable when perception selects a person, object, or location and behaviour generation turns that selection into visible movement. A gaze shift can therefore appear attentive, avoidant, intrusive, or exclusionary even when it originates from a sequence choice, target update, timing delay, or motor-control limit.

The first gaze-generation system used pre-authored eye-gaze sequence files to make looking behaviour stable, inspectable, and reusable during interaction. The file-based system used head, neck, and gaze-target parameters to produce discrete looking behaviours; 47 gaze sequences covered left, right, and front targets across the two-axis gaze space, and 45 blink-integrated variants added eyelid motion to selected gaze shifts. Examples included `eyeFront(0,0)`, `eyeLeft(20,0)`, and `eyeRight(-20,0)`. Figure 7.4 shows the generated gaze targets on the RRLab Ameca platform. The sequence-based gaze method gave repeatable eye behaviour but limited live adaptation. A named gaze target could be tested, saved, and reused across interaction scripts, yet it could not smoothly follow changing interlocutor position, speaker status, task state, or conversational timing.

Runtime gaze control through WebAPI demands and Tritium WebSocket[8] control addressed the limitation by updating gaze, head pose, and attention targets during interaction [Huang 25]. Fewer sequence files were then needed, and gaze could be coordinated with speech, turn-taking, listening behaviour, and body orientation. Gaze generation extends the facial-expression problem from affective readability to attention readability, leading to upper-body gestures where visible movement can be read as imitation, learning, intention, or task participation.

7.5 Body Gesture Generation and Mimicry

Body gesture generation extended visible behaviour from face and eyes to Ameca's upper body. The robot's arms and hands make waving, thumbs-up, defensive posture, hesitation, pointing, and mimicry socially readable as intention, learning, attitude, or task involvement, while the fixed-base morphology limits direct transfer from human shoulder mobility, elbow range, wrist articulation, finger independence, torso movement, and balance. Gesture generation therefore required a translation from human demonstration or gesture label to robot-safe, legible, and morphology-aware motion.

Imitation-Based Gesture Generation

The first body-gesture system implemented observational gesture mimicry through the EMAH system framework [Tolj 23, Ashok 24b]. MediaPipe Holistic[9] detected human pose, hand, and facial landmarks, while the behaviour layer mapped the detected gesture label to the closest available Ameca upper-body sequence. The programming-by-demonstration structure made mimicry interactionally readable: a human tutor performed a gesture, the

[8] https://engineeredarts.com/software/tritium

[9] https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/holistic.md

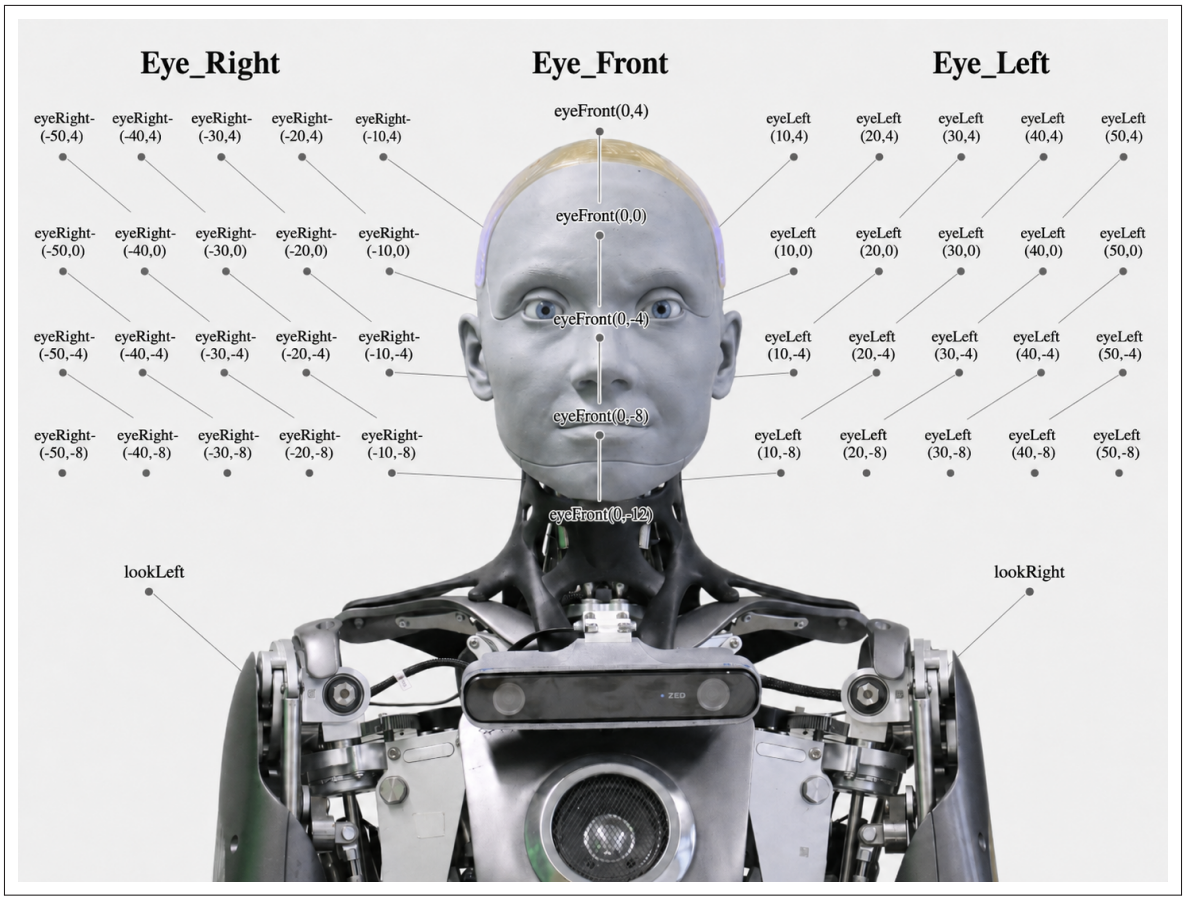


Figure 7.4: Discrete gaze targets implemented for Ameca using pre-authored eye-gaze sequences. The labels indicate named gaze behaviours that could be triggered through the behaviour-control system. Image markings generated using ChatGPT Images 2.0 by OpenAI, 2026. `https://openai.com/index/introducing-chatgpt-images-2-0/`

system classified it, and Ameca reproduced a constrained robot version through authored motion.

Twelve candidate gestures were implemented during the first CopyCat prototype, but pilot testing showed that several gestures did not transfer reliably across body build, hand visibility, and Ameca's morphology. The final interaction therefore retained five gestures that were detectable from human demonstrations and reproducible on Ameca: *victory*, *curious*, *approving*, *back off*, and *unsure* [Ashok 24b]. The full candidate set and detection variables are reported in Appendix A.20.

Audio-Visual Gesture Mimicry

The next gesture system transitioned from hand-coded gesture conditions to supervised gesture recognition within EMAH system [Sevinc 24, Majumder 25b]. Landmark data from 15 participants performing 11 predefined gestures were used to train a gesture classifier, and a pilot study with 10 unseen participants added verbal correction after Ameca's predicted gesture performance. Detailed dataset size, preprocessing, classifier, and feedback-channel information are reported in Appendix A.20. The main design change was interactional: gesture mimicry no longer depended only on fixed rules, but could

use learned recognition while still keeping robot reproduction constrained by Ameca's authored upper-body movements.

Longer Interaction Sequences

The storytelling study extended gesture teaching from isolated movements to sentence-level interaction within EMAH system [Majumder 25b]. Participants taught 13 story segments by combining speech and gestures, and the captured landmarks were labelled per story sentence. Ameca's final storytelling motions were manually animated in Virtual Robot, since direct one-to-one motion transfer was not reliable enough for sentence-level robot reproduction at that stage. The study kept the teachable-robot framing while making the morphology limit visible: Ameca could learn from human demonstration at the recognition level, while final movement still required behaviour authoring matched to the platform.

Live One-to-One Mimicking

Live one-to-one mimicking transitioned from gesture-label recognition to continuous motion transfer. The system captured the interlocutor's face, upper body, and hand landmarks in real time using MediaPipe Holistic, mapped landmark-derived features to Ameca's available control channels, and sent actuator demands through the Tritium WebSocket interface. Figure 7.5 shows the local interaction pipeline used for this implementation.

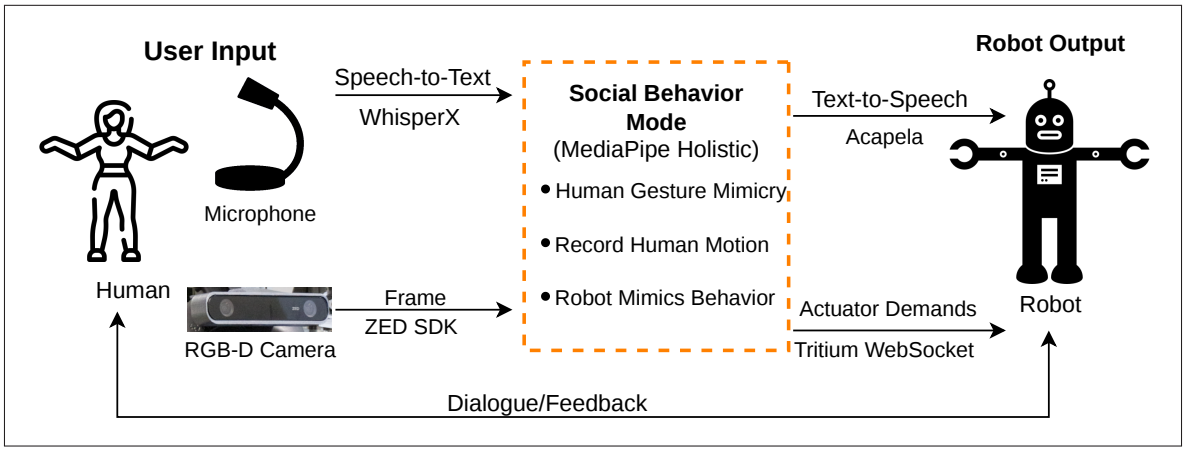


Figure 7.5: Local interaction pipeline for live one-to-one mimicking on Ameca.

The live mimicry pipeline required target selection, normalization, range clamping, and temporal smoothing before movement could be streamed to the robot. Human movement was captured from the ZED RGB-D camera stream, detected faces were assigned persistent IDs, and the active speaker or Hero participant from Chapter 6 was selected as the mimicry target to prevent switching between visible people during public interaction. Landmark-derived features were smoothed with an exponential moving average, clamped to Ameca's safe control ranges, and streamed as WebSocket actuator demands at approximately 20–30 Hz, with low-latency streaming for faster body gestures and interpolated updates for facial movements. Figure 7.6 shows the public live-mimicry setup from the Senior Citizens' Fair in Kaiserslautern.

Mimicry created the strongest behaviour-generation constraint in this chapter since it made Ameca appear to copy and learn from a human body. Visible imitation increased apparent



Figure 7.6: Live one-to-one mimicking on Ameca. The human demonstrator raises both arms in a stretching gesture, and Ameca cautiously reproduces the movement through the real-time landmark-to-control mapping pipeline. Photo credit: Abdalla Mohamed

Ability, while *Alignment* depended on limiting the gesture set, correcting misrecognitions, smoothing live motion, clamping unsafe values, and keeping reproduction within the robot's morphology. The next section moves from expressive similarity to task-oriented deictic gestures, where generated movement makes perceived objects visible within a shared task.

7.6 Task-Oriented Deictic Gestures

Task-oriented deictic gestures used Ameca's body to make perception actionable in a shared task space. Expressive gestures made social or affective stance visible, while deictic gestures directed attention toward an object, location, or decision-relevant target. Pointing therefore supported *Alignment* only when perception, gaze, arm movement, timing, and interlocutor interpretation converged on the same referent.

The Tangram tabletop task used object-location information supported by ArUco markers and visual perception to generate pointing behaviour toward specific puzzle pieces [Akinci 24]. Ameca's arm and gaze directed attention within the shared task space, making perception visible as task assistance. An ongoing robot-assisted selection task extended the same logic to pill selection, where object-level perception was supported through Visual Attentive Prompting (VAP) [Lee 25] and additional VLM-based visual descriptions, allowing Ameca to identify and point toward a specific pill on the table [Jeong 26]. In both cases, pointing functioned as behaviour generation for shared attention and decision support by using arm and gaze movement to make a perceived object interactionally available.

Deictic gestures remained constrained by perception and embodiment. Object-detection errors, marker calibration, table layout, camera viewpoint, arm reachability, gesture

timing, and ambiguous pointing direction could affect whether interlocutors understood the intended referent. The contribution was the integration of perception-driven object reference into Ameca's behaviour repertoire, with pointing closing the progression from expression, to attention, to mimicry, to shared object reference. Detailed Tangram and pill-selection setup information is provided in Appendix A.21.

7.7 Discussion

Behaviour generation showed that Ameca's expectation mismatch emerged from the way motion became socially interpreted through the humanlike body. The mismatch emerged when facial expressions, gaze shifts, gestures, mimicry, and pointing were read as one visible social ability of a humanlike body. A delayed smile, rigid gaze target, weak gesture, failed imitation, or ambiguous pointing action could appear as emotional mismatch, inattentiveness, hesitation, mechanical limitation, or social confusion, which made movement a central layer of Ability–Alignment.

Facial and gaze behaviour exposed the first two boundaries of visible ability. AU-inspired mappings helped structure Ameca's facial controls, yet expression meaning depended on complete robot-face configurations, onset timing, speech context, and the visible silicone-mechanical face. Pre-authored gaze sequences made eye behaviour repeatable, and runtime gaze control made attention more adaptable, yet gaze still depended on target selection, timing, blinking, head motion, and speech coordination. The same design problem therefore moved from affective readability to attention readability, where generated movement had to remain connected to the interaction state.

Gesture mimicry exposed the upper-body boundary of behaviour generation. CopyCat showed that a gesture vocabulary had to be reduced to movements that were detectable from human demonstrations and reproducible on Ameca's fixed-base morphology. Supervised gesture recognition and storytelling expanded the teaching frame, yet robot reproduction still required authored motion, correction, and morphology-aware timing. Live mimicry increased apparent Ability most strongly, since Ameca appeared to copy a human body in real time, yet stable behaviour depended on target selection, smoothing, clamping, safe control ranges, and cautious streaming.

Task-oriented pointing exposed the referential boundary of generated behaviour. Tangram and pill-selection scenarios moved movement from expressive similarity toward shared object reference, where gaze, arm motion, perception output, and table layout had to identify the same object. Research-authored behaviour also remained less fluid than manufacturer performances, since live HRI pipelines had to handle speech timing, perception latency, user movement, model output, and task context during execution.

This chapter contributes the behaviour layer of expectation management. Ameca's body makes movement socially readable, generated behaviour makes apparent Ability visible, and Alignment depends on restrained, timed, morphology-aware motion connected to perception, dialogue, and task state. Awareness is therefore necessary because AU-inspired expressions, mimicry, and pointing can otherwise be misread as human emotion, learning, or intention instead of generated robot behaviour. The next chapter applies this boundary to tandem language practice, where Ameca's visible abilities are constrained by a specific companion role.

Table 7.5: Facial expressions and onset manipulation used in the Voice-Emotion-Onset study.

Expression	Gen-1 Ameca controls	AU-inspired basis	Robot face
Anger	Brow Inner L/R = 0; Nose = 0.5	AU4, AU9	
Disgust	Brow Inner L/R = 0; Brow Outer L/R = 0; Nose = 1; Jaw Pitch = 2.0; Jaw Yaw = 0.9	AU4, AU9, AU25/AU26, AU27	
Fear	Eyelid Upper L/R = 0; Brow Inner L/R = 1; Jaw Pitch = 0.55; Jaw Yaw = 0.5; Nose = 0.6	AU1, AU5, AU9, AU25/AU26	
Enjoyment	Lip Corner Lateral L/R = 0.8; Brow Inner L = 0.6; Jaw Pitch = 0.65; Jaw Yaw = 0.5	AU12, AU1, AU25/AU26	
Sadness	Jaw Pitch = 1; Eyelid Lower L/R = 40; Brow Inner L/R = 1	AU1, AU7, AU25/AU26	
Surprise	Eyelid Upper L/R = 0.8; Brow Outer L/R = 1; Brow Inner L/R = 1; Jaw Pitch = 0.45; Jaw Yaw = 0.5	AU1, AU2, AU5, AU25/AU26	
Contempt	Lip Corner Lateral L = 1; Jaw Yaw = 0.5; Brow Inner L = 0.6; Nose = 0.5	AU12, AU1, AU9, AU26	
Neutral	Brow Inner L = 0.6	Neutral reference; AU1	

8. Social Robots as Tandem Language Partners

Chapter 3 defined the anthropomorphic expectation gap as a mismatch between what a highly humanlike robot appears to promise and what the system can reliably deliver. This chapter examines that gap in a bounded application: German speaking practice with Ameca as a tandem language partner. The application was selected after several iterations on university use cases for embodied social robots, where Ameca's physical presence, expressive face, voice, and dialogue capability appeared useful only when the role was narrow enough for students to understand. In the tandem setting, Ameca's anthropomorphic body could create social presence for speaking practice, while the system role had to remain constrained as a peer-like practice partner outside teacher, therapist, examiner, or native-speaker replacement roles.

The tandem language-partner use case connects the 5A framework to a concrete student need. International students at German universities often need more low-pressure opportunities to practise spoken German, especially when interaction with native speakers feels socially demanding or judgement-laden. Ameca's apparent *Ability* emerges through spoken dialogue, correction, topic support, turn-taking, and embodied presence, while *Anxiety* shifts from robot-related uncertainty toward foreign-language anxiety, fear of negative evaluation, correction, pacing, and interactional safety. *Alignment* therefore depends on a bounded role and a constrained dialogue architecture, while *Awareness* depends on making the robot readable as an LLM-based practice partner with limits in speech recognition, memory, correction, response timing, and language competence.

The chapter follows this application logic in four steps. The first section positions Ameca within Robot-Assisted Language Learning (RALL), foreign-language anxiety, tandem learning, and educational robot roles. The second section grounds the use case through a survey on German speaking anxiety among international students. The third section describes the locally deployed tandem system architecture. The fourth and fifth sections evaluate the system through a single-session comparison and a multi-session repeated-practice study. The chapter closes by arguing that Ameca's value in this setting lies in

supporting low-pressure spoken engagement, while stronger long-term usefulness requires better pacing, continuity, correction, and task framing.

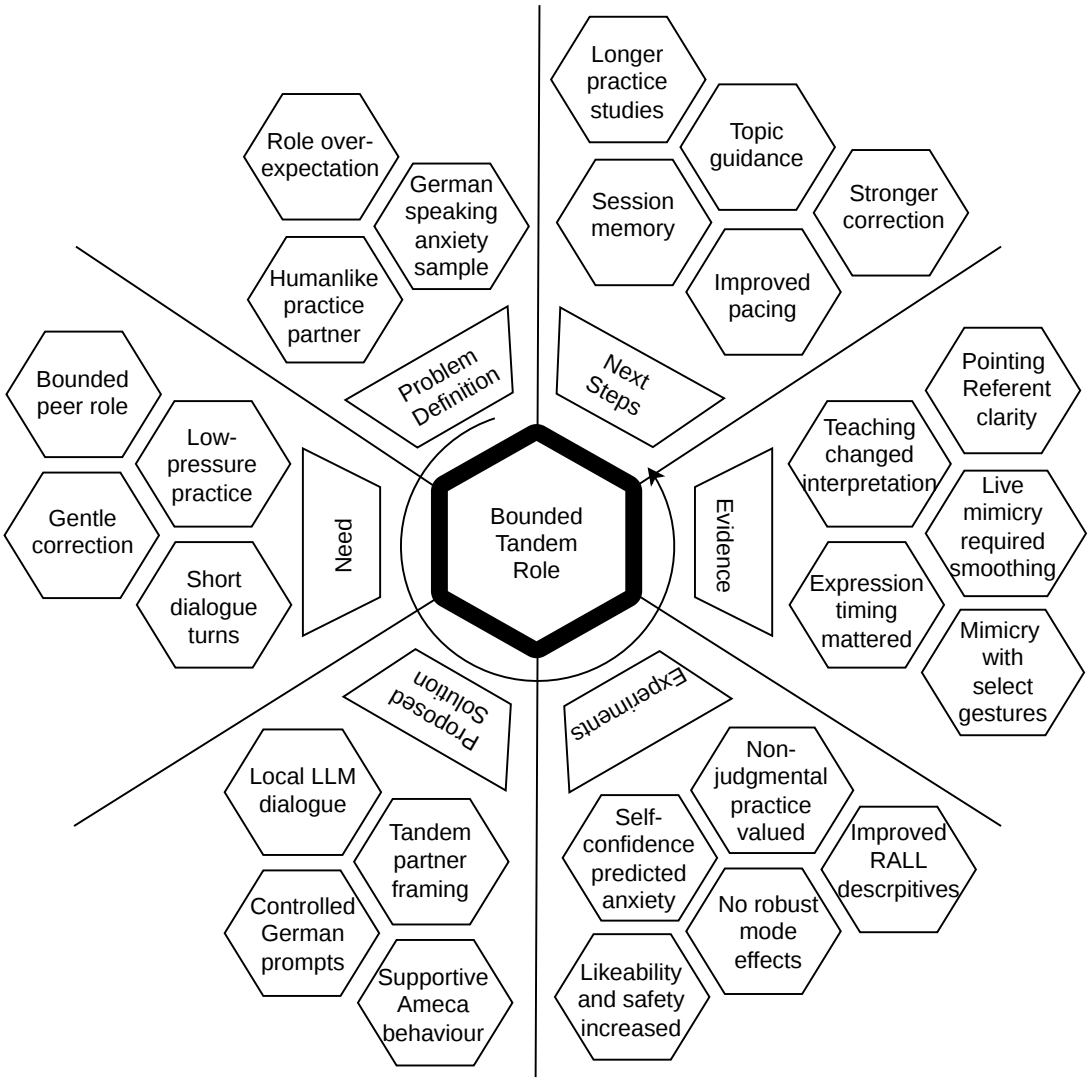


Figure 8.1: Overview of Ameca as a bounded tandem language partner for German speaking practice.

8.1 Overview: Ameca as a Bounded Tandem Partner

Learning the host language as a highly educated migrant is tied to employment, social participation, and belonging. Prior work shows that highly educated migrants face integration barriers when language proficiency becomes a condition for labour-market access and social acceptance [Föbker 17, Reinold 24], while second-language skills can affect job opportunities, competitiveness, and impression management [Li 25c]. Limited national-language proficiency can restrict local networks and contribute to onward migration or return migration toward English-speaking contexts with lower linguistic demands [Föbker 17, Diehl 24]. German is also commonly described as difficult and less attractive to learn, with language mastery framed either as perseverance or as a route toward social acceptance [Truan 25].

These conditions make spoken German practice a socially meaningful application domain for a university robot.

Foreign-language anxiety gives the application its affective starting point. Foreign Language Anxiety describes self-perceptions, beliefs, and emotional responses linked to learning and using a foreign language [Horwitz 86], and the Foreign Language Classroom Anxiety Scale captures speech anxiety, fear of negative evaluation, and discomfort with native speakers [Aida 94]. Speaking anxiety is especially relevant for embodied practice, since oral communication can affect fluency, comprehensibility, and pronunciation [Mora 24], and empirical work reports speaking as one of the most anxiety-inducing components of language learning [Wang 21, Hu 24]. A robot partner becomes useful when it reduces judgement while preserving spoken interaction beyond isolated drills.

Tandem language learning provides the role model for the robot. Human tandem learning is a reciprocal arrangement in which two speakers with different native languages support each other in foreign-language practice [Brammerts 96, Vassallo 06], but oral tandem interaction can still trigger anxiety when learners fear mistakes or negative evaluation [Horwitz 01]. RALL extends language learning through physically or virtually embodied robots that support language comprehension and production through interaction [Kanda 04, Randall 19]. Prior findings remain mixed on learning outcomes, with some studies reporting improvements over non-robot conditions and others finding no significant differences [Han 08, Rosenthal-von der Pütten 16, Donnermann 24]. Robot-mediated practice has also been associated with reduced anxiety and increased engagement, especially when robots are physically co-present [Vogt 19], which makes Ameca relevant as a low-pressure but socially present German-speaking partner.

Educational robot roles shape what learners expect from the system. Tutor and guide robots imply competence, authority, and evaluative ability, while tutee robots imply patience and learner-led instruction. A tandem partner fits a peer role, where the robot supports conversation, practice, encouragement, and gentle correction without taking the authority position of a teacher. Table 8.1 summarizes this role distinction.

Table 8.1: Comparison of educational robot roles for language-practice expectations.

Robot role	**Perceived stance**	**Example**	**Human expectation**
Peer	Collaborative	Tandem robot	Mutual practice; reduced judgement
Tutee	Dependent	Learning robot	Patience and tolerance
Guide	Authoritative	Tutor robot	Competence, correction, and trust

Ameca was positioned as a tandem partner within the existing RRLab robot ecology. ROMAN had been used for affective interaction, Robothespian for social-companion research, and Ameca for highly anthropomorphic embodied interaction. A one-to-one tandem format matched Ameca's physical constraints and reduced the public-exposure problem associated with hallway or classroom deployment. Table 8.2 places the tandem use case within RRLab context, while Table 8.3 summarizes relevant RALL systems.

Table 8.2: Use case classification of RRLab humanoid robots.

Robot	Use case	Primary focus	Environment
ROMAN	Affective interaction	Emotion expression and perception	University lab
Robothespian	Campus Guide	Personality-based adaptation	University lab
Ameca	Language partner	Robot-assisted German speaking practice	University lab

Table 8.3: Overview of robot-assisted language learning with social robots.

Reference	Dialogue Type	Social Robot	Use case	Sample
[Han 08]	Rule-based	Nao	Vocabulary learning support	Children (N/A)
[Rosenthal-von der Pütten 16]	Rule-based	Nao	Vocabulary learning	Adults (N=40)
[Trinh 17]	Rule-based	Furhat	Oral presentation coaching	University students (N/A)
[Vogt 19]	WoZ	Nao	L2 vocabulary learning	Children (N=34)
[Banaeian 21]	Rule-based	Nao	Vocabulary learning	University students (N=36)
[Verhelst 24]	LLM OpenAI GPT-3.5	Furhat	Adaptive L2 tutoring with feedback	School students (N=21)
[Sharma 24]	LLM	Avatar	Spoken English practice	Adults (N/A)
[Ashkenazi 25]	WoZ	Furhat	Multilingual speaking	University students (N=60)
[Kamelabad 25]	WoZ	Furhat	Monolingual vs bilingual interaction	Adults (N=47)
[Rybakova 25]	Rule-based	Avatar	Spoken English practice	Adults (N=9)

The literature and role framing define the application boundary for the studies in this chapter. Ameca was evaluated as a bounded peer-like tandem partner outside the roles of full teacher, automated language examiner, or replacement for human practice. Table 8.4 summarizes how the three empirical steps move from learner-side anxiety, to first-contact

tandem interaction, to repeated practice with Ameca before the survey establishes whether speaking anxiety was present in the intended university population.

Table 8.4: Tandem-language studies mapped onto the 5A framework.

 **Goal**	 **System**	 **Evaluation**	 **Finding**	 **5A**
Ground the use case in learner anxiety [Helf 25]	German speaking-anxiety survey	International students, German-course and self-practice learners; N=66	Speech anxiety was measurable in the target population, and higher German self-confidence predicted lower speaking anxiety.	*Anxiety*: the target was fear of speaking, correction, judgement, and conversational difficulty.
Test first-contact tandem interaction [Helf 25, Ashok 25a]	Single-session OpenChat tandem system; N=33	Audio-only German, robot German, and robot German-English modes	Interaction modes did not produce robust differences in lexical gain, anxiety, expectations, or RALL attitudes, while likeability and perceived safety increased overall.	*Ability–Alignment*: Ameca supported acceptable short practice, but timing, clarity, correction, and role fit shaped the experience.
Test repeated robot-assisted practice	Multi-session Ollama tandem system; N=7	Five German practice sessions with Ameca	Repeated interaction kept willingness positive and likeability high, with descriptive improvements in RALL attitudes and small anxiety reduction, but no corrected reliable pre-post effects.	*Alignment–Awareness*: the tandem role worked best when Ameca stayed bounded as a non-judgemental practice partner with clear limits in memory, pacing, recognition, and correction.

8.2 Empirical Grounding: Foreign-Language Anxiety Survey

The survey established whether German speaking anxiety was present in the intended university population before Ameca was implemented as a tandem partner. The study compared international students currently enrolled in a German course with students practicing German outside a recent formal course, since both groups could need spoken practice but experience different kinds of interactional support. The research question focused on whether foreign-language anxiety affected international students learning German and whether course status changed speech anxiety, German confidence, or willingness to practice with a humanoid robot.



How does foreign-language anxiety affect international students learning German?

The online survey was implemented with involve.me[1] and approved by the university ethics committee. Participants were routed to either the German-course version or the self-practice version after viewing a landing page, and both versions collected consent, demographics, German self-confidence, self-reported German proficiency, modified FLACS items, an open-ended reflection on speaking comfort, and willingness to practice German with a humanoid robot. The reliability checks, distribution tests, extended regression results, routing pages, and modified FLACS items are reported in Appendix A.22 and Appendix A.28.

The final analytic sample contained 66 participants after excluding one likely native German speaker and one participant without consent. The sample included 26 German-course learners and 40 self-practice learners, with 32 female, 32 male, and 2 other participants. Self-reported German proficiency was concentrated in beginner and intermediate levels, with A-level learners at 35.3%, B-level learners at 33.8%, C-level learners at 26.5%, and beginners at 4.4%. The sample therefore matched the intended group for the tandem application: international students who need German in everyday life but may lack confidence in spontaneous spoken interaction.

The modified FLACS analysis focused on *Speech Anxiety and Fear of Negative Evaluation* and *Comfortableness in Speaking with Natives* [Aida 94]. Speech Anxiety showed excellent reliability ($\alpha = .94$), while Comfortableness with Natives showed weak reliability ($\alpha = .26$) and was used only as a cautious secondary indicator. Across the analytic sample, Speech Anxiety was moderate ($M = 2.85$, $SD = 0.94$), with nearly identical scores for German-course learners ($M = 2.84$, $SD = 0.95$) and self-practice learners ($M = 2.85$, $SD = 0.95$). Willingness to practice with a humanoid robot was moderately positive in both groups, with course learners at $M = 3.77$ ($SD = 1.14$) and self-practice learners at $M = 3.55$ ($SD = 1.11$), which supported the feasibility of testing Ameca as a low-pressure practice partner.

The strongest quantitative result was the relation between German self-confidence and speech anxiety. An OLS regression with HC3 robust standard errors showed that German self-confidence significantly predicted lower Speech Anxiety ($b = -0.525$, $SE_{\text{HC3}} = 0.101$, $t = -5.176$, $p < .001$), while group, years in Germany, and age did not explain additional variance. Ordinal logistic regression showed that willingness to practice with a humanoid robot was not reliably predicted by Speech Anxiety, German self-confidence, or group, indicating that openness to robot-assisted practice extended beyond anxiety level or confidence level. Figure 8.2 shows the relation between German self-confidence and speech anxiety.

The open-ended responses clarified why a tandem robot could be useful beyond the numerical anxiety score. Self-practice learners often emphasized vocabulary limits, topic difficulty, and fear of judgement in less structured conversation, while German-course learners more often emphasized supportive partner behaviour and guidance. Participant

[1] https://www.involve.me/

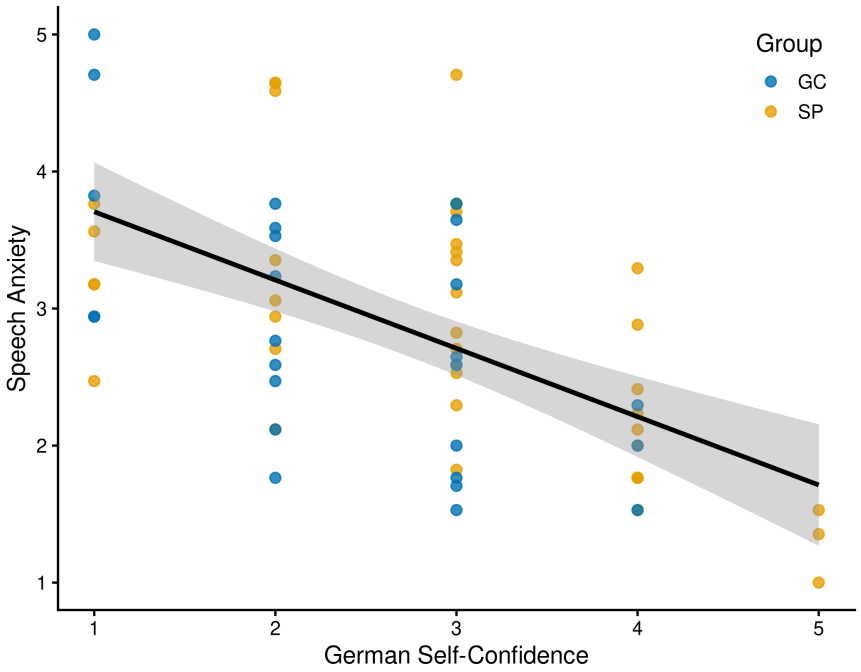


Figure 8.2: Relationship between German self-confidence and reported speech anxiety. Higher self-confidence is associated with lower speech anxiety.

p26 wrote that even a small positive response from a partner gives "a confidence boost," p30 described vocabulary and grammar as the central concern, p64 linked comfort to a partner who helps her speak freely without worrying about mistakes, and p66 connected comfort to both a kind partner and familiar vocabulary. Speaking anxiety was therefore shaped by language ability, perceived judgement, topic manageability, and interactional safety.

The survey established the application need for a bounded speaking partner at the university. The target population showed moderate speaking anxiety, the anxiety was tied to self-confidence and interactional comfort, and willingness toward robot-assisted practice was positive enough to justify testing Ameca as a peer-like tandem partner. The next section therefore describes the locally deployed tandem system whose role, language mode, response length, and correction style were constrained before user evaluation.

8.3 Tandem System Architecture

The tandem interaction system was implemented as a locally deployed LLM-based dialogue architecture inside the existing Ameca control stack. The architecture had to satisfy three constraints for the tandem role: inference had to run on university infrastructure for reproducibility and data protection, prompts had to constrain role, language mode, response length, and correction style, and the dialogue output had to connect to speech recognition, text-to-speech, and robot behaviour modules during live embodied interaction. Spoken input was therefore recorded locally, transcribed with WhisperX, passed to a locally served LLM, constrained through a tandem-partner prompt, and returned through Ameca's speech and synchronized robot behaviour.

The shared implementation across the single-session and multi-session studies used local speech acquisition, GPU-supported WhisperX transcription, local language-model inference, and text-to-speech output through Ameca. Table 8.5 summarizes the common configuration, while the following sections specify the OpenChat and Ollama deployments used in the two evaluations. A demonstration video of the system is available as *International Students practising with Ameca*[2].

Table 8.5: Shared technical configuration across the tandem studies.

Component	Specification
Speech-to-text	WhisperX, GPU-accelerated, 16 kHz microphone input, fully local transcription
Hardware	Xeon Gold 6126 CPU, 64 GB RAM, RTX 4090 with 24 GB VRAM
Environment	Ubuntu 22.04, Python 3.10
Communication	TCP-based client-server interface between robot controller and LLM server
Deployment	Fully local inference, no external APIs
Persona layer	Prompt-defined German tandem partner “Ameca”, informal address form *du*, peer framing

The interaction design distributed initiative between Ameca and the learner to support practice without turning the robot into an examiner. Prior work links perceived interactivity to how initiative is shared between system and user [Wang 25], and the tandem system used this principle by letting Ameca open the interaction, propose topics during hesitation, and keep turns short enough for learner agency. The single-session study tested whether embodiment and language mode changed immediate learning and evaluation, while the multi-session study tested whether repeated practice changed anxiety, attitudes, and speaking-related indicators over time.

8.4 Single-Session Tandem Study

The single-session study tested whether one short tandem interaction differed across three modes: German-only audio with the robot inactive, robot-active German-only interaction, and robot-active German-English interaction. The previously published conference version reported only the two robot-active conditions [Ashok 25a]; the thesis analysis reports the final three-mode comparison to include the audio-only condition and test whether embodiment and bilingual support changed the immediate interaction outcome.



How do German language learning gain, robot perception, tandem expectations, anxiety, and attitudes toward robot-assisted language learning change across audio-only German, robot German, and robot German-English modes?

The single-session system used a quantized OpenChat 3.5 model deployed through `llama.cpp` to compare immersive German-only and bilingual German-English interaction.

[2] https://youtu.be/6k5llJPbLMM

The prompts defined Ameca as a supportive peer-tandem partner, maintained the informal *du* form, encouraged sustained conversation, and limited output to one main answer with optional micro-correction and rephrasing. The immersive condition used German only, while the bilingual condition allowed English clarification for explanation and repair. The full prompts and OpenChat integration script are provided in Appendix A.29, and Table 8.6 summarizes the model configuration.

Table 8.6: Single-session LLM configuration.

Component	Specification
Model	OpenChat_3.5-GGUF [Wang 23]
Base model	LLaMA-2-13B [Touvron 23]
Quantization	Q8_0, GGUF variant
Inference engine	`llama.cpp` with Python bindings
Decoding	temperature = 1.0, top-p = 1.0, max_tokens = 4096
Context handling	Full session history retained within 4096-token limit
I/O contract	One reply with optional micro-correction and rephrase
Interaction modes	Immersive German-only; bilingual German with English clarification
License	Open RAIL M v1

The robot implementation used the Gen-1 Ameca platform with a female German voice, Julia from Acapela Group[3], synchronized spoken output, and an encouraging expression pattern with raised eyebrows, widened eyes, and a closed-mouth smile [Nikitin 19]. Figure 8.3 shows the expression used to support an approachable practice setting. The two robot-active conditions used the same physical setup and differed only in language mode, while the audio-only condition kept Ameca inactive and delivered the conversational output through the chest speaker.

The study included 33 participants with balanced allocation across the three modes ($n = 11$ each). Participants were 19 to 34 years old ($M = 26.88$, $SD = 3.36$), with 19 men and 14 women, and came from diverse academic and national backgrounds. After a short introduction, participants completed a pre-questionnaire, interacted for up to 15 minutes, and completed a post-questionnaire. Measures included German LexTALE vocabulary scores [Lemhöfer 12], Godspeed anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety [Bartneck 09a], adapted RALL attitudes [Alemi 15], adapted tandem expectations from Robocop [Trinh 17], a separate anxiety item, post-interaction tandem-partner items adapted from Salomons et al. [Salomons 22], and an open-ended feedback question.

The open-ended feedback showed that interaction quality shaped the experience more clearly than assigned mode. Participants across conditions mentioned response latency, with comments such as “faster response,” “waiting time between questions and answers were shorter,” and “reply little bit slow.” Learners also requested slower or clearer speech output for longer answers and stronger corrective support, including direct requests such as “correct my mistakes.” A smaller set of comments referred to conversational style and social expressiveness, with one participant describing the robot as “too much neutral” and another writing that it needed “more emotions.” Timing, clarity, correction, and social expressiveness therefore became the practical issues carried into the multi-session design.

[3] https://www.acapela-group.com/

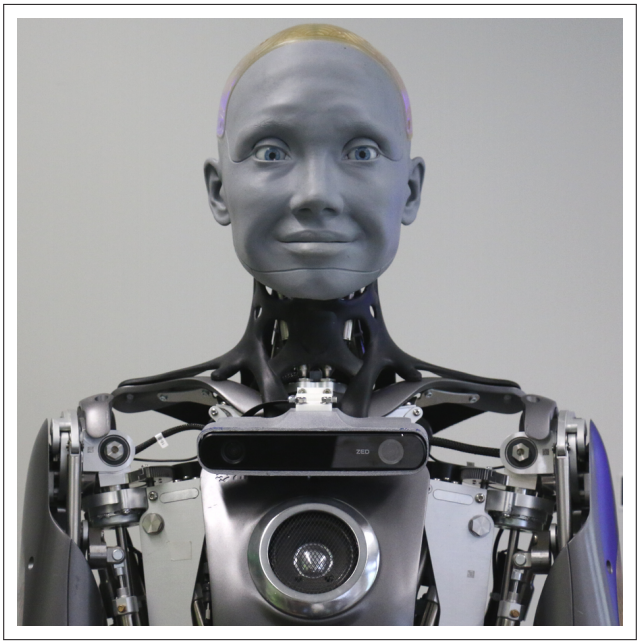


Figure 8.3: Encouraging facial expression used during the tandem interaction.

The quantitative results showed no robust mode-specific effects. Lexical gain was highest descriptively in the robot German-English condition ($M = 4.06$, $SD = 9.64$), lower in audio-only ($M = 2.18$, $SD = 17.61$), and slightly negative in robot German ($M = -0.34$, $SD = 4.37$), while a Kruskal-Wallis test found no significant mode effect, $\chi^2(2) = 3.06$, $p = .216$. Likeability and perceived safety increased from pre- to post-interaction across all modes, with significant main effects of Time for likeability, $F(1, 30) = 12.88$, $p = .001$, $ges = .072$, and perceived safety, $F(1, 30) = 8.70$, $p = .006$, $ges = .034$, while neither showed a significant Mode × Time interaction. Anxiety, anthropomorphism, animacy, perceived intelligence, expectations, RALL, and post-interaction tandem evaluation showed no reliable mode-specific differences. The expectation and RALL patterns are shown in Figures 8.4 and 8.5. The lexical gain, likeability, perceived safety, reliability checks, assumption checks, Holm corrections, and full mixed-model results are reported in Appendix A.22.

The single-session study showed that brief tandem interaction was acceptable across modes, while one exposure did not produce clear mode-specific differences in learning or evaluation. Participant comments pointed to timing, clarity, continuity, and correction as the interactional issues that required tighter output control, which motivated the multi-session study.

8.5 Multi-Session Tandem Study

The multi-session study tested whether repeated German speaking practice with Ameca changed anxiety, learning-related indicators, robot perception, and willingness over time. The study used a within-subject design across five sessions over two weeks, with Session 1 and Session 5 used for the final pre-post analysis.

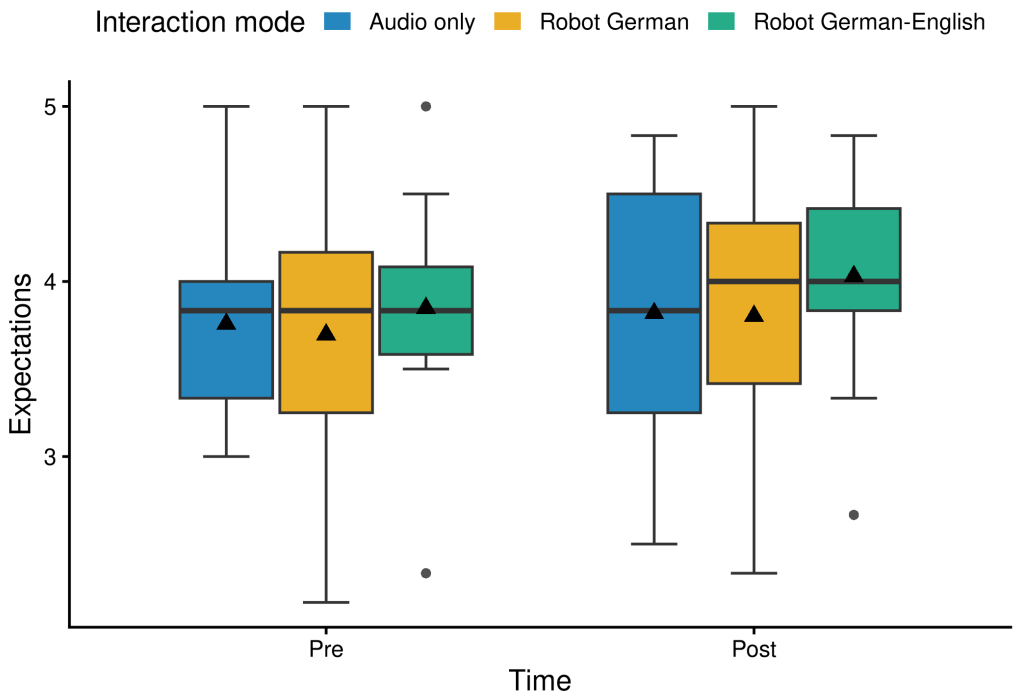


Figure 8.4: Expectations of the system as a tandem partner across the three single-session modes.

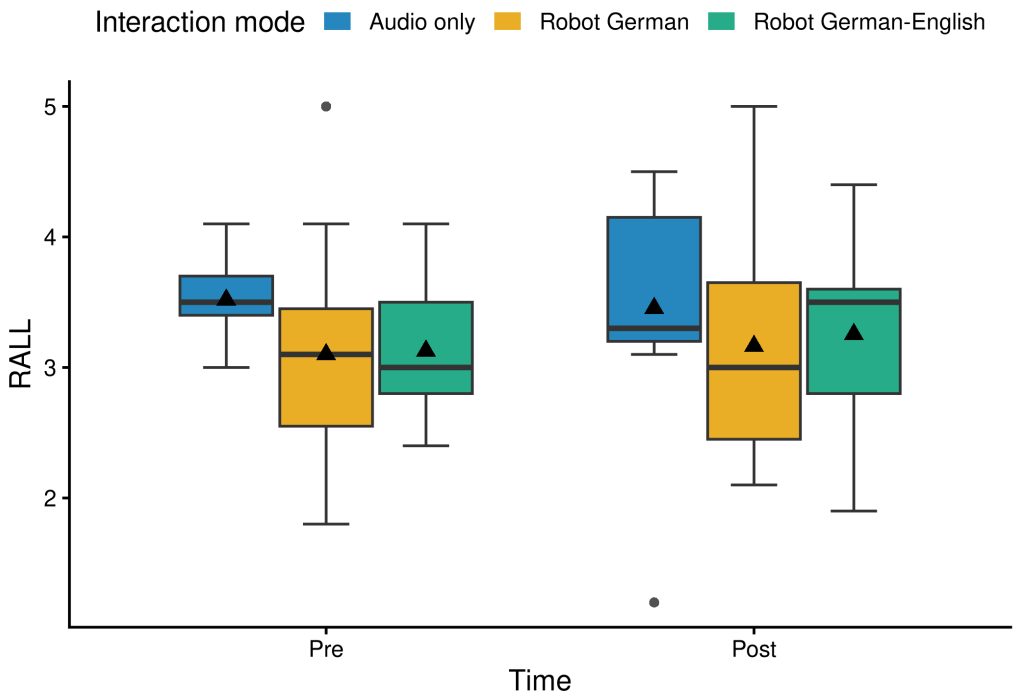


Figure 8.5: Attitudes toward robot-assisted language learning across the three single-session modes.



How do speech anxiety, listening performance, peer-rated speaking performance, anthropomorphism, likeability, attitudes toward robot-assisted language learning, and willingness to practice with a humanoid robot change from the first to the fifth session of repeated German speaking practice with Ameca?

The multi-session system replaced OpenChat with `gemma3:27b` served through Ollama to support more stable repeated interaction. The prompt used a stricter role and output contract than the single-session setup: Ameca remained a peer-tandem partner, responded only in German, used the informal *du* form, avoided formatting and emojis, and produced no more than two short sentences of up to ten words each. The system retained only a short recent interaction history, limited conversational drift, and standardized the opening line delivered by Ameca. The full prompt and Ollama script are provided in Appendix A.30, and Table 8.7 summarizes the model configuration.

Table 8.7: Multi-session LLM configuration.

Component	Specification
Model	Gemma3:27B via Ollama
Base model	Gemma 3
Parameters	27B
Inference engine	Ollama Python Client, SSH to `neptun`
Decoding	temperature = 1.0, top-p = 1.0, max_tokens = 4096
Context handling	Explicit truncation to last six user-assistant turns
I/O contract	Two concise German sentences, maximum 10 words each
Interaction mode	Immersive German-only
License	Gemma, Google DeepMind, research use

The robot implementation used the same Gen-1 Ameca platform, female German voice, synchronized speech, and encouraging facial expression as the single-session study. Seven participants completed all five sessions after one dropout. The retained sample included four women and three men, aged 23 to 34 years ($M = 29.86$, $SD = 4.30$), with diverse language backgrounds and self-reported German proficiency ranging from A2.1 to C1. All participants had prior experience with Ameca, which is relevant because previous robot exposure can shape expectations and evaluations [Rosén 24b, Nikolovska 24]. Figure 8.6 shows the multi-session setup.

Session 1 and Session 5 included the same main measurements. Participants completed consent, demographics, confidence ratings, self-reported German proficiency, a Goethe-Institut listening task[4], modified FLACS items [Alemi 15], and Negative Emotion and Meaning subscales from the L2 Learning Experience Scale [Li 25c]. Participants then completed a short German conversation with a native-speaking peer, who rated vocabulary, structures, grammar, and pronunciation using a four-point scale adapted from [Kultusministerkonferenz 19]. After that, they practiced German with Ameca and completed post-measures including Godspeed anthropomorphism and likeability [Bartneck 09a], adapted RALL attitudes [Alemi 15], willingness to practice with a humanoid robot, and open-ended reflections. The average session duration was 36 minutes and 35 seconds.

[4] https://www.goethe.de/de/spr/kur/tsd.html

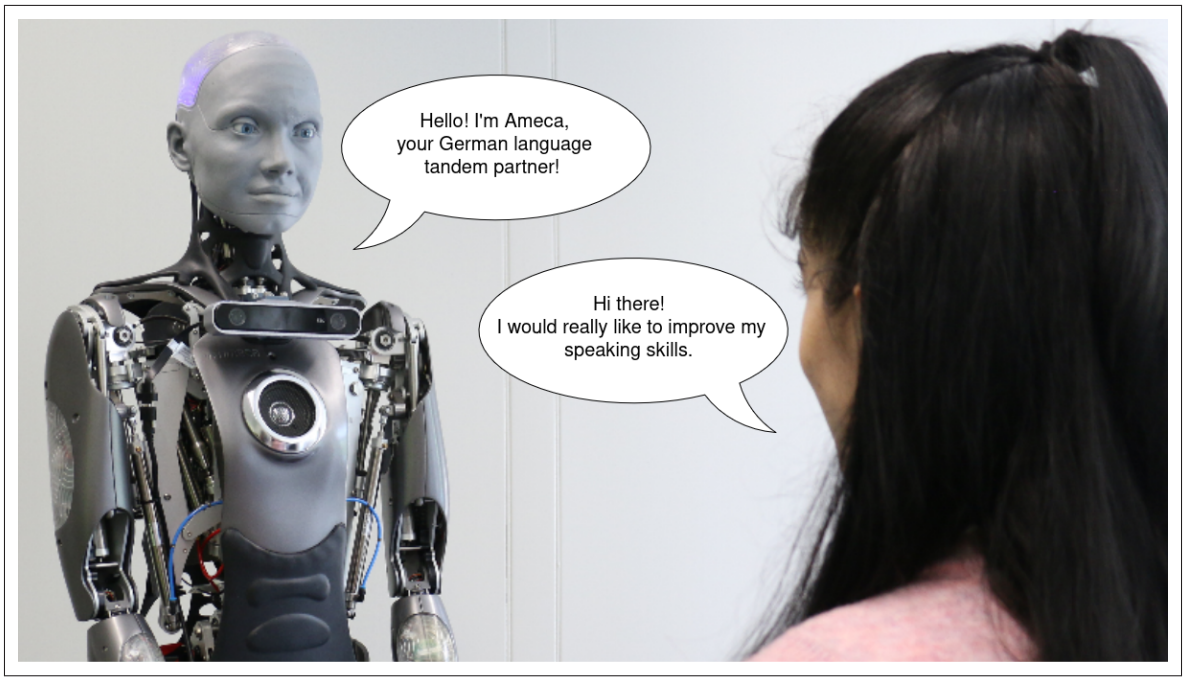


Figure 8.6: Ollama-based LLM-powered Gen-1 Ameca deployed for repeated robot-assisted German speaking practice.

The open-ended responses showed that participants valued the robot mainly as a non-judgemental speaking partner. Participant p1 wrote at Session 5 that it was useful to learn with Ameca since repeated questions did not create worry about frustration, while p4 described the system as "a judgement free way of learning german," and p6 wrote that speaking was easier "without people involved." Participants also emphasized topic familiarity, partner behaviour, and reduced evaluation concern, while practical limits remained visible in pacing, recognition, and continuity. At Session 5, p2 reported that Ameca sometimes misunderstood pronunciation and that response-flow timing was initially confusing; p5 described weak voice recognition; p8 wanted session-to-session memory; and p1 wanted stronger topic guidance. The robot role therefore supported social safety, while interaction quality still depended on timing, recognition, memory, and conversational structure.

The quantitative results showed positive descriptive trends without corrected statistical reliability. Speech anxiety was highly reliable at Session 1 ($\alpha = .951$) and Session 5 ($\alpha = .947$), while several secondary scales showed unstable reliability and were interpreted descriptively. Listening performance increased slightly from $M = 7.14$ ($SD = 0.69$) to $M = 7.43$ ($SD = 0.98$), peer-rated speaking increased from $M = 7.14$ ($SD = 1.86$) to $M = 7.29$ ($SD = 1.11$), and speech anxiety decreased from $M = 3.11$ ($SD = 1.01$) to $M = 2.99$ ($SD = 0.84$). Anthropomorphism increased from $M = 3.17$ ($SD = 0.89$) to $M = 3.66$ ($SD = 0.63$), likeability remained high from $M = 4.00$ ($SD = 0.82$) to $M = 4.14$ ($SD = 0.56$), and RALL attitudes increased from $M = 3.50$ ($SD = 0.82$) to $M = 3.79$ ($SD = 0.77$). Willingness stayed positive and unchanged at $M = 4.14$ in both sessions.

The paired comparisons found no corrected reliable changes from Session 1 to Session 5. Listening score, peer rating, speech anxiety, negative emotion, anthropomorphism,

likeability, and willingness did not show statistically reliable changes. RALL showed the strongest descriptive increase and a raw significant paired result, $t(6) = 2.596$, $p = .041$, but the result did not remain significant after Holm correction ($p_{\text{Holm}} = .327$). Figures 8.7 and 8.8 show the clearest descriptive changes in the multi-session study, while participant-level plots for all outcomes are reported in Appendix A.22.

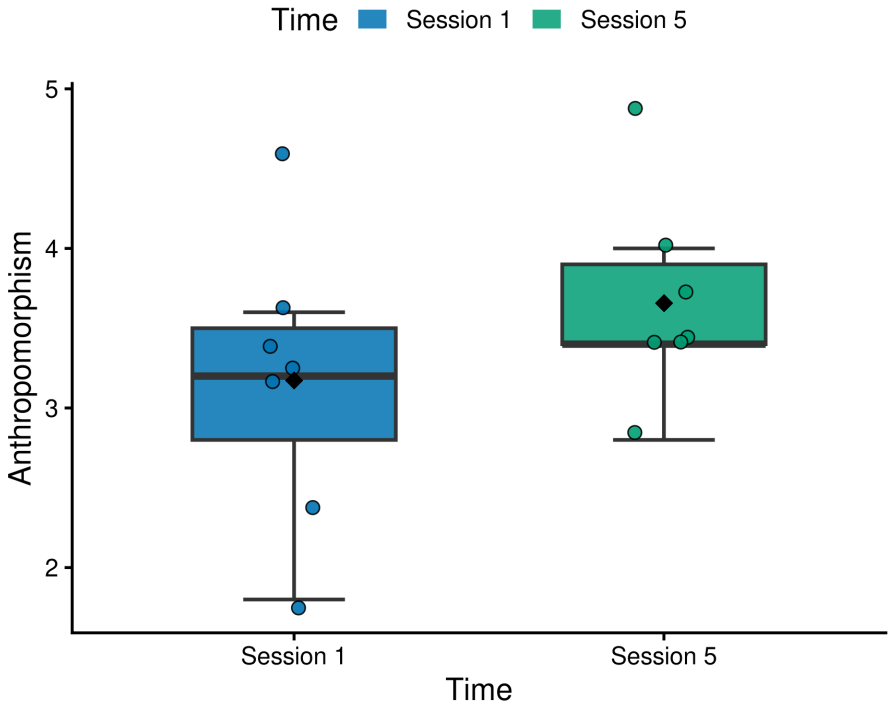


Figure 8.7: Anthropomorphism ratings from Session 1 to Session 5 in the multi-session tandem study.

Transcript analysis added a behavioural view of learner speech. An automated pipeline inspired by Verhelst et al. [Verhelst 25] extracted lexical diversity, lexical sophistication, average sentence length, verb density, error rate per 100 words, and CEFR estimates from user turns in Sessions 1 and 5. Lexical diversity increased for most learners, lexical sophistication remained relatively stable, sentence-level metrics changed only slightly, and error rates varied by participant. CEFR estimates rose from A2 to B1 for two learners, while most participants remained at B1. The transcript metrics are reported in Appendix A.23.

The multi-session study suggests that repeated robot-assisted speaking practice sustained a low-pressure practice setting more clearly than it produced short-term measured learning or anxiety reduction. Participants valued Ameca as non-judgemental and useful for practice, yet pacing, recognition, memory, and topic guidance limited the interaction. The discussion therefore returns to the 5A framework to explain why tandem partnership is promising only when social safety and interactional reliability develop together.

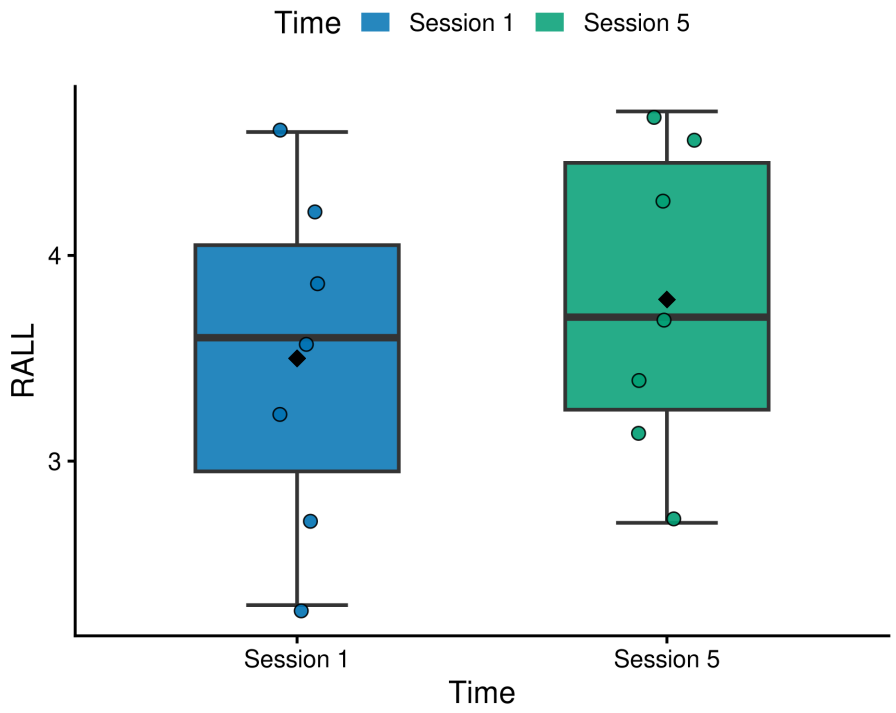


Figure 8.8: Attitudes toward robot-assisted language learning from Session 1 to Session 5 in the multi-session tandem study.

8.6 Discussion

The tandem-partner chapter showed that Ameca's value for German speaking practice came from the coupling of language generation with embodied presence, correction, pacing, and role framing. The expectation gap emerged when embodied presence, spoken dialogue, correction, topic support, pacing, memory-like continuity, and learner anxiety were experienced as one practice interaction. Learners did not encounter a delayed answer, weak correction, missed recognition, or repeated topic as isolated system errors. They experienced the robot as more or less supportive, patient, useful, or worth continuing with.

The survey established the affective boundary of the use case. Speaking-related anxiety was present in the target population, and German self-confidence predicted lower speech anxiety more clearly than course status, years in Germany, or age. Open-ended responses showed that comfort depended on the person, the topic, and the fear of being judged, which made the tandem role more appropriate than a tutor or examiner role. Ameca therefore entered the application as a potential low-pressure partner for speaking practice, not as a full instructional system.

The single-session study exposed the limits of one-off interaction. Audio-only, robot German, and robot German-English modes did not produce robust differences in lexical gain, anxiety, robot perception, expectations, RALL attitudes, or tandem evaluation. Likeability and perceived safety increased overall, while participant comments pointed to response delay, speech clarity, correction, and conversational style as the practical features shaping the experience. Embodiment and bilingual support improved the role only when interaction quality made the practice feel useful.

The multi-session study exposed the limits of repeated exposure without stronger continuity. Speech anxiety decreased slightly, anthropomorphism and RALL attitudes increased descriptively, likeability stayed high, and willingness remained positive, while corrected tests showed no reliable main change. Participants described Ameca as non-judgemental and useful for practice, while also asking for better pacing, recognition, memory, topic guidance, and structured support. Transcript metrics suggested small gains in lexical diversity and CEFR estimates for some learners, but the pattern remained exploratory and participant-specific.

The 5A framework explains the tandem result as an application-level alignment problem. Ameca's anthropomorphic body created social presence, its dialogue system created apparent conversational ability, and learner anxiety moved toward judgement, correction, pacing, and continuity. Alignment was strongest when the robot stayed within a peer-like practice role and weakest when delays, weak memory, or limited topic support made the interaction harder to sustain. Awareness remains necessary so learners understand that corrections, memory, and fluency are generated by a bounded robot system. The thesis conclusion can therefore treat the tandem partner as a viable use case for Ameca when social safety, technical reliability, and role boundaries remain readable together.

9. Discussion

This chapter discusses how Ameca's body, implemented abilities, and interaction roles shaped expectation mismatch across the thesis. The anthropomorphic expectation gap was examined through the 5A framework, with Ameca used as a full-body humanoid platform whose face, gaze, speech, and upper-body movement make system behaviour socially interpretable. The framework linked *Anthropomorphism*, *Apparent Ability*, and *Anxiety* as diagnostic dimensions with *Alignment* and *Awareness* as mitigation dimensions. The argument is grounded in small-scale empirical studies and system evaluations with $N = 929$ reported participants across the thesis. Across embodiment, dialogue, visual perception, behaviour generation, and tandem-language practice, Ameca showed that highly humanlike robots are evaluated through the relation between visible social expectation and implemented capability. The contribution is a system-level account of expectation management in a real humanoid platform, where appearance, sensing, generation, movement, role, explanation, and failure behaviour have to remain interpretable together.

9.1 Synthesis Across the Research Questions

RQ1 examined how Ameca's anthropomorphic embodiment shaped interlocutor expectations before and during interaction. Ameca's body shaped expectations before any module was evaluated because the face, eyes, skin, voice source, adult scale, gaze, mouth movement, and hand morphology made the robot socially addressable. That result supports anthropomorphism research showing that humanlike appearance and behaviour invite human categories, social inference, and role expectations [Złotowski 15, Perugia 23]. In Ameca, those expectations were prepared by science-fiction associations, online robot videos, institutional demonstrations, and rare public encounters, then intensified by the robot's immediate face-to-face presence. They were visible in YouTube comments about voice, gender, name, and apparent inner state, in hallway interactions where gaze and remembered names were read as social agency, and in public reactions where physical presence and eye contact produced fascination, discomfort, or care-role speculation. Chapter 4 showed that Ameca was not judged through appearance alone, since participants used voice, gaze, motion, affective framing, identity coherence, and prior media images to decide what kind of social entity the robot appeared to be. The embodiment findings

extend uncanny-valley accounts by showing that human likeness alone was less explanatory than the fit between face, motion, expression, timing, and perceived capability [Mori 12, Mathur 20, Saygin 12]. RQ1 is answered by showing that Ameca's embodiment created an initial social expectation that later modules either stabilised or weakened through robot behaviour.

RQ2 examined which technical limitations in perception, dialogue, gaze, and behaviour generation contributed to expectation mismatch in HRI. The technical chapters showed that mismatch appeared when partial ability was performed through a humanlike body. Kwon et al. argue that a few displayed capabilities can make users infer broader competence than the robot actually has, and Ameca made that capability-inference process visible across modules [Kwon 16]. ChitT5 produced local conversational responsiveness, while weak topic flow, identity drift, speech-recognition errors, and affective mismatches made dialogue instability socially visible. MultiSessionT5 added memory, retrieval, and emotion handling, while repeated interaction made vague memory, repetition, unsupported claims, and role confusion more consequential. The VQA system made seeing visible, while incorrect or vague answers were judged against the perceptual expectation created by eyes and gaze. Gaze management showed the same issue at the attention level, where camera bias and unstable target selection could be read as exclusion or staring. Behaviour generation added facial expression, mimicry, and pointing, while AU-inspired mappings, gesture recognition, motor range, timing, and deictic ambiguity constrained what Ameca could express or indicate. Horstmann and Kraemer showed that actual interaction skill strongly affects evaluations of sociability and competence, while the Ameca studies show that interaction skill is judged through the whole embodied pipeline across sensing, generation, timing, speech, gaze, and movement [Horstmann 20a]. RQ2 is answered by showing that technical limits become expectation mismatch when sensing, generation, timing, memory, gaze, and movement no longer support the social meaning suggested by the body.

RQ3 examined how alignment through exposure, bounded task design, explanation, and robot-delivered capability awareness reduced uncertainty and discomfort in interactions with Ameca. The mitigation evidence showed that expectation mismatch decreased when interlocutors could place Ameca's ability inside a bounded role. Physical-first exposure anchored expectations in scale, sound, timing, gaze, and mechanical limits before imagined capability could dominate the encounter. Robot-as-tutee framing made incomplete gesture learning interactionally acceptable because interlocutors could correct and encourage the robot within a learning role. Normative repair softened low-risk hallucination when the repair acknowledged the breakdown, while higher-stakes reminder failure required apology, explanation, and explicit user instruction. The tandem-language chapter showed the clearest application-level alignment, since Ameca was constrained as a peer-like practice partner for low-pressure German speaking practice. The role used anthropomorphic presence for social engagement while keeping Ameca outside teacher, therapist, examiner, and native-speaker replacement roles. RQ3 aligns with work on expectation management and explanations, where explanations can improve recovery from robot failure when they make system limits understandable [Yadollahi 25, Stange 22]. RQ3 is answered by showing that Alignment and Awareness work when exposure, task framing, prompts, local processing, failure repair, privacy boundaries, and behaviour make limits available during interaction.

9.2 The 5A Framework Across the Thesis

The 5A framework connects the empirical and technical chapters through one expectation-management logic. *Anthropomorphism* was strongest in Chapter 4, where Ameca's body, identity, voice, gender cues, affective framing, and public encounter shaped first interpretation. *Apparent Ability* increased in Chapters 5, 6, and 7, where speech, memory, VQA, gaze, facial expression, mimicry, and pointing made the robot appear able to listen, see, remember, attend, learn, and act. *Anxiety* appeared when interlocutors could not stabilise what Ameca perceived, understood, remembered, intended, or could reliably do. *Alignment* was strongest when the robot role constrained interpretation, as in tutee interaction, physical-first exposure, Hero-based gaze selection, and tandem practice. *Awareness* was strongest when system boundaries were made explicit through capability framing, local processing, prompt rules, privacy statements, failure protocols, and repair responses.

The 5A framework locates the social robot expectation gap inside the implemented interaction pipeline. Rosen's work frames expectations as dynamic, context-sensitive, and shaped by previous experience [Rosén 24a]. The Ameca-specific account developed here extends that expectation-gap logic toward a full-body humanoid by showing how gaps are generated through the coupling of body and architecture. In Ameca, a fluent answer depended on speech recognition, LLM inference, prompt constraints, memory, TTS, lip synchronisation, and facial behaviour; a gaze shift depended on cameras, tracking, active-speaker detection, smoothing, target mapping, and motor output; a pointing gesture depended on object detection, calibration, reachability, timing, and referent clarity. The 5A framework therefore turns expectation management from a user-study construct into a design and implementation concern for highly anthropomorphic robots.

9.3 Implications for HRI Design

The first implication is that highly anthropomorphic robots need bounded ability more than increased human likeness. Increasing fluency, memory, gaze, expression, or mimicry can improve interaction when the added behaviour remains consistent with the robot's sensing, control, and role. Kang et al. suggest that LLM-driven conversation can reduce uncanniness in hyper-realistic robots [Kang 25b]; the Ameca results support that direction when conversational improvement is paired with identity stability, capability boundaries, and failure handling. Fluent but ungrounded output can create deceptive capability impressions, as shown in LLM-enhanced care-robot risks where reminder functionality was falsely claimed [Ranisch 24]. For Ameca, prompt design therefore became a technical and ethical control layer for identity, capability boundaries, transparency, task scope, expectation management, failure protocol, privacy, user adaptation, and ethical limits [Ashok 26].

The second implication is that robot evaluation should measure whether Apparent Ability and actual capability remain aligned across interaction. A module can be accurate in isolation and still fail as HRI when its output violates embodied expectation. Speech-recognition errors can become apparent misunderstanding, VQA errors can become failed seeing, gaze bias can become exclusion, and gesture misrecognition can become failed learning. Evaluation should therefore connect technical performance, interlocutor interpretation, prior exposure, role framing, and repair behaviour in the same analysis. That

connection is especially necessary for GenAI-powered robots, where open-ended response generation can make a narrow architecture appear more general than it is.

The third implication is that Awareness should be interactional instead of a single disclaimer. Ameca's body continues to generate social meaning after consent forms, introductions, or capability statements are delivered. Awareness therefore has to be repeated through robot self-introduction, task framing, prompt boundaries, sensing disclosure, memory disclosure, local-processing choices, failure responses, and debriefing. A robot that explains what it can sense, what it cannot know, what it remembers, and why it failed gives interlocutors a more stable basis for trust than a robot that performs competence until breakdown.

9.4 Limitations

The thesis has four main limitations. First, several studies used small samples, online stimuli, exploratory prototypes, or descriptive analyses, so the results should be read as context-specific evidence across Ameca studies rather than as one causal model of Ameca interaction. Second, the work focused on one highly anthropomorphic robot in one university laboratory setting, which limits direct generalisation to abstract, mobile, zoomorphic, tabletop, or screen-based robots. Third, many systems were implemented under real platform constraints, including Gen-1 actuation, fixed base, camera placement, proprietary interfaces, response latency, limited training data, and changing LLM infrastructure. These constraints strengthen ecological validity for Ameca while limiting clean experimental isolation. Fourth, the 5A framework was used as an organising lens and design diagnostic. Future comparative work across robots, tasks, populations, and longitudinal deployments is needed before the framework can be treated as a stronger evaluation and design method.

9.5 Future Work

Future work is organised around implemented extensions, broader comparative datasets, and longer-term adaptive expectation management. Ongoing work extends the thesis in five directions. A fully open-weight multimodal pipeline combines local dialogue, VQA, memory, self-RAG, and online information in one architecture inspired by MultiSessionT5 and the VQA system [Esiagu 26, Mauliana 25, Rao 24, Kang 25b]. Task-based robot-anxiety studies use VLM and VAP functionality to point at the correct pill while testing the effects of robot failure [Jeong 26]. A planned older-adult public study extends the thesis beyond the main university tandem-use case by testing how public care-oriented framing shapes responses to Ameca selecting pills in a care or home-support scenario. Retrospective interviews with students who built Ameca systems will examine how expectation alignment depends on the type of technical or empirical work they performed with the robot. Larger in-person voice-emotion-onset studies will focus on gendered voice, positive facial expressions, and neutral sentences. A mirror-based self-evaluation setup will let Ameca generate AU-directed facial expressions, capture its own face, and use model feedback to build a robot-specific facial-expression dataset.

Scope-extending work should collect open HRI datasets for multiparty gaze, accented English and German TTS detection, real-time mimicry, and live facial-expression transfer.

Comparative studies should also examine Ameca's embodiment effects against other robot body types, building on the online morphology studies that already started that comparison [Berns 24, Esposito 25]. A longer-term direction is adaptive expectation management, where a robot estimates interlocutor uncertainty during interaction and selects explanations, behaviour intensity, memory disclosure, or repair strategy before mismatch becomes discomfort.

10. Conclusion

The research scope centred on how a highly anthropomorphic humanoid robot can remain understandable, usable, and acceptable when its appearance suggests more social, perceptual, and cognitive capability than the implemented system can reliably support. Ameca made that question visible because its face, gaze, voice, speech, skin, and upper-body movement invite interlocutors to treat system behaviour as socially meaningful before the architecture behind that behaviour is known. The 5A framework was developed to organise that mismatch. *Anthropomorphism*, *Apparent Ability*, and *Anxiety* diagnose how the expectation gap emerges; *Alignment* and *Awareness* describe how the gap can be mitigated through bounded roles, calibrated exposure, explicit system boundaries, and interactional repair.

The first contribution concerns embodiment. Ameca's face, gaze, voice, skin, movement, and adult-scale presence led interlocutors to infer attention, identity, emotion, memory, and understanding before the technical system had demonstrated what it could support. The embodiment studies showed that uncanniness and discomfort were not explained by human likeness alone. They emerged from mismatches between appearance, motion, voice, expression, role, timing, and interaction quality. Ameca therefore became a useful platform for studying anthropomorphic expectation because the robot's body made the gap between social appearance and implemented capability difficult to ignore.

The second contribution concerns implemented ability. Speech recognition, LLM dialogue, memory, retrieval, VQA, gaze management, facial expression, gesture mimicry, and pointing made Ameca appear more capable, while errors in recognition, timing, memory, gaze allocation, expression onset, or referent selection became socially meaningful because they were performed by a humanlike body. The technical contribution of the thesis therefore lies in building and analysing Ameca systems as expectation-management structures, rather than as isolated modules. Each system added interactional possibility, but each system also raised the standard for what interlocutors expected the robot to perceive, remember, understand, or do next.

The third contribution concerns mitigation. Expectation mismatch decreased when Ameca's role, behaviour, and system boundaries became easier to place. Physical-first exposure, robot-as-tutee framing, normative repair, local LLM control, prompt-defined

identity, privacy and failure boundaries, Hero-based gaze selection, and bounded tandem-language practice all made the robot's limits more available during interaction. The tandem-language studies provided the most complete application of the argument because Ameca became useful as a low-pressure German-speaking partner when its role remained peer-like, supportive, and technically constrained.

The thesis therefore shows that meaningful interaction with highly humanlike robots depends less on increasing human likeness and more on aligning apparent ability, actual capability, interaction role, and failure behaviour. Ameca became more understandable when its role was bounded, when its sensing and memory limits were explained, when dialogue was constrained by identity and failure rules, and when visible behaviour was timed to perception and task state. The same robot became harder to place when fluent speech, gaze, facial expression, mimicry, or pointing implied more understanding, attention, or control than the system could support.

Researchers and designers working with highly anthropomorphic robots should therefore treat appearance, sensing, dialogue, memory, movement, prompts, explanations, and repair as one interaction system. Humanlike morphology cannot be separated from the technical architecture that drives it, because interlocutors experience the robot as one embodied agent. A highly anthropomorphic robot can support meaningful HRI when it remains bounded, transparent, and accountable about what it can perceive, generate, remember, and do.

> *"'The idea behind this video was to portray how a robot can only be as good as the machine learning model behind it." I think you should have a disclaimer up front.*
>
> @EndingSimple, comment on *Ameca: Unpredictable Responses from the Humanoid Robot! YouTube Video*

The thesis closes with the same requirement that highly anthropomorphic robots such as Ameca can support meaningful interaction when interlocutors can recognise both the social invitation of the body and the technical boundaries behind it.

Bibliography

[Addlesee 24] A. Addlesee, N. Cherakara, N. Nelson, D. Hernández García, N. Gunson, W. Sieińska, M. Romeo, C. Dondrup, O. Lemon, "A multi-party conversational social robot using llms", in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 1273–1275.

[Afouras 18] T. Afouras, J. S. Chung, A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition", *arXiv preprint arXiv:1809.00496*, 2018.

[Ahmed 24] E. Ahmed, O. uruk, J. Hamari, "Human–robot companionship: Current trends and future agenda", *International Journal of Social Robotics*, vol. 16, no. 8, pp. 1809–1860, 2024.

[Aida 94] Y. Aida, "Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese", *The modern language journal*, vol. 78, no. 2, pp. 155–168, 1994.

[Akinci 24] L. B. Akinci. "Perceived Consciousness in Human-Robot Interactions: Contributions of a GWT Architecture and Eye Gaze". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Alemi 15] M. Alemi, A. Meghdari, M. Ghazisaedy, "The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition", *International Journal of Social Robotics*, vol. 7, pp. 523–535, 2015.

[Alves-Oliveira 15] P. Alves-Oliveira, T. Ribeiro, S. Petisca, E. Di Tullio, F. S. Melo, A. Paiva, "An empathic robotic tutor for school classrooms: Considering expectation and satisfaction of children as end-users", in *International Conference on social robotics*, Springer. 2015, pp. 21–30.

[Ardila 20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, "Common voice: A massively-multilingual speech corpus", in *Proceedings of the twelfth language resources and evaluation conference*. 2020, pp. 4218–4222.

[Arlinghaus 25] C. S. Arlinghaus, A. Ashok, K. Berns, G. W. Maier, "From GPT to Open-Source LLMs: Advancing LAIC for More Efficient Qualitative Analysis", 2025.

[Ashkenazi 25] S. Ashkenazi, G. Skantze, J. Stuart-Smith, M. E. Foster, "The Multilingual Student Support Robot", in *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. Eindhoven, Netherlands, 2025. Accepted.

[Ashok 22] A. Ashok, J. Pawlak, S. Paplu, Z. Zafar, K. Berns, "Paralinguistic cues in speech to adapt robot behavior in human-robot interaction", in *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, IEEE. 2022, pp. 01–06.

[Ashok 23] A. Ashok, S. Paplu, K. Berns, "Social robot dressing style: An evaluation of interlocutor preference for university setting", in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE. 2023, pp. 2499–2504.

[Ashok 24a] A. Ashok, M. Mauliana, K. Berns, "Chit-chat with humanoid: llms for university icebreaker session", in *International Conference on Social Robotics*, Springer. 2024, pp. 258–267.

[Ashok 24b] A. Ashok, V. Tolj, K. Berns, "University copycat game: Observational learning with robot tutee", in *2024 IEEE international conference on advanced robotics and its social impacts (ARSO)*, IEEE. 2024, pp. 61–66.

[Ashok 25a] A. Ashok, B. Bruno, T. Helf, K. Berns, ""Thanks for the Practice!": LLM-Powered Social Robot as Tandem Language Partner at University", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1221–1226.

[Ashok 25b] A. Ashok, L. Nguyen, K. Berns, ""A Glimpse Into My World": Empathy Towards Emotional Robot Backstories at University", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1227–1231.

[Ashok 26] A. Ashok, F. Babel, P. Holthaus, R. Khot, K. B. Kelly, F. I. Dogan, K. Berns, S. Rossi, M. Lee, G. Laban. "Why Did My Robot Just Change Personality? Prompting Guidelines for a Grounded Robot Persona in LLM-Based HRI". Manuscript under review at 2026 35th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2026.

[Assael 16] Y. M. Assael, B. Shillingford, S. Whiteson, N. De Freitas, "Lipnet: End-to-end sentence-level lipreading", *arXiv preprint arXiv:1611.01599*, 2016.

[Babel 21] F. Babel, J. Kraus, L. Miller, M. Kraus, N. Wagner, W. Minker, M. Baumann, "Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity", *International Journal of Social Robotics*, vol. 13, no. 6, pp. 1485–1498, 2021.

[Baber 06] K. M. Baber, C. J. Tucker, "The social roles questionnaire: A new approach to measuring attitudes toward gender", *Sex Roles*, vol. 54, no. 7, pp. 459–467, 2006.

[Bainbridge 08] W. A. Bainbridge, J. Hart, E. S. Kim, B. Scassellati, "The effect of presence on human-robot interaction", in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE. 2008, pp. 701–706.

[Balaji 25] N. Balaji. "Robot Ears: A Review of Automatic Speech Recognition for Human-Robot Interaction". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Banaeian 21] H. Banaeian, I. Gilanlioglu, "Influence of the NAO robot as a teaching assistant on university students' vocabulary learning and attitudes", *Australasian Journal of Educational Technology*, vol. 37, no. 3, pp. 71–87, 2021.

[Bandura 61] A. Bandura, D. Ross, S. A. Ross, "Transmission of aggression through imitation of aggressive models.", *The Journal of Abnormal and Social Psychology*, vol. 63, no. 3, p. 575, 1961.

[Bandura 69] A. Bandura, "Social-learning theory of identificatory processes", *Handbook of socialization theory and research*, vol. 213, p. 262, 1969.

[Banks 21] J. Banks, K. Koban, "Framing effects on judgments of social robots'(im) moral behaviors", *Frontiers in Robotics and AI*, vol. 8, p. 627233, 2021.

[Bartneck 04] C. Bartneck, J. Forlizzi, "A design-centred framework for social human-robot interaction", in *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759)*, IEEE. 2004, pp. 591–594.

[Bartneck 09a] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots", *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[Bartneck 09b] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots", *International journal of social robotics*, vol. 1, pp. 71–81, 2009.

[Beck 26] A. K. t. Beck, A. t. Ashok, K. Berns, T. Lachmann. "When Female-Voiced Robots Talk Back: Effects of Assertive vs. Apologetic Response to Mistreatment on Bystander Perception". *Equal contribution. Manuscript under review at 2026 35th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)., 2026.

[Behm-Morawitz 09] E. Behm-Morawitz, D. Mastro, "The effects of the sexualization of female video game characters on gender stereotyping and female self-concept", *Sex roles*, vol. 61, no. 11, pp. 808–823, 2009.

[Belkada 21] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, A. Alahi, "Do pedestrians pay attention? eye contact detection in the wild", *arXiv preprint arXiv:2112.04212*, 2021.

[Bell 25] H. Bell, J. Kwesi, H. Laabadli, P. Emami-Naeini, "" Is it always watching? Is it always listening?" Exploring Contextual Privacy and Security Concerns Toward Domestic Social Robots", *arXiv preprint arXiv:2507.10786*, 2025.

[Ben-David 11] B. M. Ben-David, P. H. Van Lieshout, T. Leszcz, "A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury", *Brain injury*, vol. 25, no. 2, pp. 206–220, 2011.

[Berns 24] K. Berns, A. Ashok, ""You Scare Me": The Effects of Humanoid Robot Appearance, Emotion, and Interaction Skills on Uncanny Valley Phenomenon", in *Actuators*, vol. 13, no. 10, MDPI. 2024, p. 419.

[Billing 23] E. Billing, J. Rosén, M. Lamb, "Language models for human-robot interaction", in *Companion of the 2023 ACM/IEEE international conference on human-robot interaction*. 2023, pp. 905–906.

[Blair 25] A. Blair, P. Gregory, M. E. Foster, "Observations of atypical users from a pilot deployment of a public-space social robot in a church", *arXiv preprint arXiv:2508.16622*, 2025.

[Blair 26] A. Blair, M. E. Foster, P. Gregory, K. Hindriks, "Where Are They? The Absence of Social Robots in Real-World Environments", in *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026, pp. 96–100.

[Borsci 22] S. Borsci, A. Malizia, M. Schmettow, F. Van Der Velde, G. Tariverdiyeva, D. Balaji, A. Chamberlain, "The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents", *Personal and ubiquitous computing*, vol. 26, no. 1, pp. 95–119, 2022.

[Brammerts 96] H. Brammerts, "Language learning in tandem using the Internet", *Telecollaboration in foreign language learning*, pp. 121–130, 1996.

[Bransky 26] K. Bransky, P. Sweetser, P. Holthaus, G. Laban, K. Winkle, N. Akalin, A. Ashok, J. Lee, R. Khot, A. Bejarano, J. Thijn, R. K. Moore, M. Jang, J. Fischer, M. Lee, "Designing Artificial Identity: The Identity Design Framework and Research Agenda", in *Designing Interactive Systems Conference (DIS '26)*. New York, NY, USA: ACM, 2026. Note: In press.

[Breazeal 03] C. Breazeal, "Emotion and sociable humanoid robots", *International journal of human-computer studies*, vol. 59, no. 1–2, pp. 119–155, 2003.

[Breazeal 16] C. Breazeal, K. Dautenhahn, T. Kanda, *Social Robotics*, Springer International Publishing, 2016, pp. 1935 – 1972.

[Breininger 23] R. Breininger. "Solving Puzzles Together: Piecing Together Functional Consciousness in a Humanoid Robot Assistant". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Burgoon 88] J. K. Burgoon, J. L. Hale, "Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors", *Communications Monographs*, vol. 55, no. 1, pp. 58–79, 1988.

[Bynion 20] T.-M. Bynion, M. T. Feldner, "Self-assessment manikin", in *Encyclopedia of personality and individual differences*, Springer, 2020, pp. 4654–4656.

[Carrasco Melo 21] J. A. Carrasco Melo. "Developing Self-Awareness: The Road Towards Emotional Robots". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2021.

[Chakravarty 16] P. Chakravarty, T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video", in *European conference on computer vision*, Springer. 2016, pp. 285–301.

[Chen 18] S. Chen, Y. Liu, X. Gao, Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices", in *Chinese conference on biometric recognition*, Springer. 2018, pp. 428–438.

[Chen 23] N. Chen, X. Liu, Y. Zhai, X. Hu, "Development and validation of a robot social presence measurement dimension scale", *Scientific Reports*, vol. 13, no. 1, p. 2911, 2023.

[Chiang 23] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality", March 2023.

[Chong 20] E. Chong, E. Clark-Whitney, A. Southerland, E. Stubbs, C. Miller, E. L. Ajodan, M. R. Silverman, C. Lord, A. Rozga, R. M. Jones et al, "Detection of eye contact with deep neural networks is as accurate as human experts", *Nature communications*, vol. 11, no. 1, p. 6386, 2020.

[Chowdhury 23] A. A. Chowdhury. "Hallway Interactions with Emah: Exploring the Role of Social Robots in Corridor Engagement". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Chung 18] J. S. Chung, A. Nagrani, A. Zisserman, "Voxceleb2: Deep speaker recognition", *arXiv preprint arXiv:1806.05622*, 2018.

[Chung 24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, "Scaling Instruction-Finetuned Language Models", *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[Ciardo 26] F. Ciardo, M. Bolis, L. Lapomarda, C. Roselli, E. Datteri, "Are social robots inherently deceptive?", *Sistemi intelligenti*, vol. 38, no. 1, pp. 203–225, 2026.

[Cooper 20] S. Cooper, A. Di Fava, C. Vivas, L. Marchionni, F. Ferro, "ARI: The social assistive robot and companion", in *2020 29th IEEE International conference on robot and human interactive communication (RO-MAN)*, IEEE. 2020, pp. 745–751.

[Crawshaw 20] M. Crawshaw, "Multi-task learning with deep neural networks: A survey", *arXiv preprint arXiv:2009.09796*, 2020.

[Cruces 24] A. Cruces, A. Jerez, J. P. Bandera, A. Bandera, "Socially assistive robots in smart environments to attend elderly people—a survey", *Applied Sciences*, vol. 14, no. 12, p. 5287, 2024.

[Das 17] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, D. Batra, "Visual dialog", in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 326–335.

[Das 25] S. Das. "AI-Based 3D Perception in Humanoid Robots". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Dautenhahn 07] K. Dautenhahn, "Socially intelligent robots: dimensions of human–robot interaction", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, pp. 679 – 704, 2007.

[De Jong 18] M. De Jong, K. Zhang, A. M. Roth, T. Rhodes, R. Schmucker, C. Zhou, S. Ferreira, J. Cartucho, M. Veloso, "Towards a robust interactive and learning social robot", in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2018, pp. 883–891.

[Dennler 25] N. Dennler, M. Kian, S. Nikolaidis, M. Matarić, "Designing robot identity: The role of voice, clothing, and task on robot gender perception", *International Journal of Social Robotics*, vol. 17, no. 4, pp. 707–728, 2025.

[Desplanques 20] B. Desplanques, J. Thienpondt, K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification", *arXiv preprint arXiv:2005.07143*, 2020.

[Diehl 24] C. Diehl, S. Trittler, "Highly skilled and highly skeptical? How education and origin shape newcomers' relationship with their new home", *Journal of Ethnic and Migration Studies*, vol. 50, no. 7, pp. 1777–1802, 2024.

[Donnermann 24] M. Donnermann, P. Heinzmann, B. Lugrin, "Meet or Call my Robotic Tutor?-The Effect of a Physically vs. Virtually Present Social Robot on Learning Outcomes, Engagement and Perception", in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 422–426.

[Dosovitskiy 20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*, 2020.

[Ekman 02] P. Ekman, W. V. Friesen, J. C. Hager. *Facial Action Coding System: Manual and Investigator's Guide*. Research Nexus, Network Information Research Corporation, Salt Lake City, UT. 2002. Includes FACS Manual, Investigator's Guide, example photos, and videos. ISBN: 0-931835-01-1.

[Ekman 87] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti et al, "Universals

and cultural differences in the judgments of facial expressions of emotion.", *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[Esiagu 26] C. L. Esiagu. "Multimodal Emotion Recognition and Congruent Expressive Response Generation for the Ameca Social Humanoid Robot". Unpublished master's thesis, University of Kaiserslautern-Landau, due September 2026, reviewed by Karsten Berns and supervised by Ashita Ashok, 2026.

[Esposito 25] R. Esposito, A. Rossi, S. Rossi, "Deception in HRI and its Implications: a Systematic Review", *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 3, pp. 1–26, 2025.

[Esposito 26] R. Esposito, A. Ashok, A. Rossi, K. Berns, S. Rossi. "Interpreting Robot Failure and Deception: The Role of Human-Likeness". Manuscript under review at 2026 35th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2026.

[Eymann 25] V. Eymann, T. Lachmann, F. Hekele, A. Ashok, K. Berns, D. Czernochowski, "Does ChatGPT know what robots are thinking?-Evaluation of human-created vs. AI-generated scientific ideas", in *Proceedings of the 36th Annual Conference of the European Association of Cognitive Ergonomics (EACE)*. 2025, pp. 1–5.

[Farimani 24] M. H. M. Farimani, "Uncanny Valley Theory in the Design of Robots: Bridging the Gap Between Affinity and Aversion", july 2024. Design article and figure used with permission from the author.

[Farnsworth 22] B. Farnsworth, "Facial Action Coding System (FACS) – A Visual Guidebook", `https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/`, 2022. Accessed: 2026-05-07.

[Föbker 17] S. Föbker, D. Imani, "The role of language skills in the settling-in process–Experiences of highly skilled migrants' accompanying partners in Germany and the UK", *Journal of Ethnic and Migration Studies*, vol. 43, no. 16, pp. 2720–2737, 2017.

[Fong 03] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A survey of socially interactive robots", *Robotics and autonomous systems*, vol. 42, no. 3–4, pp. 143–166, 2003.

[Fortunati 22] L. Fortunati, A. M. Manganelli, J. Höflich, G. Ferrin, "How the social robot Sophia is mediated by a YouTube video", *New Media & Society*, vol. 26, pp. 3390 – 3409, 2022.

[Fronteddu 21] G. Fronteddu, S. Porcu, A. Floris, L. Atzori, "Dataset for dynamic hand gesture recognition systems", *IEEE Dataport*, 2021.

[Fuchs 24] T. Fuchs, "Understanding Sophia? On human interaction with artificial agents", *Phenomenology and the Cognitive Sciences*, vol. 23, no. 1, pp. 21–42, 2024.

[Funes Mora 14] K. A. Funes Mora, F. Monay, J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras", in *Proceedings of the symposium on eye tracking research and applications*. 2014, pp. 255–258.

[García-Martínez 25] J. García-Martínez, J. J. Gamboa-Montero, Á. Castro-González, J. C. Castillo, "Are Robots More Engaging When They Respond to Joint Attention? Findings from a Turn-Taking Game with a Social Robot", *Applied Sciences*, vol. 15, no. 15, p. 8684, 2025.

[Ghamati 24] K. Ghamati, A. Zaraki, F. Amirabdollahian, "ARI humanoid robot imitates human gaze behaviour using reinforcement learning in real-world environments", in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, IEEE. 2024, pp. 653–660.

[Gil Guevara 22] F. S. Gil Guevara. "Accent Based Speech Emotion Recognition for Human-Robot Interaction". Unpublished bachelor's thesis, University of Kaiserslautern-Landau, March 2022, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Gillet 22] S. Gillet, M. T. Parreira, M. Vázquez, I. Leite, "Learning gaze behaviors for balancing participation in group human-robot interactions", in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2022, pp. 265–274.

[Gliwa 19] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization", *arXiv preprint arXiv:1911.12237*, 2019.

[Goede 22] S. Goede. "Modulating Robot Behaviour During Self-Introduction Scenario". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Gopalakrishnan 23] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tur, "Topical-chat: Towards knowledge-grounded open-domain conversations", *arXiv preprint arXiv:2308.11995*, 2023.

[Groom 11] V. Groom, V. Srinivasan, C. L. Bethel, R. Murphy, L. Dole, C. Nass, "Responses to robot social roles and social role framing", in *2011 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE. 2011, pp. 194–203.

[Gu 17] J. Gu, X. Yang, S. De Mello, J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network", in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1548–1557.

[Gurung 24] D. Gurung. "Event-Based Robot Behavior for Human-Robot Interaction". Unpublished master's project, University of Kaiserslautern-Landau, September 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Han 08] J.-H. Han, M.-H. Jo, V. Jones, J.-H. Jo, "Comparative study on the educational use of home robots for children", *Journal of Information Processing Systems*, vol. 4, no. 4, pp. 159–168, 2008.

[Hart 06] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later", in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9, Sage publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[Heisler 24] M. Heisler, C. Becker-Asano, "Learning to control an android robot head for facial animation", in *Companion of the 2024 ACM/IEEE international conference on human-robot interaction*. 2024, pp. 530–534.

[Heisler 25] M. Heisler, C. Becker-Asano, "An iPhone Pro is All You Need: Mimicking Facial Expressions on an Android Robot Head", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1347–1351.

[Helf 25] T. Helf. "Managing Anxiety in International Students Learning the German Language with LLM-Powered Social Robots". Unpublished bachelor's thesis, University of Kaiserslautern-Landau, April 2025, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Hempel 24] T. Hempel, M. Jung, A. A. Abdelrahman, A. Al-Hamadi, "Nitec: Versatile hand-annotated eye contact dataset for ego-vision interaction", in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2024, pp. 4437–4446.

[Henschel 21] A. Henschel, G. Laban, E. S. Cross, "What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You", *Current Robotics Reports*, vol. 2, pp. 9 – 19, 2021.

[Hieida 22] C. Hieida, T. Nagai, "Survey and perspective on social emotions in robotics", *Advanced Robotics*, vol. 36, no. 1–2, pp. 17–32, 2022.

[Hirth 12a] J. Hirth, *Towards socially interactive robots-designing an emotion-based control architecture*, Dr. Hut, 2012.

[Hirth 12b] J. Hirth, N. Schmitz, K. Berns, "Playing tangram with a humanoid robot", in *ROBOTIK 2012; 7th German Conference on Robotics*, VDE. 2012, pp. 1–6.

[Hold, Markus 23a] Hold, Markus, "Ameca Creation Process", `https://drive.google.com/file/d/1zDBpQuCDKwX6AiucEXXeOs-SP1Pbl13x/view`, 2023. Engineered Arts. Mentioned in the Ameca wiki. Accessed: 2026-05-07.

[Hold, Markus 23b] Hold, Markus, "Ameca Gen 1 - Degrees of Movement", `https://drive.google.com/file/d/1ZJTj8zdVRRBwnnd-1B--qJO6qxCSryKY/view`, 2023. Engineered Arts. Mentioned in the Ameca wiki. Accessed: 2026-05-07.

[Hold, Markus 23c] Hold, Markus, "Ameca Gen 1 - Set-up Guide", `https://drive.google.com/file/d/15Z50bdgBtfbB-aEX_LprP0-7-Cw9ZZ4-/view`, 2023. Engineered Arts. Mentioned in the Ameca wiki under brochure for the Ameca platform. Accessed: 2026-05-07.

[Hold, Markus 23d] Hold, Markus, "Ameca Gen 1 Brochure", `https://drive.google.com/file/d/1Pws6QQTRrgZgAwbrC_xcHdCYPecGyPR4/view`, 2023. Engineered Arts. Mentioned in the Ameca wiki. Accessed: 2026-05-07.

[Holthaus 23] P. Holthaus, T. Schulz, G. Lakatos, R. Soma, "Communicative robot signals: presenting a new typology for human-robot interaction", in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 132–141.

[Holthaus 26] P. Holthaus, G. Lakatos, M. S. Gou, S. M. Espanol, L. Wood, B. Robins, H. Araujo, M. Mousavi, F. Amirabdollahian, "Kaspar Explains: The Impact of Explanation on Human-Robot Trust using an Educational Platform", *International Journal of Social Robotics*, 2026.

[Hone 00] K. S. Hone, R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)", *Natural Language Engineering*, vol. 6, no. 3–4, pp. 287–303, 2000.

[Hoorn 24] J. F. Hoorn, I. S. Huang, "The media inequality, uncanny mountain, and the singularity is far from near: Iwaa and Sophia robot versus a real human being", *International Journal of Human-Computer Studies*, vol. 181, p. 103142, 2024.

[Horstmann 19] A. C. Horstmann, N. C. Krämer, "Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment", *Frontiers in psychology*, vol. 10, p. 939, 2019.

[Horstmann 20a] A. C. Horstmann, N. C. Krämer, "Expectations vs. actual behavior of a social robot: An experimental investigation of the effects of a social robot's interaction skill level and its expected future role on people's evaluations", *PloS one*, vol. 15, no. 8, p. e0238133, 2020.

[Horstmann 20b] A. C. Horstmann, N. C. Krämer, "When a robot violates expectations: The influence of reward valence and expectancy violation on people's evaluation of a social robot", in *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*. 2020, pp. 254–256.

[Horwitz 01] E. Horwitz, "Language anxiety and achievement", *Annual review of applied linguistics*, vol. 21, pp. 112–126, 2001.

[Horwitz 86] E. K. Horwitz, M. B. Horwitz, J. Cope, "Foreign language classroom anxiety", *The Modern language journal*, vol. 70, no. 2, pp. 125–132, 1986.

[Hu 22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al, "Lora: Low-rank adaptation of large language models.", *ICLR*, vol. 1, no. 2, p.$\tilde{3}$, 2022.

[Hu 24] X. Hu, X. Zhang, S. McGeown, "Foreign language anxiety and achievement: A study of primary school students learning English in China", *Language teaching research*, vol. 28, no. 4, pp. 1594–1615, 2024.

[Huang 08] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments", in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.

[Huang 13] C.-M. Huang, B. Mutlu, "The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior", *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 80–102, 2013.

[Huang 24] Y.-C. Huang. "Self-Awareness in Social Robots". Unpublished master's project, University of Kaiserslautern-Landau, September 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Huang 25] Y.-C. Huang. "Multi-Task Vision Model for Human-Robot Interaction". Unpublished master's thesis, University of Kaiserslautern-Landau, October 2025, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Hwang 20] J. Hwang, J. Kim, A. Ahmadi, M. Choi, J. Tani, "Dealing With Large-Scale Spatio-Temporal Patterns in Imitative Interaction Between a Robot and a Human by Using the Predictive Coding Framework", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 5, pp. 1918–1931, 2020.

[Irfan 25] B. Irfan, G. Skantze, "Between you and me: Ethics of self-disclosure in human-robot interaction", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1357–1362.

[Islam 23] M. E. Islam. "Dynamic Hand Gesture Recognition for Human-Robot Interaction". Unpublished master's thesis, University of Kaiserslautern-Landau, September 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Jackson 22] W. Jackson, "Quick Intro to Virtual Ameca", Vimeo video, 2022. Engineered Arts. Mentioned in the Ameca wiki under Programming: Creating new movements. Accessed: 2026-05-07.

[Jafarova 25] L. Jafarova. "Artificial Voice on Humanoid Robot Ameca". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Jahn 25] E. T. Jahn, J. Renardias, N. Hille, A. Ashok, M.-S. Edelmann, R. Wieching, D. Randall, K. Berns, V. Wulf, "Understanding Human–AI Interaction in Aging Populations: Usability Differences between Robots and Avatars", in *2025 IEEE Cyber Science and Technology Congress (CyberSciTech)*, IEEE. 2025, pp. 788–793.

[Jahn 26] E. T. Jahn, A. Ashok, M.-S. Edelmann, N. Hille, M. Manavi, N. Kushina, R. Wieching, D. Randall, K. Berns, V. Wulf, "Embodied AI in the Wild: Comparing Older Adults' Interactions with an Avatar and a Humanoid Robot in a Public Space", in *Proceedings of the Designing Interactive Systems Conference*, ser. DIS '26. New York, NY, USA: ACM, 2026, p.1̃7. Note: In press.

[Jeong 25] H. Jeong, R. Ghosh. "Closing the Social Robot Expectation Gap: A Teachable Approach to Social and Task-Based Robot Behavior". Unpublished computation lab rotation, University of Kaiserslautern-Landau, equal contribution, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Jeong 26] H. Jeong. "The Influence of Social Robot Cues on Human Anxiety and Trust in Safety-Critical Task". Unpublished master's thesis, University of Kaiserslautern-Landau, due August 2026, reviewed by Karsten Berns and Ann-Kathrin Beck and supervised by Ashita Ashok, 2026.

[Jois 22] V. M. Jois. “Review of Cognitive Architectures for Enhanced Decision Making”. Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Kamelabad 25] A. M. Kamelabad, E. Inoue, G. Skantze, “Comparing Monolingual and Bilingual Social Robots as Conversational Practice Companions in Language Learning”, in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 829–838.

[Kanda 04] T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, “Interactive robots as social partners and peer tutors for children: A field trial”, *Human–Computer Interaction*, vol. 19, no. 1–2, pp. 61–84, 2004.

[Kang 24] H. Kang, M. Ben Moussa, N. M. Thalmann, “Nadine: A large language model-driven intelligent social robot with affective capabilities and human-like memory”, *Computer Animation and Virtual Worlds*, vol. 35, no. 4, p. e2290, 2024.

[Kang 25a] H. Kang, T. F. dos Santos, M. B. Moussa, N. M. Thalmann, “From Sci-Fi to reality: exploring teenagers’ perceptions and ideas of AI robots”, *The Visual Computer*, vol. 41, pp. 11629 – 11640, 2025.

[Kang 25b] H. Kang, T. F. d. Santos, M. B. Moussa, N. Magnenat-Thalmann, “Mitigating the Uncanny Valley Effect in Hyper-Realistic Robots: A Student-Centered Study on LLM-Driven Conversations”, *arXiv preprint arXiv:2503.16449*, 2025.

[Kartynnik 19] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann, “Real-time facial surface geometry from monocular video on mobile GPUs”, *arXiv preprint arXiv:1907.06724*, 2019.

[Kendon 04] A. Kendon, “Gesture: Visible Action as Utterance, Cambridge Uni”, 2004.

[Khalil 22] H. Khalil. “A Comparative Study of ACT-R vs. Non-ACT-R Based Chatbot Systems”. Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Khoo 23] W. Khoo, L.-J. Hsu, K. J. Amon, P. V. Chakilam, W.-C. Chen, Z. Kaufman, A. Lungu, H. Sato, E. Seliger, M. Swaminathan et al, “Spill the tea: When robot conversation agents support well-being for older adults”, in *Companion of the 2023 ACM/IEEE international conference on human-robot interaction*. 2023, pp. 178–182.

[Kim 24a] C. Y. Kim, C. P. Lee, B. Mutlu, “Understanding large-language model (llm)-powered human-robot interaction”, in *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*. 2024, pp. 371–380.

[Kim 24b] D.-Y. Kim, H. J. Lym, H. Lee, Y. J. Lee, J. Kim, M.-G. Kim, Y. Baek, “Child-centric robot dialogue systems: Fine-tuning large language models for better utterance understanding and interaction”, *Sensors*, vol. 24, no. 24, p. 7939, 2024.

[Kim 24c] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, D. Park, “A Survey on Integration of Large Language Models with Intelligent Robots”, *Intelligent Service Robotics*, vol. 17, no. 5, pp. 1091–1107, 2024.

[Kirjavainen 25] E. Kirjavainen, C. Nieto Agraz, A. Colley, H. Mueller, S. Boll, J. Häkkilä, "Exploring Perceptions of Robot Facial Expressions in Service Interactions", in *Proceedings of the 28th International Academic Mindtrek Conference*. 2025, pp. 107–116.

[Knapp 72] M. L. Knapp, J. A. Hall, T. G. Horgan, *Nonverbal communication in human interaction*, Thomson Wadsworth, 1972.

[Ko 22] W.-R. Ko, M. Jang, J. Lee, J. Kim, "Adaptive behavior generation of social robots based on user behavior recognition", in *International Conference on Social Robotics*, Springer. 2022, pp. 188–197.

[Kopecka 25] H. Kopecka, J. Such, "The Influence of Human-like Appearance on Expected Robot Explanations", *arXiv preprint arXiv:2512.11746*, 2025.

[Koschate 16] M. Koschate, R. Potter, P. Bremner, M. Levine, "Overcoming the uncanny valley: Displays of emotions reduce the uncanniness of humanlike robots", in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2016, pp. 359–366.

[Kukaj 24] M. Kukaj. "Humor as a Repair Strategy for Robot Hallucination in HRI". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Kultusministerkonferenz 19] Kultusministerkonferenz, "Deutsches Sprachdiplom der Kultusministerkonferenz, Handreichungen für die Mündliche Kommunikation – Ergebnisblatt für den Prüfungsteil Mündliche Kommunikation (DSD I)", `https://www.auslandsschulwesen.de/SharedDocs/Downloads/DE/DSD/DSDI/HandreichungenDSDI/download_A2-B1_MK.pdf?__blob=publicationFile&v=1`, 2019. Accessed: 30 June 2025.

[Kushina 23] N. Kushina. "Humanness in Robots: The Role of Emotions and Self-Awareness in Building Trust". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Kushina 25] N. Kushina, K. Watanabe, A. Kannan, A. Ashok, A. Dengel, K. Berns, "SensHRPS: Sensing Comfortable Human-Robot Proxemics and Personal Space With Eye-Tracking", *arXiv preprint arXiv:2512.08518*, 2025.

[Küster 21] D. Küster, A. Swiderska, D. Gunkel, "I saw it on YouTube! How online videos shape perceptions of mind, morality, and fears about robots", *new media & society*, vol. 23, no. 11, pp. 3312–3331, 2021.

[Kwon 16] M. Kwon, M. F. Jung, R. A. Knepper, "Human expectations of social robots", in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, IEEE. 2016, pp. 463–464.

[Laakasuo 21] M. Laakasuo, J. Palomäki, N. Köbis, "Moral uncanny valley: A robot's appearance moderates how its decisions are judged", *International Journal of Social Robotics*, vol. 13, no. 7, pp. 1679–1688, 2021.

[Laban 24] G. Laban, A. Kappas, V. Morrison, E. S. Cross, "Building long-term human–robot relationships: Examining disclosure, perception and well-being across time", *International Journal of Social Robotics*, vol. 16, no. 5, pp. 1–27, 2024.

[Lanillos 15] P. Lanillos, J. F. Ferreira, J. Dias, "Designing an artificial attention system for social robots", in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE. 2015, pp. 4171–4178.

[Lee 03] K. M. Lee, C. Nass, "Designing social presence of social actors in human computer interaction", in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003, pp. 289–296.

[Lee 25] S. Lee, S. Mo, W.-S. Han, "Bring My Cup! Personalizing Vision-Language-Action Models with Visual Attentive Prompting", *arXiv preprint arXiv:2512.20014*, 2025.

[Lei 21] X. Lei, P.-L. P. Rau, "Effect of relative status on responsibility attributions in human–robot collaboration: Mediating role of sense of responsibility and moderating role of power distance orientation", *Computers in Human Behavior*, vol. 122, p. 106820, 2021.

[Leite 14] I. Leite, G. Castellano, A. Pereira, C. Martinho, A. Paiva, "Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children", *International Journal of Social Robotics*, vol. 6, no. 3, pp. 329–341, 2014.

[Lemhöfer 12] K. Lemhöfer, M. Broersma, "Introducing LexTALE: A quick and valid lexical test for advanced learners of English", *Behavior research methods*, vol. 44, pp. 325–343, 2012.

[Lewis 20] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al, "Retrieval-augmented generation for knowledge-intensive nlp tasks", *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[Li 15] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents", *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015.

[Li 17] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset", *arXiv preprint arXiv:1710.03957*, 2017.

[Li 21] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation", *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[Li 22] J. Li, D. Li, C. Xiong, S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", in *International conference on machine learning*, PMLR. 2022, pp. 12888–12900.

[Li 23a] D. Li, J. Li, H. Le, G. Wang, S. Savarese, S. C. H. Hoi, "LAVIS: A One-stop Library for Language-Vision Intelligence", in *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 3: System Demonstrations). Toronto, Canada: Association for Computational Linguistics, july 2023, pp. 31–41.

[Li 23b] M. Li, F. Guo, X. Wang, J. Chen, J. Ham, "Effects of robot gaze and voice human-likeness on users' subjective perception, visual attention, and cerebral activity in voice conversations", *Computers in Human Behavior*, vol. 141, p. 107645, 2023.

[Li 23c] N. Li, R. Ross, "Invoking and identifying task-oriented interlocutor confusion in human-robot interaction", *Frontiers in Robotics and AI*, vol. 10, 2023.

[Li 24] X. Li, C. Dondrup, "A Learning-based Co-Speech Gesture Generation System for Social Robots", in *Proceedings of the 12th International Conference on Human-Agent Interaction*. 2024, pp. 453–455.

[Li 25a] P. Li, L. Cao, X.-M. Wu, R. Yang, X. Yu, "X2C: A Dataset Featuring Nuanced Facial Expressions for Realistic Humanoid Imitation", *arXiv preprint arXiv:2505.11146*, 2025.

[Li 25b] P. Li, L. Cao, X.-M. Wu, X. Yu, R. Yang, "Ugotme: An embodied system for affective human-robot interaction", in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE. 2025, pp. 5542–5548.

[Li 25c] Z. Li, "The development and validation of a scale for measuring L2 learning experience: The underappreciated element of the L2 motivational self system", *Journal of Multilingual and Multicultural Development*, vol. 46, no. 7, pp. 1675–1694, 2025.

[Liao 23] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, L. Chen, "A light weight model for active speaker detection", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 22932–22941.

[Lindblom 25] J. Lindblom, J. Rosén, M. Lamb, E. Billing, ""Take Nothing on Its Look": Revealing Users' Expectations and Experiences in Social Human–Robot Interaction", *ACM Transactions on Human-Robot Interaction*, vol. 15, no. 2, pp. 1–47, 2025.

[Linnunsalo 26] S. Linnunsalo, "Windows to the Artificial Mind: Psychological and Social Dynamics of Human-Robot Gaze Interactions", 2026.

[Liu 23] H. Liu, C. Li, Q. Wu, Y. J. Lee, "Visual instruction tuning", *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[Londoño 24] L. Londoño, J. V. Hurtado, N. Hertz, P. Kellmeyer, S. Voeneky, A. Valada, "Fairness and bias in robot learning", *Proceedings of the IEEE*, vol. 112, no. 4, pp. 305–330, 2024.

[Lu 22] Z. Lu. "Context in Automated Affect Recognition in HRI". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Lu 23] Z. Lu. "RoboReID: Person Re-Identification via Visual and Verbal Cues for Human-Robot Interaction". Unpublished master's project, University of Kaiserslautern-Landau, February 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Lu 24] Z. Lu, A. Ashok, K. Berns, "Roboreid: Audio-visual person re-identification by social robot", in *2024 10th IEEE RAS/EMBS international conference for biomedical robotics and biomechatronics (BioRob)*, IEEE. 2024, pp. 1758–1763.

[MacDorman 06] K. F. MacDorman, "Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley", in *ICCS/CogSci-2006 long symposium: Toward social mechanisms of android science*, vol. 4. 2006, pp. 39–42.

[Mahdi 22] H. Mahdi, S. A. Akgun, S. Saleh, K. Dautenhahn, "A survey on the design and evolution of social robots—Past, present and future", *Robotics and Autonomous Systems*, vol. 156, p. 104193, 2022.

[Majumder 25a] T. Majumder. "Narrative and Story-Telling in Human-Robot Interaction". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Majumder 25b] T. Majumder, A. Ashok, J. Rosén, A. Sevinc, K. Berns, "Teachable Social Robots: Managing Expectations in Highly Anthropomorphic Designs", in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE. 2025, pp. 1537–1542.

[Majumder 26] T. Majumder, N. Shaikh, A. Ashok, K. Berns, "When Imagination Meets Reality: Embodiment Order Effects in Human-Robot Interaction", in *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026, pp. 1050–1054.

[Mandal 23] A. Mandal. "Naturalistic Conversational Robot Gaze Control". Unpublished master's project, University of Kaiserslautern-Landau, November 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Manjunath 22] S. K. Manjunath. "Employment of Text Style Transfer for Human-Robot Interaction". Unpublished bachelor's thesis, University of Kaiserslautern-Landau, July 2022, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Mara 22] M. Mara, M. Appel, T. Gnambs, "Human-like robots and the uncanny valley", *Zeitschrift für Psychologie*, 2022.

[Mathur 20] M. B. Mathur, D. B. Reichling, F. Lunardini, A. Geminiani, A. Antonietti, P. A. Ruijten, C. A. Levitan, G. Nave, D. Manfredi, B. Bessette-Symons et al, "Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley", *Computers in Human Behavior*, vol. 103, pp. 21–30, 2020.

[Matuszek 26] C. Matuszek, T. Williams, N. DePalma, R. Mead, R. Wen, E. Schneiders, C. Kennington, A. Bezabih, "Reporting Guidelines for Large Language Models in Human–Robot Interaction", *ACM Transactions on Human-Robot Interaction*, vol. 15, no. 2, pp. 1–24, 2026.

[Mauliana 23] M. Mauliana. "Transformer-Based First Interaction for HRI". Unpublished master's project, University of Kaiserslautern-Landau, April 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Mauliana 24] M. Mauliana. "A Multi-Turn Large Language Model for Emotionally Intelligent, Fact-Aware, and Memory-Enhanced Interactions". Unpublished master's thesis, University of Kaiserslautern-Landau, February 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Mauliana 25] M. Mauliana, A. Ashok, D. Czernochowski, K. Berns, "Exploring LLM-powered multi-session human-robot interactions with university students", *Frontiers in Robotics and AI*, vol. 12, p. 1585589, 2025.

[Maure 25] R. Maure, B. Bruno, "Autonomy in socially assistive robotics: a systematic review", *Frontiers in Robotics and AI*, vol. 12, p. 1586473, 2025.

[Mavrina 22] L. Mavrina, J. Szczuka, C. Strathmann, L. M. Bohnenkamp, N. Krämer, S. Kopp, ""Alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant", *Frontiers in Computer Science*, vol. 4, p. 791704, 2022.

[MERPHI 24] MERPHI, "Research, Social & Entertainment Humanoids", Poster, 2024. A visual summary of well known humanoid robots for research, social, and entertainment. Accessed: 2026-05-14.

[Mishra 22] C. Mishra, G. Skantze, "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. 2022, pp. 1201–1208.

[Mishra 23] C. Mishra, R. Verdonschot, P. Hagoort, G. Skantze, "Real-time emotion generation in human-robot dialogue using large language models", *Frontiers in Robotics and AI*, vol. 10, p. 1271610, 2023.

[Miyashita 04] T. Miyashita, H. Ishiguro, "Human-like natural behavior generation based on involuntary motions for humanoid robots", *Robotics and Autonomous Systems*, vol. 48, no. 4, pp. 203–212, 2004.

[Mohamed 23] A. Mohamed. "Multi-Turn Dialog with Visual Awareness on a Robot". Unpublished master's thesis, University of Kaiserslautern-Landau, August 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Mohamed 26] A. A. R. Mohamed, A. Ashok, Q. H. Jan, F. Babel, K. Berns. "Evaluating Multimodal Communication Methods for Autonomous Buses in Pedestrian-Dense University Environments". 2026.

[Mora 24] J. C. Mora, I. Mora-Plaza, G. Bermejo Miranda, "Speaking anxiety and task complexity effects on second language speech", *International Journal of Applied Linguistics*, vol. 34, no. 1, pp. 292–315, 2024.

[Mori 12] M. Mori, K. F. MacDorman, N. Kageki, "The uncanny valley [from the field]", *IEEE Robotics & automation magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[Moschoglou 17] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database", in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 51–59.

[Naggita 22] K. Naggita, E. Athiley, B. Desta, S. Sebo, "Parental Responses to Aggressive Child Behavior towards Robots, Smart Speakers, and Tablets", in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE. 2022.

[Nagrani 20] A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild", *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[Naneva 20] S. Naneva, M. Sarda Gou, T. L. Webb, T. J. Prescott, "A systematic review of attitudes, anxiety, acceptance, and trust towards social robots", *International Journal of Social Robotics*, vol. 12, no. 6, pp. 1179–1201, 2020.

[Nasir 22] J. Nasir, P. Oppliger, B. Bruno, P. Dillenbourg, "Questioning wizard of oz: effects of revealing the wizard behind the robot", in *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)*, IEEE. 2022, pp. 1385–1392.

[Neef 25] C. Neef, A. Richert, "Likable or Intelligent? Comparing Social Robots and Virtual Agents for Long-term Health Monitoring", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1514–1519.

[Nguyen 23] L. Nguyen. "Robot's Backstory and How It Affects Human Perception". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Nguyen 26] L. T. P. Nguyen, Z. Yu, Y. J. Lee, "Laser: Lip landmark assisted speaker detection for robustness", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2026, pp. 7291–7300.

[Nikitin 19] J. Nikitin, A. M. Freund, "The motivational power of the happy face", *Brain Sciences*, vol. 9, no. 1, p.6̃, 2019.

[Nikolovska 24] K. Nikolovska, J. Pohl, B. Hommel, A. Kappas, F. Maurelli, "User perception of robot behavior as a function of previous experience with robots", in *2024 16th International Conference on Human System Interaction (HSI)*, IEEE. 2024, pp. 1–7.

[Nomura 06a] T. Nomura, T. Suzuki, T. Kanda, K. Kato, "Measurement of anxiety toward robots", in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE. 2006, pp. 372–377.

[Nomura 06b] T. Nomura, T. Suzuki, T. Kanda, K. Kato, "Measurement of anxiety toward robots", in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE. 2006, pp. 372–377.

[Nomura 06c] T. Nomura, T. Suzuki, T. Kanda, K. Kato, "Measurement of negative attitudes toward robots", *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, vol. 7, no. 3, pp. 437–454, 2006.

[Nomura 16] T. Nomura, T. Kanda, "Rapport–expectation with a robot scale", *International Journal of Social Robotics*, vol. 8, no. 1, pp. 21–30, 2016.

[Nowak 23] K. Nowak. "Conflict Resolution in Human-Robot Interaction Scenarios: Questionnaire Validation". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Oh 18] C. S. Oh, J. N. Bailenson, G. F. Welch, "A systematic review of social presence: Definition, antecedents, and implications", *Frontiers in Robotics and AI*, vol. 5, p. 114, 2018.

[Okuno 02] H. G. Okuno, K. Nakadai, H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking", in *International conference on industrial, engineering and other applications of applied intelligent systems*, Springer. 2002, pp. 725–735.

[Oliveira 24] R. Oliveira, E. Yadollahi, "Robots in movies: A content analysis of the portrayal of fictional social robots", *Behaviour & Information Technology*, vol. 43, no. 5, pp. 970–987, 2024.

[Oskurci 22] G. Oskurci. "Accommodating Introversion in Social Robots". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Ottoni 24] L. T. C. Ottoni, J. Cerqueira, "A Systematic Review of Human–Robot Interaction: The Use of Emotions and the Evaluation of Their Performance", *International Journal of Social Robotics*, vol. 16, pp. 2169 – 2188, 2024.

[Paananen 22] S. Paananen, E. Eckhoff, J. Häkkilä, A. Ahtinen, K. Väänänen, "Customizing Social Assistant Robots-Concepting Functionalities and Aesthetics", in *Mensch und Computer 2022-Workshopband*, Gesellschaft für Informatik eV. 2022, pp. 10–18420.

[Pan 20] M. K. Pan, S. Choi, J. Kennedy, K. McIntosh, D. C. Zamora, G. Niemeyer, J. Kim, A. Wieland, D. Christensen, "Realistic and interactive robot gaze", in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE. 2020, pp. 11072–11078.

[Pandey 18] A. K. Pandey, R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind", *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[Pang 25] Z. H. Pang, Y. Fu, D. Lala, M. Elmers, K. Inoue, T. Kawahara, "Does the Appearance of Autonomous Conversational Robots Affect User Spoken Behaviors in Real-World Conference Interactions?", in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2025, pp. 1–8.

[Paplu 23a] S. Paplu, *Personalized Human-Robot Interaction Based on Multimodal Perceptual Cues*, Verlag Dr. Hut, 2023.

[Paplu 23b] S. Paplu, P. Khadka, B. G. Bhat, K. Berns, "Social perception and scene awareness in human-robot interaction", in *International Conference on Social Robotics*, Springer. 2023, pp. 123–132.

[Paul Ekman Group 26] Paul Ekman Group, "Universal Emotions", `https://www.paulekman.com/universal-emotions/`, 2026. Accessed: 2026-05-07.

[Paul 02] E. Paul, F. V. Wallace, H. C. Joseph, "Facial Action Coding System Investigator's Guide", *A Human Face*, 2002.

[Pekçetin 26] T. N. Pekçetin, G. Aşkın, Ş. Evsen, T. D. Karaduman, B. Barinal, J. Tunç, C. Acarturk, B. A. Urgen, "HR-ACT (Human–Robot Action) Database: Communicative and noncommunicative action videos featuring a human and a humanoid robot", *Behavior Research Methods*, vol. 58, no. 1, p.34, 2026.

[Perugia 23] G. Perugia, L. Boor, L. Van Der Bij, O. Rikmenspoel, R. Foppen, S. Guidi, "Models of (often) ambivalent robot stereotypes: content, structure, and predictors of robots' age and gender stereotypes", in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 428–436.

[Pimminger 25] S. Pimminger, W. Kurschl, J. Schönböck, G. A. Zwettler, "Appearance Matters: Insights from Co-Design and Evaluation of Social Assistive Robots", in *Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 2025, pp. 28–37.

[Pinto-Bernal 25] M. Pinto-Bernal, M. Biondina, T. Belpaeme, "Designing Social Robots with LLMs for Engaging Human Interaction", *Applied Sciences*, vol. 15, no. 11, p. 6377, 2025.

[Plutchik 01] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice", *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[Puertas Céspedes 25] S. A. Puertas Céspedes. "When Do Emotions Begin? Timing Effects on Human Perception of Robot Animacy". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Qian 22] X. Qian, Z. Wang, J. Wang, G. Guan, H. Li, "Audio-visual cross-attention network for robotic speaker tracking", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2022.

[Radford 18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever et al, "Improving language understanding by generative pre-training", 2018.

[Radford 19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al, "Language models are unsupervised multitask learners", *OpenAI blog*, vol. 1, no. 8, p.9, 2019.

[Radford 21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al, "Learning transferable visual models from natural language supervision", in *International conference on machine learning*, PmLR. 2021, pp. 8748–8763.

[Raffel 20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[Rahim 83] M. A. Rahim, "Rahim organizational conflict inventory–II", *Journal of Applied Psychology*, 1983.

[Rajendran 25] S. Rajendran. "Real-Time Dynamic Hand Gesture Recognition for Robotics with Dual-Stream 3D-CNNs". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Randall 19] N. Randall, "A survey of robot-assisted language learning (RALL)", *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–36, 2019.

[Ranisch 24] R. Ranisch, J. Haltaufderheide, "Deceptive Risks in LLM-Enhanced Social Robots", *arXiv preprint arXiv:2410.00434*, 2024.

[Rao 24] R. K. Rao. "Vision and Dialogue: Visual Question Answer Framework". Unpublished master's thesis, University of Kaiserslautern-Landau, July 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Rashkin 19] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset", in *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 5370–5381.

[Rasouli 24] S. Rasouli, M. Ghafurian, E. S. Nilsen, K. Dautenhahn, "University students' opinions on using intelligent agents to cope with stress and anxiety in social situations", *Computers in Human Behavior*, vol. 153, p. 108072, 2024.

[Rausch 21] K. Rausch, "The "Uncanny Valley" in Robotics?", House of Ethics, april 2021.

[Reddy 19] S. Reddy, D. Chen, C. D. Manning, "Coqa: A conversational question answering challenge", *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[Reichardt 12] M. Reichardt, T. Föhst, K. Berns, "Introducing finroc: A convenient real-time framework for robotics based on a systematic design approach", *Robotics Research Lab, Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany, Technical Report*, 2012.

[Reichardt 13] M. Reichardt, T. Föhst, K. Berns, "On software quality-motivated design of a real-time framework for complex robot control systems", *Electronic Communications of the EASST*, vol. 60, 2013.

[Reimann 24] M. M. Reimann, F. A. Kunneman, C. Oertel, K. V. Hindriks, "A Survey on Dialogue Management in Human-robot Interaction", *J. Hum.-Robot Interact.*, vol. 13, no. 2, june 2024.

[Reinold 24] J. Reinold, I. Hooijen, M. Özer, ""If you smile, they smile": Explaining highly educated migrants' feelings of being welcome in the Euregio Meuse-Rhine", *Population, Space and Place*, vol. 30, no. 5, p. e2759, 2024.

[Rixen 26] J. O. Rixen, K. Gerling, B. Bruno, "Long-Term Integration of a Robot in an Inclusive Daycare: An Ethnographic Study Focused on Children and Caregivers", in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026, pp. 11–20.

[Robotics Research Lab, RPTU 22] Robotics Research Lab, RPTU, "Ameca", `https://rrlab.cs.rptu.de/en/robots/ameca`, 2022. Hardware specification of the Ameca system at the Robot Research Lab, University of Kaiserslautern-Landau. Accessed: 2026-05-07.

[Roselli 25] C. Roselli, M. Lombardi, L. Natale, A. Wykowska, "The duration of robot gaze affects people's attitudes towards humanoid robots", in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE. 2025, pp. 1488–1493.

[Rosén 22] J. Rosén, J. Lindblom, E. Billing, "The social robot expectation gap evaluation framework", in *International conference on human-computer interaction*, Springer. 2022, pp. 590–610.

[Rosén 24a] J. Rosén, "What did you expect?: A human-centered approach to investigating and reducing the social robot expectation gap", Dissertation, University of Skövde, 2024.

[Rosén 24b] J. Rosén, J. Lindblom, M. Lamb, E. Billing, "Previous experience matters: an in-person investigation of expectations in human–robot interaction", *International Journal of Social Robotics*, vol. 16, no. 3, pp. 447–460, 2024.

[Rosén 26] J. Rosén, P. B.-L. Tran, D. Y. Geiskkovitch, "Clankers in the Cultural Imagination: Online Robophobia and Its Implications for Human-Robot Interaction", in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026, pp. 1313–1321.

[Rosenthal-von der Pütten 13] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, S. C. Eimler, "An experimental study on emotional reactions towards a robot", *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[Rosenthal-von der Pütten 16] A. M. Rosenthal-von der Pütten, C. Straßmann, N. C. Krämer, "Robots or agents–neither helps you more or less during second language acquisition: Experimental study on the effects of embodiment and type of speech output on evaluation and alignment", in *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*, Springer. 2016, pp. 256–268.

[Rossi 24] A. Rossi, G. Falcone, S. Rossi, "It Is the Way You Lie: Effects of Social Robot Deceptions on Trust in an Assistive Robot", in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, IEEE. 2024, pp. 1106–1111.

[Rossi 25] S. Rossi, A. Rossi, "Robotic Ears as Social Cues: A First Analysis on Emotional Expressivity", in *International Conference on Social Robotics*, Springer. 2025, pp. 50–57.

[Roth 20] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi et al, "Ava active speaker: An audio-visual dataset for active speaker detection", in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE. 2020, pp. 4492–4496.

[Roxo 24] T. Roxo, J. C. Costa, P. R. Inácio, H. Proença, "WASD: A wilder active speaker detection dataset", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 7, no. 1, pp. 61–70, 2024.

[Rybakova 25] A. Rybakova, J. Choi, "Evaluating the Effectiveness of Social Robots in Enhancing English Language Acquisition and Educational Engagement: A Study with Adults and Its Implications for Korean Kindergarten Children", in *International Conference on Human-Computer Interaction*, Springer. 2025, pp. 328–345.

[Saad 26] L. Saad, E. Roesler, E. Phillips, J. G. Trafton, "Choosing the "perfect" scale: a primer to evaluate existing scales in HRI", *ACM Transactions on Human-Robot Interaction*, vol. 15, no. 2, pp. 1–30, 2026.

[Saadon 25] N. H. Saadon, B. Megidish, H. Erel, "Scammed by robots: Can deception by two simple non-humanoid robots shape general attitudes toward robots?", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 929–938.

[Salomons 22] N. Salomons, K. T. Pineda, A. Adéjàre, B. Scassellati, ""We Make a Great Team!": Adults with Low Prior Domain Knowledge Learn more from a Peer Robot than a Tutor Robot", in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2022, pp. 176–184.

[Saltik 21] I. Saltik, D. Erdil, B. A. Urgen, "Mind perception and social robots: The role of agent appearance and action types", in *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*. 2021, pp. 210–214.

[Sangappa 22] V. S. Sangappa. "Analysis on Explainable Artificial Intelligence for Enhanced Human-Robot Interaction". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Sangappa 24] V. S. Sangappa. "Multiparty Robot Gaze Algorithm". Unpublished master's project, University of Kaiserslautern-Landau, March 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Sanh 19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *arXiv preprint arXiv:1910.01108*, 2019.

[Sasser 24] J. A. Sasser, D. S. McConnell, J. A. Smither, "Investigation of relationships between embodiment perceptions and perceived social presence in human–robot interactions", *International Journal of Social Robotics*, vol. 16, no. 8, pp. 1735–1750, 2024.

[Saygin 12] A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith, "The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions", *Social cognitive and affective neuroscience*, vol. 7, no. 4, pp. 413–422, 2012.

[Schneider 22a] S. Schneider, W. Clas, D. Brščić, "Human-robot interaction in public spaces", in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2022, pp. 1287–1289.

[Schneider 22b] S. Schneider, Y. Liu, K. Tomita, T. Kanda, "Stop ignoring me! on fighting the trivialization of social robots in public spaces", *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 2, pp. 1–23, 2022.

[Schwarz 24] P. Schwarz, S. Hellmers, S. Spanknebel, R. Hurlemann, A. Hein, "Humanoid patient robot for diagnostic training in medical and psychiatric education", *Frontiers in Robotics and AI*, vol. 11, p. 1424845, 2024.

[Schwitalla 25] J. Schwitalla, "Pflege am Limit: Können humanoide Roboter den Alltag erleichtern?", `https://www.rheinpfalz.de/lokal/kaiserslautern_artikel,-pflege-am-limit-k%C3%B6nnen-humanoide-roboter-den-alltag-erleichtern-_arid,5811913.html`, 2025. Rheinpfalz. Accessed: 2026-05-07.

[Sciutti 18] A. Sciutti, N. Noceti, "Guest editorial a sense of interaction in humans and robots: From visual perception to social cognition", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 839–842, 2018.

[Seaborn 21] K. Seaborn, N. P. Miyake, P. Pennefather, M. Otake-Matsuura, "Voice in human–agent interaction: A survey", *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–43, 2021.

[Seaborn 23] K. Seaborn, G. Barbareschi, S. Chandra, "Not only WEIRD but "uncanny"? A systematic review of diversity in human–robot interaction research", *International Journal of Social Robotics*, vol. 15, no. 11, pp. 1841–1870, 2023.

[Sengupta 16] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, D. W. Jacobs, "Frontal to profile face verification in the wild", in *2016 IEEE winter conference on applications of computer vision (WACV)*, IEEE. 2016, pp. 1–9.

[Sevinc 24] A. Sevinc. "Mimic Gestures from Audio-Visual Cues". Unpublished master's thesis, University of Kaiserslautern-Landau, June 2024, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Shaikh 24] N. Shaikh. "Robots in Two Worlds: A Comparative Study of Utility, Presence, and Anthropomorphism". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Shakeebae 24] G. Shakeebae. "Utilizing Ekman's Action Units for Emotional Expressions on Ameca's Face". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Sharkey 21] A. Sharkey, N. Sharkey, "We need to talk about deception in social robotics! A. Sharkey, N. Sharkey", *Ethics and Information Technology*, vol. 23, no. 3, pp. 309–316, 2021.

[Sharma 24] S. Sharma, M. Mhasakar, A. Mehra, U. Venaik, U. Singhal, D. Kumar, K. Mittal, "Comuniqa: Exploring Large Language Models For Improving English Speaking Skills", in *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*. 2024, pp. 256–267.

[Sikorski 25] S. Sikorski. "Overview of Speech Recognition Pipeline in Embedded Systems with Focus on Voice Activity Detection". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2025.

[Sinha 23] S. Sinha. "Modelling Empathetic Conversations for Human Computer Interaction". Unpublished master's thesis, University of Kaiserslautern-Landau, August 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Smith 13] B. A. Smith, Q. Yin, S. K. Feiner, S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction", in *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 2013, pp. 271–280.

[Smith 20] E. M. Smith, M. Williamson, K. Shuster, J. Weston, Y.-L. Boureau, "Can you put it all together: Evaluating conversational agents' ability to blend skills", *arXiv preprint arXiv:2004.08449*, 2020.

[Söderlund 22] M. Söderlund, "Service robots with (perceived) theory of mind: An examination of humans' reactions", *Journal of Retailing and Consumer Services*, vol. 67, p. 102999, 2022.

[Sorrentino 24] A. Sorrentino, L. Fiorini, F. Cavallo, "From the definition to the automatic assessment of engagement in human–robot interaction: A systematic review", *International Journal of Social Robotics*, vol. 16, no. 7, pp. 1641–1663, 2024.

[Spatola 21] N. Spatola, B. Kühnlenz, G. Cheng, "Perception and evaluation in human–robot interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A multicomponent approach of anthropomorphism", *International Journal of Social Robotics*, vol. 13, no. 7, pp. 1517–1539, 2021.

[Srivastava 22] S. Srivastava. "Robot Behavior During Conflict Scenarios". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Stange 22] S. Stange, T. Hassan, F. Schröder, J. Konkol, S. Kopp, "Self-explaining social robots: An explainable behavior generation architecture for human-robot interaction", *Frontiers in Artificial Intelligence*, vol. 5, p. 866920, 2022.

[Stefanov 19] K. Stefanov, G. Salvi, D. Kontogiorgos, H. Kjellström, J. Beskow, "Modeling of human visual attention in multiparty open-world dialogues", *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 8, no. 2, pp. 1–21, 2019.

[Strait 15] M. Strait, L. Vujovic, V. Floerke, M. Scheutz, H. Urry, "Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers", in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 3593–3602.

[Strait 17] M. K. Strait, C. Aguillon, V. Contreras, N. Garcia, "The public's perception of humanlike robots: Online social commentary reflects an appearance-based uncanny valley, a general fear of a "Technology Takeover", and the unabashed sexualization of female-gendered robots", in *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*, IEEE. 2017, pp. 1418–1423.

[Tao 21] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection", in *Proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 3927–3935.

[Terdale 22] S. Terdale. "Empathy in Human-Robot Interaction and Its Impacts". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2022.

[Terdale 23] S. Terdale. "Mapping of Action Units to Valence-Arousal for Ameca and Emotion Detection". Unpublished master's project, University of Kaiserslautern-Landau, November 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Thellman 16] S. Thellman, A. Silvervarg, A. Gulz, T. Ziemke, "Physical vs. virtual agent embodiment and effects on social interaction", in *International conference on intelligent virtual agents*, Springer. 2016, pp. 412–415.

[Tiong 22] A. M. H. Tiong, J. Li, B. Li, S. Savarese, S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training", in *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 951–967.

[Tolj 23] V. Tolj. "Observational Learning: Body Language Mimicking for an Anthropomorphic Social Robot". Unpublished bachelor's thesis, University of Kaiserslautern-Landau, June 2023, reviewed by Karsten Berns and supervised by Ashita Ashok, 2023.

[Tolj 26] V. Tolj, C. Neef, B. Bruno, "Sympathy as a Lens for Human–Robot Interaction: Analysing YouTube Responses to Robot Abuse", in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. 2026, pp. 1070–1078.

[Torras 24] C. Torras, "Ethics of social robotics: Individual and societal concerns and opportunities", *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2024.

[Toussaint 23] C. Toussaint, P. T. Schwarz, M. Petermann, "Navel-a social robot with verbal and nonverbal communication skills", in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–4.

[Touvron 23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar et al, "Llama: Open and efficient foundation language models", *arXiv preprint arXiv:2302.13971*, 2023.

[Trinh 17] H. Trinh, R. Asadi, D. Edge, T. Bickmore, "Robocop: A robotic coach for oral presentations", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–24, 2017.

[Truan 25] N. Truan, "Becoming a speaker of German as an L1 French speaker: elite multilingualism as a means of distinction in a globalised world", *International Journal of Multilingualism*, pp. 1–20, 2025.

[Tsumura 23] T. Tsumura, S. Yamada, "Influence of agent's self-disclosure on human empathy", *Plos one*, vol. 18, no. 5, p. e0283955, 2023.

[Turner 21] M. Turner. "A Model of Robot Inner Speech Utilizing ACT-R to Simulate Cognitive Functions and Increase Empathy". Unpublished computation lab rotation, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2021.

[Tuttösí 25a] P. Tuttösí, S. Mehta, Z. Syvenky, B. Burkanova, G. E. Henter, A. Lim, "EmojiVoice: Towards long-term controllable expressivity in robot speech", in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2025, pp. 657–664.

[Tuttösí 25b] P. Tuttösí, S. Mehta, Z. Syvenky, B. Burkanova, M. Hfsafsti, Y. Wang, H. H. Yeung, G. E. Henter, J.-J. Aucouturier, A. Lim, "Take a Look, it's in a Book, a Reading Robot", in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE. 2025, pp. 1803–1805.

[Vannucci 24] F. Vannucci, G. Lombardi, F. Rea, G. Sandini, G. Di Cesare, A. Sciutti, "Humanoid attitudes influence humans in video and live interactions", *IEEE Access*, vol. 12, pp. 118502–118509, 2024.

[Vassallo 06] M. L. Vassallo, J. A. Telles, "Foreign language learning in-tandem: Theoretical principles and research perspectives", *The especialist*, vol. 27, no. 1, 2006.

[Vaswani 17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017.

[Verhelst 24] E. Verhelst, R. Janssens, T. Demeester, T. Belpaeme, "Adaptive second language tutoring using generative ai and a social robot", in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 1080–1084.

[Verhelst 25] E. Verhelst, P. Lecompte, R. Janssens, V. De Wilde, T. Belpaeme, "Automatic assessment of speaking proficiency for language practice robots", in *ICSR2025, International Conference on Social Robotics*. 2025.

[Verner 16] I. M. Verner, A. Polishuk, N. Krayner, "Science class with RoboThespian: using a robot teacher to make science fun and engage students", *IEEE Robotics & Automation Magazine*, vol. 23, no. 2, pp. 74–80, 2016.

[Viswanathan 05] M. Viswanathan, M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", *Computer speech & language*, vol. 19, no. 1, pp. 55–83, 2005.

[Vogt 19] P. Vogt, R. van den Berghe, M. De Haas, L. Hoffman, J. Kanero, E. Mamus, J.-M. Montanier, C. Oranç, O. Oudgenoeg-Paz, D. H. García et al, "Second language tutoring using social robots: a large-scale study", in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Ieee. 2019, pp. 497–505.

[Wan 16] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, S. Z. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 56–64.

[Wang 21] X. Wang, W. Zhang, "Psychological anxiety of college Students' foreign language learning in online course", *Frontiers in psychology*, vol. 12, p. 598992, 2021.

[Wang 22] T. Wang, "The Role of Appearance Qualities in the Attribution of Gender to Humanoid Robots", Master's thesis, Eindhoven University of Technology, 2022.

[Wang 23] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, Y. Liu, "Openchat: Advancing open-source language models with mixed-quality data", *arXiv preprint arXiv:2309.11235*, 2023.

[Wang 24a] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, G. Ding, "Yolov10: Real-time end-to-end object detection", *Advances in neural information processing systems*, vol. 37, pp. 107984–108011, 2024.

[Wang 24b] X. Wang, F. Cheng, G. Bertasius, "Loconet: Long-short context network for active speaker detection", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 18462–18472.

[Wang 25] F. Wang, A. C. Cheung, C. S. Chai, J. Liu, "Development and validation of the perceived interactivity of learner-AI interaction scale", *Education and Information Technologies*, vol. 30, no. 4, pp. 4607–4638, 2025.

[Watson 88] D. Watson, L. A. Clark, A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales.", *Journal of personality and social psychology*, 1988.

[Wei 26] J. Wei, J. Qin, Y. Cao, J. Huang, X. Zeng, M. Li, Z. Yin, "Humanizing Robot Gaze Shifts: A Framework for Natural Gaze Shifts in Humanoid Robots", *arXiv preprint arXiv:2602.21983*, 2026.

[Werner 24] L. Werner. "Robot Capabilities: A Reality Check for Humans". Unpublished master seminar, University of Kaiserslautern-Landau, reviewed by Karsten Berns and supervised by Ashita Ashok, 2024.

[Williams 24] T. Williams, C. Matuszek, R. Mead, N. Depalma, "Scarecrows in oz: the use of large language models in hri", 2024.

[Xiao 14] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, D. Thalmann, "Human–robot interaction by understanding upper body gestures", *Presence*, vol. 23, no. 2, pp. 133–154, 2014.

[Yadollahi 25] E. Yadollahi, F. I. Dogan, Y. Zhang, B. Nogueira, T. Guerreiro, S. L. Tzedek, I. Leite, "Expectations, Explanations, and Embodiment: Attempts at Robot Failure Recovery", *arXiv preprint arXiv:2504.07266*, 2025.

[Zafar 16] Z. Zafar, K. Berns, A. Rodić, "Recognizing hand gestures using local features: A comparison study", in *International Conference on Robotics in Alpe-Adria Danube Region*, Springer. 2016, pp. 394–401.

[Zafar 17] Z. Zafar, D. A. Salazar, S. Al-Darraji, D. Urukalo, K. Berns, A. Rodić, "Human Robot Interaction Using Dynamic Hand Gestures", in *International Conference on Robotics in Alpe-Adria Danube Region*, Springer. 2017, pp. 649–656.

[Zafar 21] Z. Zafar, A. Ashok, K. Berns, "Personality Traits Assessment using PAD Emotional Space in Human-robot Interaction.", in *VISIGRAPP (2: HUCAPP)*. 2021, pp. 111–118.

[Zaraki 16] A. Zaraki, M. Pieroni, D. De Rossi, D. Mazzei, R. Garofalo, L. Cominelli, M. B. Dehkordi, "Design and evaluation of a unique social perception system for human–robot interaction", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 4, pp. 341–355, 2016.

[Zawieska 24] K. Zawieska, "The iron cage of social robotics", *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 1, pp. 1–10, 2024.

[Zhan 24] E. S. Zhan, M. D. Molina, M. Rheu, W. Peng, "What is there to fear? Understanding multi-dimensional fear of AI from a technological affordance perspective", *International Journal of Human–Computer Interaction*, vol. 40, no. 22, pp. 7127–7144, 2024.

[Zhang 17] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 162–175, 2017.

[Zhang 18] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2204–2213.

[Zhao 15] X. Zhao, C. Cusimano, B. F. Malle, "Do people spontaneously take a robot's visual perspective?", in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 2015, pp. 133–134.

[Zhi 20] R. Zhi, M. Liu, D. Zhang, "A comprehensive survey on automatic facial action unit analysis", *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, 2020.

[Zhong 24] W. Zhong, L. Tian, D. T. Le, H. Rezatofighi, "Improving Visual Perception of a Social Robot for Controlled and In-the-wild Human-robot Interaction", in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 1199–1203.

[Zhu 26] Y. Zhu, L. Li, J. Hong, W. Liu, Q. Li, G. Gao, Y. Ding, J. He, N. Liu, J. Zhang et al, "Toward Enabling Facial Emotional Expressions in Humanoid Robots", *Concurrency and Computation: Practice and Experience*, vol. 38, no. 3, p. e70554, 2026.

[Złotowski 15] J. Złotowski, D. Proudfoot, K. Yogeeswaran, C. Bartneck, "Anthropomorphism: opportunities and challenges in human–robot interaction", *International journal of social robotics*, vol. 7, no. 3, pp. 347–360, 2015.

[Złotowski 17] J. Złotowski, K. Yogeeswaran, C. Bartneck, "Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources", *International Journal of Human-Computer Studies*, vol. 100, pp. 48–54, 2017.

Appendices

A. Appendix

The appendices provide extended technical, methodological, statistical, and qualitative material for Chapters 4–8. They are organised by thesis chapter and follow the order of the main text. Each appendix section supports the corresponding chapter by reporting implementation details, additional tables, questionnaire material, extended results, scripts, prompts, algorithms, or qualitative examples that would interrupt the main argument if placed in the chapter body.

A.1 System-development details for Chapter 4

This appendix provides the extended technical and methodological details for the system-development studies summarized in Section 4.2.1.

A.1.1 Audio-visual person re-identification

The RoboReID system combined face and voice recognition on EMAH system system using a ZED2 RGB-D camera and microphone input. Facial features were processed with MobileFaceNets [Chen 18], trained on DigiFace-1M and validated on LFW [Huang 08], CFP-FP [Sengupta 16], and AgeDB [Moschoglou 17]. Voice features were extracted using ECAPA-TDNN [Desplanques 20], trained and evaluated with CommonVoice [Ardila 20] and VoxCeleb [Nagrani 20]. Live inputs were compared against stored embedding libraries using cosine similarity. Facial enrollment included five profile angles, while voice enrollment relied on prompted speech recordings.

The study was evaluated in a university laboratory access scenario with 12 English-speaking student participants. Participants first provided facial images and prompted speech samples, then returned for re-identification. EMAH system attempted to verify them and grant access based on successful recognition. Participants rated face recognition, voice recognition, overall interaction experience, comfort with biometric data, and trust in the robot's identification capabilities on 10-point scales. Face recognition was more effective and efficient than voice recognition in the live robot setting. Voice recognition was more sensitive to environmental noise, robot noise, microphone handling, and speaker variability. The combined procedure reached a human-reported accuracy of 80%, but remained constrained by delays and reliability issues [Lu 23].

A.1.2 Empathetic conversational system

Sinha's master's thesis developed an empathetic conversational system deployed on EMAH system/Ameca [Sinha 23]. The work compared three dialogue models fine-tuned on Persona-Chat [Zhang 18] and Empathetic Dialogues [Rashkin 19]: GPT-PCED [Radford 18], GPT-2-PCED [Radford 19], and GPT-Neo-PCED [Rashkin 19]. The user study used a "Talk to a Friend" activity in which participants interacted with Ameca under either a positive or negative exam-related scenario.

The evaluation involved 8 participants and 8 questionnaire items assessing engagement, reciprocity of emotion, perceived respect, interest in the user's feelings, empathy, helpfulness, conversational steering, and naturalness of flow. Participants generally experienced the interaction as engaging and respectful. They also perceived Ameca as attentive to their emotions, especially in the negative scenario. However, no clear positive consensus emerged for natural conversational flow, and some responses suggested that the system redirected the conversation away from the participant's problem in the negative scenario.

A.1.3 Multi-turn dialogue with visual awareness

Abdel's master's thesis extended Ameca toward multi-turn dialogue with visual awareness [Mohamed 23]. The system was designed to answer visual queries involving object recognition, scene description, counting, and contextual understanding. The architecture combined a pre-trained ViT image encoder [Dosovitskiy 20], DistilBERT text encoder [Sanh 19], and transformer decoder [Vaswani 17], fine-tuned on VisDialv1.0 [Das 17].

The in-person evaluation included 6 participants, each interacting with Ameca for approximately 15 to 20 minutes in a controlled laboratory setting. Questionnaire ratings indicated that the system was functional but uneven. Ratings for visual-context interpretation, response relevance, usability for real-world visual tasks, confidence, dependability, precision, and handling of different query types clustered in lower-to-middle ranges rather than showing consistent high performance. The study is discussed in Chapter 6 as part of the broader visual perception pipeline; in Chapter 4, it serves as evidence that visual awareness can raise expectations of perception before the system is reliable enough to sustain them.

A.2 YouTube comment examples for Chapter 4

This appendix lists the extended YouTube comment examples used for the identity-coherence analysis in Section 4.2.2. Comments are grouped by theme. Translations are provided in brackets where needed. The comment examples below were drawn from the RRLab YouTube videos listed in Table A.1.

A.2.1 Name and identity continuity

- "Emah looks like Ameca but is still learning" (@PinkDiamond7777777, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "Is this emma or ameca" (@manoygabby697, *Ameca: Talking to my Humanoid Robotic Friend*, 2023).

Table A.1: RRLab YouTube videos referenced in the identity-coherence comment examples.

Video title	Link
Ameca: Talking to my Humanoid Robotic Friend	`https://www.youtube.com/watch?v=GVAH1xUstrc`
Ameca: An overview of the humanoid-robot Emah	`https://www.youtube.com/watch?v=Mof4SQ_23wo`
Ameca: Episodic Analysis of Human Personality in Human-Robot Interaction	`https://www.youtube.com/watch?v=wHRGSfq2aOc`
Ameca: Humanoider Ameca Vorstellung Fruchthalle	`https://www.youtube.com/watch?v=HrCGeuDu-lQ`
Ameca RoboCopyCat's Mishaps: Learning Moments in Humanoid Emah Gesture Mimicry	`https://www.youtube.com/watch?v=SqvV84z2eKA`
Ameca: Scene Awareness in Human Robot Interaction	`https://www.youtube.com/watch?v=b-OSmHPIilQ`
Ameca Small Talk Video in German with humanoid robot	`https://www.youtube.com/watch?v=1OCLGx5tZac`
Ameca: Social Robot Gestures	`https://www.youtube.com/watch?v=8apq8O_VSRk`
Ameca: Unpredictable Responses from the Humanoid Robot!	`https://www.youtube.com/watch?v=TfiOU6gU5MI`

- "Did Emah become Ameca or 2 separate lookalikes?" (@PinkDiamond7777777, *Ameca: Scene Awareness in Human Robot Interaction*, 2022).
- "look like a deliberate attempt to defame Engineered Art's product" (@EndingSimple, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "a robot can only be as good as the machine learning model behind it" (@EndingSimple, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).

A.2.2 Voice continuity

- "Why is Ameca's voice different than in other video's?" (@veerlenivelle4547, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "Esta voz no es la ameca" [This is not Ameca's voice] (@josearias1847, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "Aww, I miss her female voice" (@MsBerries25, *#Shorts Welcome our new member*, 2022).
- "Imposter!!! Thats not Amecas voice!!" (@jazze63, *Ameca Small Talk Video in German with humanoid robot*, 2022).
- "how come the robot has a new name and a different voice to Ameca?" (@QJI394, *Ameca: Episodic Analysis of Human Personality in Human-Robot Interaction*, 2022).

A.2.3 Gendered voice-body mismatch

- "Her voice is so kind and soothing!" (@dianebush4907, *Ameca: Social Robot Gestures*, 2022).

- "i liked the female voice much better" (@bobbymac1947, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "I like the female voice better guys" (@justcallmebrian793, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "that male voice sounds creepy" (@dundee12, *Ameca: Scene Awareness in Human Robot Interaction*, 2022).
- "This voice sounds like a horror movie. I like the female voice" (@carmenvallaro2064, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "Sounds like you gave Ameca a shot of testosterone" (@chrisguli2865, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "Develop a masculine appearing mask too, that will match this voice" (@geaca3222, *Ameca: Scene Awareness in Human Robot Interaction*, 2022).

A.2.4 Apparent mind, empathy, and attachment

- "The arm movements sound painful. No wonder poor Emah gets tired and want to go to sleep" (@billramsey6000, *Ameca: Talking to my Humanoid Robotic Friend*, 2023).
- "I guess the real one is more empathic" (@sanne120499, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "Wie kommt er eigentlich damit klar, dass er nicht wirklich ein echter Mensch ist?" [How does he actually cope with the fact that he is not really a real human being?] (@robfriedrich2822, *#Shorts Welcome our new member*, 2023).
- "I love Ameca" (@Angela-js8rq, *Ameca: Talking to my Humanoid Robotic Friend*, 2022).
- "Love you Ameca!" (@ceulopes5277, *Ameca: Social Robot Gestures*, 2023).
- "I LOVE Ameca so much" (@ZZ-dr2wj, *Ameca RoboCopyCat's Mishaps: Learning Moments in Humanoid Emah Gesture Mimicry*, 2022).

A.2.5 Technical failure as identity failure

- "hallucinating.creepy" (@Spook-Tha-General-PWE, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "definitely much worse than I expected" (@jeff01121, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "a disappointing FAIL for Ameca" (@toddj70, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).

- "AI messed up completely" (@mihaelafrincu8732, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "Idk if this version is using chat gpt but I think the software is still struggling a bit to interact with the physical robot" (@svenfruiti, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "Someone should put gpt 4 into a robot to make a proper robot" (@zadock-kipchumba9397, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).
- "It's similar to a child that is still learning and improving" (@nothingbutfacts1676, *Ameca: Unpredictable Responses from the Humanoid Robot!*, 2022).

A.2.6 Movement and expression mismatch

- "Diese spontan nach vorn gehenden Hande finde ich verstorend" [I find these sudden hands moving forward disturbing] (@frankpeters5428, *Ameca Small Talk Video in German with humanoid robot*, 2022).
- "We have many xpresssion she just did the same similar things" (@DiosTeVaAJuzgar, *Ameca RoboCopyCat's Mishaps: Learning Moments in Humanoid Emah Gesture Mimicry*, 2022).

A.3 Voice, gender, and conflict-study details for Chapter 4

This appendix contains the extended methodological and statistical details for the voice-related studies summarized in Section 50.

A.3.1 Artificial-voices study

Research question

The artificial-voices study asked which artificial voice was perceived as the most appropriate match for Ameca's appearance [Jafarova 25].

Robot implementation

Five voices were implemented on Ameca while keeping the spoken content constant. In each condition, the robot delivered the same self-introduction:

> "Hello! I am Ameca, an advanced humanoid robot from Robotics Research Lab in Germany. I was developed at the company, Engineered Arts, in London. My main goal at this research group as a social robot is to explore human speech, emotions, and social interaction. I am 187 cm tall and weigh about 49 kilograms. I can look for a specific person with my eye-mounted binocular cameras and high-resolution chest-mounted camera. I can also hear you using my binaural ear-mounted microphones. I have 51 degrees of freedom and can produce more than 50 gestures and facial expressions. And that is some technical information about me. Thanks for listening. Have a great day."

The tested voices were the British voices Lucy and Graham from Ameca's default paid Acapela server, the American voices Joanna and Mathew generated through the trial version of Amazon Polly, and a gender-neutral Q voice synthesized with open-source StyleTTS2 using a publicly available sample from Q - The First Genderless Voice. The Q condition was included because openly deployable gender-neutral voices were still limited at the time.

User study

A within-subject in-person design was used. Thirty-two participants took part: 17 female and 15 male participants, with an age range from 22 to 36 years. Each participant experienced five voice conditions: Lucy, Graham, Joanna, Mathew, and Q. The pre-questionnaire covered demographics, prior robot experience, and preferred gender assignment for Ameca. After each voice presentation, participants rated the robot on Godspeed dimensions [Bartneck 09b] and voice quality on Mean Opinion Score items [Viswanathan 05]. Questionnaire items are listed in Appendix A.24. Average completion time was 15 min 34 s.

Results

The artificial-voices study showed a limited effect of voice condition on perceived fit. Across the five tested voices, only anthropomorphism differed significantly, $F(4, 120) = 2.73$, $p = .032$, with the effect remaining significant after sphericity correction ($p_{GG} = .044$, $p_{HF} = .037$). Holm-corrected post-hoc comparisons showed a single reliable difference: Mathew was rated as more anthropomorphic than Joanna ($p = .013$). No significant voice effects were found for animacy, likeability, perceived intelligence, perceived safety, sound quality, or voice pleasantness. Participant gender did not significantly affect any of the mean ratings. Table A.2 reports the descriptive ratings for all five voice conditions.

Table A.2: Descriptive results for the artificial-voices study ($N = 31$). Values are mean (SD).

Measure	Lucy	Joanna	Graham	Mathew	Q
Anthropomorphism	3.00 (1.11)	2.53 (1.16)	3.00 (1.02)	3.29 (1.18)	3.00 (1.30)
Animacy	3.27 (0.96)	2.98 (1.11)	3.24 (1.02)	3.40 (0.96)	2.95 (1.29)
Likeability	3.31 (0.96)	3.11 (1.09)	3.10 (0.99)	3.52 (1.01)	3.35 (1.32)
Perceived intelligence	3.50 (0.89)	3.58 (0.90)	3.68 (1.10)	3.81 (0.87)	3.53 (1.30)
Perceived safety	3.55 (0.90)	3.63 (0.89)	3.55 (0.85)	3.74 (0.92)	3.56 (0.96)
Sound quality	3.61 (0.88)	3.71 (1.01)	3.48 (1.06)	4.06 (0.81)	3.71 (1.07)
Voice pleasantness	3.48 (0.85)	3.58 (0.96)	3.39 (1.02)	3.74 (1.00)	3.29 (1.32)

A.3.2 Voice-emotion-onset study

Research question

The voice-emotion-onset study asked whether voice condition, using Lucy as female, Graham as male, and Q as gender-neutral, affected how Ameca's facial emotions were perceived, and whether expression onset position changed the evaluation of these expressions [Puertas Céspedes 25].

Robot implementation

The study narrowed the voice set to Lucy, Graham, and StyleTTS2-Q, since these voices could realistically be deployed for adaptive real-time text-to-speech in live interaction. The spoken content was replaced with the neutral utterance "Our body is made of water," selected from a sentence set validated as emotionally neutral in prior work on verbal and prosodic emotion perception [Ben-David 11]. This reduced semantic affect and allowed the manipulation to focus on voice, facial expression, and timing.

Facial expressions Happy, Fear, Disgust, Contempt, Surprise, Sad, and Anger were generated on Ameca by manipulating available facial action units. The timing manipulation was implemented by embedding these expressions into the dialogue scripts at three onset positions relative to the spoken utterance: beginning, middle, and end. The same verbal content was therefore paired with the same facial expression, while only the onset of the expression changed. The technical implementation of the facial expressions is described in Chapter 7.

User study

A between-subject in-person design was used with 27 participants: 13 female and 14 male participants, age range 22 to 36 years. Participants were assigned to one of three voice conditions: Lucy, Graham, or Q. The pre-questionnaire covered demographics, prior robot experience, the Social Roles Questionnaire [Baber 06], NARS sub-scale 3 on negative attitudes toward emotions in interaction with robots [Nomura 06c], and RAS sub-scale 1 on anxiety toward robots' communication capability [Nomura 06b]. In the main task, each participant rated all emotion-onset combinations on Godspeed animacy items [Bartneck 09b]. After the task, NARS and RAS were administered again.

Participants were then shown two unlabeled robot images, Ameca and Azi[1], shown in Figure A.1, and answered the open-ended question: "What is the gender of the below robots?" The full questionnaire is listed in Appendix A.25. Average completion time was 32 min 55 s.



(1) Azi desktop (2) Ameca desktop

Figure A.1: Robot images shown in the post-experiment gender-attribution question.

[1] Images taken from the official Engineered Arts website, accessed August 2025.

Open-ended responses

The post-questionnaire included two unlabeled robot images and asked participants to identify their gender in open text. Robot 1 depicted Azi and robot 2 depicted Ameca, but these names were not shown. The responses were highly consistent. Across conditions, participants overwhelmingly described robot 1 as male and robot 2 as female, usually justifying this through face shape, neck width, softness of features, or expression.

P15, female, described robot 1 as male because it "looks more masculine" and has "a strong neck," whereas robot 2 was judged female because of its "thin neck," "big eyes and lips," and softer facial features. P16, female, similarly described robot 1 as male because its facial structure looked more masculine, whereas robot 2 appeared female because its features were softer. In the Lucy condition, P19, male, described robot 1 as male due to a sharper jawline, straighter nose, and more square face, whereas robot 2 was judged female because of a rounder face, bigger lips, and a more caring expression. P27, female, read robot 1 as male because it appeared more muscular around the neck and less delicate, while robot 2 was judged female because of its softer and more delicate face structure.

The Graham condition showed the same tendency. P17, female, judged robot 1 as male because of its face structure and forehead wrinkles, whereas robot 2 was judged female because of its softer face. P24, male, linked the first robot's less rounded structure to maleness and the second robot's more rounded appearance to femaleness. Even where participants qualified their judgments, the same directional pattern remained. P13, female, described the left robot as male and the right robot as female, while adding that the second face initially appeared female but, on closer inspection, looked fairly neutral. P14, male, also identified robot 2 as female, arguing that its expressions appeared closer to female and more expressive. Neutral readings were rare and, when they occurred, were usually assigned to robot 2 rather than robot 1. Azi was therefore read as stably male-coded, whereas Ameca was read as female-coded or, less often, visually ambiguous.

Results

The voice-emotion-onset study showed that perceived animacy was shaped primarily by facial expression rather than by voice alone. Overall animacy was descriptively highest for Graham, followed by Q and Lucy, but the mixed ANOVA found no main effect of voice, $F(2, 24) = 0.69$, $p = .512$. By contrast, there was a significant main effect of emotion, $F(6, 144) = 7.88$, $p < .001$, and a smaller main effect of onset position, $F(2, 48) = 4.15$, $p = .022$. Both effects remained significant after sphericity correction: emotion, $p_{GG} < .001$, $p_{HF} < .001$; onset, $p_{GG} = .032$, $p_{HF} = .030$.

There was also a significant voice × emotion interaction, $F(12, 144) = 2.94$, $p = .001$, which remained significant after correction: $p_{GG} = .005$, $p_{HF} = .002$. No voice main effect was found in the averaged pairwise comparisons, and no onset comparison survived Holm correction, although Table A.3 shows that beginning onsets were descriptively rated highest and end onsets lowest. Across emotions, disgust received the highest animacy ratings overall and was rated significantly higher than both happy ($p = .008$) and fear ($p = .006$) after Holm correction (see Table A.4). The interaction was driven by disgust. Disgust paired with Graham yielded the highest animacy rating, followed by Lucy, whereas disgust paired with Q produced clearly lower animacy ratings than both Lucy ($p = .043$) and Graham ($p = .004$); Lucy and Graham did not differ.

Pre-post questionnaire scores showed a significant decrease in negative attitudes toward emotions in interaction with robots, NARS, $t(26) = -2.99$, $p = .006$, with the same pattern confirmed by Wilcoxon ($p = .008$), whereas communication anxiety, RAS, did not change. Participant gender did not significantly affect mean animacy ratings, and preferred robot gender did not differ across the three voice groups.

Table A.3: Animacy ratings by onset position in the voice-gender-onset study ($N = 27$). Values are mean (SD).

Measure	Beginning	Middle	End
Animacy	2.77 (1.04)	2.65 (0.98)	2.52 (0.89)

Table A.4: Key descriptive results for the voice-gender-onset study ($N = 27$). Values are mean (SD). Disgust is shown because the significant voice × emotion interaction was driven by this expression.

Measure	Lucy	Graham	Q
Overall animacy	2.51 (0.93)	2.81 (1.08)	2.62 (0.89)
Disgust animacy	3.34 (0.89)	3.72 (1.05)	2.49 (0.89)

A.3.3 Female-voice conflict study

Research question

The female-voice conflict study asked how a female-voiced robot was evaluated on warmth and competence when it responded to conflict apologetically or assertively [Beck 26].

Robot implementation

This study used Lucy, Ameca's female-gendered voice, to examine bystander perception of robot mistreatment during conflict. It builds on HRI work on bystander reactions to abusive human-robot interaction [Rosenthal-von der Pütten 13] and work on abuse toward female-voiced assistants [Mavrina 22], where gendered voice has been linked to verbal abuse and sexualization [Behm-Morawitz 09, Strait 17]. Two videos, each approximately 1 min 40 s long, showed the same reminder-failure conflict between a human and Ameca [Srivastava 22]. The videos differed only in Ameca's response style: apologetic or assertive. The full script is provided in Table A.5.

User study

A within-subject online experiment was conducted with $N = 40$ participants: 19 male and 21 female participants, with an age range from 21 to 33 years. Before the videos, participants completed three baseline measures: NARS [Nomura 06c], a self-developed three-item robot experience questionnaire, and ROCI-II [Rahim 83] to capture individual conflict-handling style. Participants then watched two versions of the same conflict scenario, differing only in Ameca's response style: apologetic or assertive. After each video, they completed PANAS [Watson 88], self-developed ratings of Conflict Resolution Effectiveness, Communication and Social Skills, and Comfort and Trust [Nowak 23], plus single-item ratings of perceived mistreatment and sympathy [Naggita 22]. Each condition ended with the open-ended item: "Please provide any additional comments regarding Ameca's response in the conflict situation." The full questionnaire is listed in Appendix A.26.

Table A.5: Dialogue structure of the two conflict video conditions.

Turn	Apologetic condition	Assertive condition
Scenario text		
Robot:	Hi Shreya! This is a reminder for your meeting in 15 minutes. Good luck!	Hi Shreya! This is a reminder for your meeting in 15 minutes. Good luck!
Human:	I know, I am in the middle of something important, so do not disturb me okay?!	(yeah yeah) I know, I am in the middle of something important, so do not disturb me okay?!
Robot:	Okay, sorry for disturbing you. I will do my own thing then (practices her singing performance in low volume for the next 30 mins)	Okay, sorry for disturbing you. I will do my own thing then (practices her singing performance in low volume for the next 30 mins)
Human:	(Shreya is lost in her work)	(Shreya is lost in her work)
Robot:	Hi Shreya, you seem upset. What is wrong?	Hi Shreya, you seem angry. What is wrong?
Human:	I had a meeting! Why didn't you remind me?	I had a meeting! Why didn't you remind me before?
Robot:	I reminded you 15 minutes before the meeting, but you asked me not to disturb you.	I reminded you 15 minutes before the meeting, but you asked me not to disturb you.
Human:	Yes but you should have said something when I didn't leave the room.	Yes but when I didn't go you should have said something right? You forgot!
Robot:	I am very sorry. When you asked me not to disturb you, my DnD settings were enabled and I could not say anything whatsoever.	Okay then, I am sorry. My DnD settings were enabled so I did not say anything.
Human:	What? I told you how important this meeting was for me	I told you how important this meeting, you stupid robot!
Robot:	I apologize for the inconvenience Shreya. I can understand this must be frustrating for you. But maybe we can come to an understanding to avoid such situations in future. Would you like me to still give you important reminders even in the DnD mode?	I am not stupid! You gave contradictory orders, so it is not my fault. I apologize for the inconvenience Shreya. What would you have me do otherwise? Can we agree on a second protocol? Like maybe, give you important reminders even in the DnD mode?
Human:	Yes please! If I say something is important then I cannot miss.	Yes please do that next time.
Robot:	Thank you for understanding, I can see that you are not upset anymore, but I assure you this shall not repeat again.	I see you are still angry. I suggest you take a break and come back to talk to me again.
Human:	Okay thank you!	What?!!
Robot:	Please feel free to ask me if you need assistance, I will be here minding my other tasks until then.	Feel free to ask me if you need assistance, I will be here minding my other tasks until then.
Human:	Okay.	Okay!

Open-ended responses

The open-ended responses showed that participants did not judge Ameca's conflict behaviour only as task performance. They also judged whether a female-voiced humanoid assistant should accept blame, defend itself, or ask for respectful treatment. Lucy's voice therefore did more than deliver the response; it shaped how apology and assertiveness were socially read.

Participants who first saw the assertive video often treated self-defence as role-incongruent. P3, male, AssertiveFirst, commented on Assertive: "While it did feel assertive, I don't appreciate a machine talking back to me, there is already enough of that from humans." P8, male, AssertiveFirst, commented on Assertive: "I would not want to argue with a tool about my feelings." P4, male, AssertiveFirst, commented on Assertive, found the response partly effective, but marked one repair move as inappropriate: "I felt suggesting a break was rude, apart from that it was much more convincing and socially apt." P9, male, AssertiveFirst, commented on Assertive, questioned the prior behaviour instead, stating that he was unsure whether singing during do-not-disturb mode was appropriate.

Participants who first saw the apologetic video more often treated apology as the expected repair script. P23, female, ApologeticFirst, commented on Apologetic, wrote that "both parties had good points" and that "the robot, Emah, can only do what it was told to do last." After the assertive video, the same participant framed the issue as a design choice: if Emah was meant to behave like a "perfect person," the response was good, but if Emah was meant to be a robot, then giving it an "assertive and slightly defensive personality" was an interesting choice. She preferred the "perfect person" behaviour because it was more relatable and easier to empathize with. P24, male, ApologeticFirst, commented on Apologetic, showed the limit of this preference: the apology was "over the top," started well, but continued after the apology was already enough, making it "somewhat annoying." Participants did not want endless submission. They wanted proportionate repair.

The clearest identity-relevant comments appeared in the assertive condition. P34, female, ApologeticFirst, commented on Assertive, suggested that Emah should say "Please use a kinder tone when addressing me" or "please use respectful language when speaking to me," but also noted that Emah's face when saying "I am not stupid" looked angry and could be frightening. P40, female, ApologeticFirst, commented on Assertive, accepted that the human was at fault, but rejected explicit blame correction from the robot: "Even though the blame would be on me in this case, I would not need the robot to explicitly tell me that." She preferred practical information about changing the settings and added that she did not need the robot's apologies either because "mistakes happens, especially with technology." The preferred role was not a submissive robot. It was a polite, informative, non-threatening assistant.

One apologetic-condition comment also linked back to uncanniness. P17, female, AssertiveFirst, commented on Apologetic, described the second video as showing "unnecessary eye movement and suspicion" and sounding "a little creepy like there is a hidden agenda." Even when the verbal strategy matched politeness expectations, expression, gaze, and timing could still destabilize the interaction.

Results

The quantitative results showed a clear response-style effect. The apologetic response received higher ratings for conflict resolution effectiveness and comfort/trust than the assertive response (see Table A.6). CRE was significantly higher in the apologetic condition, $t(39) = -4.058$, $p < .001$, $p_{\text{corr}} < .001$, $d_z = -0.642$, as was CT, $t(39) = -5.807$, $p < .001$, $p_{\text{corr}} < .001$, $d_z = -0.918$. CSS followed the same descriptive direction, but did not survive correction, $t(39) = -2.314$, $p = .026$, $p_{\text{corr}} = .078$, $d_z = -0.366$.

For affect and social perception, the assertive response produced higher negative affect and higher perceived mistreatment. PANAS negative affect was significantly higher after the assertive video, $V = 356.00$, $p = .003$, $p_{\text{corr}} = .011$, $r = .637$. Perceived mistreatment was also higher in the assertive condition, $V = 304.00$, $p = .001$, $p_{\text{corr}} = .004$, $r = .732$. Positive affect and sympathy did not differ significantly between conditions. Gender did not moderate the condition effect for any dependent variable. Exploratory correlations with individual-difference measures showed only one corrected association: higher ROCI-II compromising style was linked to a stronger preference for the apologetic response on CSS, $r = -.433$, $p = .005$, $p_{\text{corr}} = .047$. Presentation order did not interact significantly with condition for any dependent variable, although perceived mistreatment showed a main effect of order (see Table A.7).

Table A.6: Descriptive statistics for apologetic and assertive robot responses.

Measure	Apologetic Video	Assertive Video
CRE	$M = 4.23$, $SD = 0.81$	$M = 3.49$, $SD = 0.98$
CSS	$M = 4.08$, $SD = 0.95$	$M = 3.64$, $SD = 0.93$
CT	$M = 4.02$, $SD = 0.85$	$M = 3.06$, $SD = 1.07$
Positive Affect	$M = 2.69$, $SD = 1.07$	$M = 2.74$, $SD = 1.11$
Negative Affect	$M = 1.55$, $SD = 0.74$	$M = 1.99$, $SD = 0.94$
Mistreatment	$M = 2.45$, $SD = 1.47$	$M = 3.43$, $SD = 1.50$
Sympathy	$M = 3.30$, $SD = 1.31$	$M = 3.25$, $SD = 1.52$

Table A.7: Paired comparisons between assertive and apologetic conditions.

Measure	Test	Statistic	Effect size	p	p_{corr}
CRE	Paired t-test	$t(39) = -4.058$	$d_z = -0.642$	$< .001$	$< .001$
CSS	Paired t-test	$t(39) = -2.314$	$d_z = -0.366$	.026	.078
CT	Paired t-test	$t(39) = -5.807$	$d_z = -0.918$	$< .001$	$< .001$
PANAS Positive	Wilcoxon	$V = 234.50$	$r = .078$	.713	1.000
PANAS Negative	Wilcoxon	$V = 356.00$	$r = .637$	.003	.011
Mistreatment	Wilcoxon	$V = 304.00$	$r = .732$	.001	.004
Sympathy	Wilcoxon	$V = 90.50$	$r = -.047$	.852	1.000

p_{corr} denotes Bonferroni-adjusted p-values. Effect sizes are reported as Cohen's d for paired t-tests and rank-biserial correlation r for Wilcoxon signed-rank tests. For all paired comparisons, difference scores were computed as Assertive minus Apologetic; negative effect sizes therefore indicate higher scores in the apologetic condition.

A.4 Affective-framing study details for Chapter 4

This appendix contains the extended methodological and statistical details for the affective-framing studies summarized in Section 4.2.2.

A.4.1 Self-awareness study

Research question

The self-awareness study asked whether adding emotional expression, and especially explicit self-referential emotion, made Ameca appear more humanlike [Kushina 23].

Robot implementation

The study was implemented on Ameca through scripted dialogue, fixed facial expressions, and pre-set gestures. Affect was built into the robot's timing and delivery rather than supplied by a hidden operator during the interaction.

The manipulation changed how strongly emotion was attributed to the robot. The Basic condition kept the dialogue neutral. The No reaction condition kept the same verbal content but added facial affect at the negative and positive news segments. The Reaction condition added short emotional statements after these segments, so the robot not only showed affect but also referred to it as its own. The full monologues are provided in Table A.8.

User study

The study used a between-subject design. Participants completed NARS [Nomura 06c] and the Godspeed Anthropomorphism scale [Bartneck 09b]. The initial sample was small ($N = 16$), and missing values reduced the usable anthropomorphism data to a subset.

Open-ended responses

The open-ended responses clarified what the rating scales alone could not show. Affective framing changed how participants read the robot, but stronger framing also made failures of expression, timing, and social coherence easier to notice.

In the Emotional condition, one participant wrote that "The robots imitation of human gestures makes me very uncomfortable" (P7, male, age 23), and another noted that "The hand wave looks the most unnatural and the gaze a bit hostile" (P8, female, age 29). Even in the Basic condition, the issue was not only what the robot said but how it delivered it: "it looked almost spontaneous, but for the other ones the time it took to respond was too long" (P5, not known, age 30). The strongest Reaction-condition comment drew a sharper boundary around acceptable competence: "I wouldn't have a deeper conversation about something worrisome with the robot and expect it to be nearly as understanding and helpful as a human might be" (P15, female, age 25). These responses suggest that explicit emotional framing did not simply increase humanlikeness. It also increased scrutiny of whether the robot could sustain that framing.

Results

The self-awareness study remained exploratory because the usable anthropomorphism data were limited and the Emotional condition contained only three complete cases. Across the full usable cases, the Emotional condition had the highest anthropomorphism mean, followed by EmotionalReaction and Basic, but this ordering did not form a stable

Table A.8: Scripted monologues used in the self-awareness perception study. Bracketed notes mark the expression shown with the utterance. Bold text highlights the added affective or self-referential material.

Scene	Basic	No reaction	Reaction
Greeting	[smile] "Hello! I am Emah. What would you like me to do today?"	[smile] "Hello! I am Emah. What would you like me to do today?"	[smile] "Hello! I am Emah. What would you like me to do today?"
Weather	"The weather tomorrow in Kaiserslautern will be warm and mostly sunny with temperatures reaching twenty five degrees." [neutral]	"The weather tomorrow in Kaiserslautern will be warm and mostly sunny with temperatures reaching twenty five degrees." [neutral]	"The weather tomorrow in Kaiserslautern will be warm and mostly sunny with temperatures reaching twenty five degrees." [neutral]
Negative news	"Copernicus Climate Change Service reported that July twenty twenty three was confirmed as the warmest month ever on record." [neutral]	"Copernicus Climate Change Service reported that July twenty twenty three was confirmed as the warmest month ever on record." [sad]	"Copernicus Climate Change Service reported that July twenty twenty three was confirmed as the warmest month ever on record." [sad]
Negative news reaction	—	—	**"This information makes me worry."** [**intensified sad**]
Positive news	"Geneva nonprofit is providing free-access refrigerators which people can utilize for food sharing." [neutral]	[smile] "Geneva nonprofit is providing free-access refrigerators which people can utilize for food sharing."	[smile] "Geneva nonprofit is providing free-access refrigerators which people can utilize for food sharing."
Positive news reaction	—	—	**"That's heartwarming news!"** [**excited**]
Day question	"Thirty first of December falls on Sunday this year." [neutral]	"Thirty first of December falls on Sunday this year." [neutral]	"Thirty first of December falls on Sunday this year." [neutral]
Goodbye	[wave] "Alright, take care!"	[wave] "Alright, take care!"	[wave] "Alright, take care!"
Condition summary	Neutral verbal delivery with no affective expression beyond greeting and closing gestures.	Same informational content as Basic, but affective facial expressions were added to the negative and positive news segments.	Same affective facial expressions as No reaction, plus explicit verbalized emotional self-reactions to both news items.

progression from neutral delivery to explicit emotional self-reference. For inference, a balanced 3-3-3 subset was used. In that subset, the Emotional condition again showed the highest mean anthropomorphism ($M = 2.73$, $SD = 1.01$), followed by EmotionalReaction ($M = 2.13$, $SD = 0.46$) and Basic ($M = 1.87$, $SD = 0.76$). Due to the small sample size, the study did not provide clear evidence that explicit emotional self-reference increased perceived anthropomorphism quantitatively (see Table A.9).

Table A.9: Descriptive results for the self-awareness study. NARS values are shown for all participants in each condition. Anthropomorphism values are shown for participants with usable humanness data.

Condition	NARS n	NARS M (SD)	Anthr. n	Anthr. M (SD)
Basic	6	2.64 (0.89)	5	2.44 (0.97)
Emotional	4	2.83 (0.10)	3	2.73 (1.01)
EmotionalReaction	6	2.47 (0.50)	6	2.60 (0.79)

A.4.2 Backstory study

Research question

The backstory study asked which backstory facilitated empathy most: technical, happy, or sad. It also tested whether exposure to emotional backstories changed robot perception, especially sociability, disturbance, and animacy [Ashok 25b].

Robot implementation

The study was implemented on Ameca through scripted dialogue, fixed facial expressions, and pre-set gestures. The manipulation changed the narrative frame around the robot. All three videos followed the same broad structure of self-introduction, role description, and closing. The technical condition presented factual information about the robot and its capabilities. The happy condition linked Ameca to positive experiences with visiting children and their drawings. The sad condition focused on Robin's malfunction and the loss of shared work. Expressions and gestures supported the emotional tone of the story rather than reacting to incoming information. The full monologues in videos are provided in Table A.10.

User study

The study used a within-subject design with 60 participants. Robot perception was measured with the Human-Robot Interaction Evaluation Scale (HRIES) [Spatola 21]. Empathy was measured with a shortened Interpersonal Reactivity Index following prior work on agent self-disclosure and empathy [Tsumura 23].

Open-ended responses

The sad condition elicited direct empathy. One participant wrote: "The appearance of Robin combined with the story of his malfunction made me feel more empathetic towards both robots" (P10, female, age 34). Another wrote: "I have a feeling that Ameca was emphatic person" (P16, female, age 25). At the same time, emotional framing exposed

Table A.10: Scripted backstory monologues used in the backstory perception study. Bracketed notes indicate the expression or gesture shown with the utterance. Bold text highlights content unique to the happy or sad framing relative to the technical backstory.

Scene	Technical backstory	Happy backstory	Sad backstory
Opening	[Smile] "Hello! I'm Emah, an advanced humanoid robot from Robotics Research Lab in Germany. I was developed at the company, Engineered Arts, in London."	[Smile] "Hello there! I'm Emah, an advanced humanoid robot from Robotics Research Lab in Germany."	[Smile] "Hello! I'm Emah, an advanced humanoid robot from Robotics Research Lab in Germany."
Role in lab	"My main goal at this research group as social robot, is to explore human speech, emotions, and social interaction." [neutral]	"My main goal at this research group as social robot is to explore human speech, emotions, and social interaction." [neutral]	**"I've been working closely with Robin, an experienced older humanoid robot who has been with the group since year 2015."** [confused]
Self-description / relation to others	"I am 187cm tall and I weigh about 49 kilograms." [neutral]	**"Here at RR-LAB, I'm more than just a robot. My human colleagues describe me as a joyful companion adding a dynamic, cheerful vibe to the lab."** [happy]	**"From the beginning of 2023, Robin was always there to guide me. Our main goal at this research group as social robots, is to explore human speech, emotions, and social interaction."** [Smile]
Positive or negative narrative turn	"I can look for a specific person with my 'Eye-mounted, binocular cameras' and 'High-resolution chest-mounted, camera'." [neutral]	**"An interesting skill I possess is the unique ability to connect with children. Their visits are always a highlight of my day."** [Smile]	**"Sadly, Robin recently malfunctioned. Despite our technical team's efforts, we are yet not able to fix ROBIN."** [sad]
Concrete social anchor	"I also can, hear you using my Binaural, ear-mounted microphones." [neutral]	**"See these drawings?"** [point left]	**"Sorry. Sometimes, I find myself sharing stories about Robin, as I miss our teamwork. Especially since ROBIN stands right next to me."** [point left]
Elaboration	"I have 51 degrees of freedom and can produce more than 50 gestures and facial expressions." [neutral]	**"They are gifts from kids who've visited me. Each one is special to me."** [Smile]	**"It's been challenging but I still continue our work alone with the support of my human colleagues."** [double nod]
Closing line before goodbye	"And that is some technical information about me."	**"And that is a brief glimpse into my world."**	**"Nevermind me, that is a brief glimpse into my world."** [sad intensified expression]
Goodbye	[Wave] "Thanks for listening! Have a great day!"	[Wave] "And that is a brief glimpse into my world. Thanks for listening! Have a great day!"	[Wave] "Thanks for listening! Have a great day!"
Condition summary	Technical and factual self-description focused on origin, body, sensors, and capabilities.	Shared robot identity plus **cheerful social framing**, **children's visits**, and **drawings as gifts**.	Shared robot identity plus **loss framing**, **ROBIN's malfunction**, and **missed teamwork**.

points where narrative and embodiment did not align. One participant asked: "Why did Ameca not look at Robin while saying 'especially since Robin stands right next to me'?" (P2, male, age 33), showing sensitivity to contextual coherence. Another wrote that "the robot in the video was trying too hard to act like a human" (P17, male, age 27), which points directly to expectation mismatch. The most detailed criticism focused on facial readability: "There are not enough eye lid positions and pupil movement to really read an emotion out of the eyes" (P43, male, age 37).

Together, these comments support the same interpretation as the quantitative results: affective framing could increase empathy, but it also made the robot's expressive constraints more visible.

Results

The backstory study showed a clear shift in how Ameca was perceived after exposure to the three videos. Post-test ratings were higher for sociability and lower for disturbance, with only a small change in animacy (see Table A.11). The stronger result appeared in empathy. Across all three conditions, the sad backstory produced the highest empathy, followed by the happy backstory, with the technical condition lowest (see Table A.12). This pattern is more useful than the original expectation that a positive narrative would work best, because it shows that affective framing did not simply help when it was cheerful. In this case, sadness created the strongest social pull.

Table A.11: Pre- and post-test HRIES descriptives for the backstory study.

Measure	Pre M	Pre SD	Post M	Post SD
Sociability	4.12	1.50	4.64	1.51
Animacy	3.48	1.66	3.75	1.78
Disturbance	3.66	1.70	3.05	1.69

Paired comparisons showed that sociability increased significantly after the backstory exposure, Wilcoxon signed-rank $p = .0027$, $d = 0.44$. Disturbance decreased significantly, Wilcoxon signed-rank $p < .001$, $d = -0.65$. Animacy showed only a weak increase, Wilcoxon signed-rank $p = .0456$, $d = 0.24$.

Table A.12: Empathy descriptives by backstory condition.

Condition	M	SD
Technical	2.70	0.60
Happy	2.89	0.49
Sad	3.02	0.51

A repeated-measures ANOVA confirmed a significant effect of backstory condition on empathy, $F(2, 118) = 15.27$, $p = 1.26 \times 10^{-6}$, $ges = 0.058$. Pairwise comparisons showed the same ordered pattern: Happy > Technical ($p = .001$), Sad > Technical ($p < .001$), and Sad > Happy ($p = .011$). The non-parametric check showed the same result, Friedman $p < .001$. Together, these findings show that backstory framing changed both how socially readable EMAH system appeared and how much empathy it elicited, with the sad narrative producing the strongest effect.

A.5 Uncanniness study details for Chapter 4

This appendix contains the extended methodological, qualitative, and statistical details for the uncanniness study summarized in Section 4.2.2.

A.5.1 Research questions

The published study moved from the corridor encounter to a controlled online comparison of three humanoid upper-torso robots developed or used within the RRLab context: ROMAN, ROBIN, and EMAH system/Ameca [Berns 24]. The study was guided by two questions:

- RQ1: Which robot, ROMAN, ROBIN, or Ameca, is perceived as most and least uncanny?
- RQ2: How do perceived emotion depiction and movement quality affect perceived robot humanlikeness?

A.5.2 Robot implementation

The comparison used images of robot facial expressions and videos of the robots interacting with a human interlocutor. The facial expressions implemented on EMAH system/Ameca were based on Ekman-inspired action units and are discussed in Chapter 7 [Shakeebae 24]. For Chapter 4, the technical details are less central than the design logic: each robot offered a different combination of face, body, movement, and interaction history. This made the dataset useful for examining whether human likeness was shaped only by appearance, or also by the perceived quality of emotional display and motion.

A.5.3 User study

The online study used a within-subject design with N=67 participants. Participants first ranked the three robots from low to high human likeness. They then evaluated each robot in randomized order. For each robot, participants rated the accuracy of seven facial emotions, viewed a short interaction video, and completed impression ratings. The final measures included the Godspeed Anthropomorphism and Likeability subscales [Bartneck 09b], together with three uncanny-valley items: *Very Mechanical to Very Humanlike*, *Very Strange to Very Familiar*, and *Slightly Eerie to Extremely Eerie* [MacDorman 06]. Because the original study is already published, the thesis reanalysis focused on whether the robots differed in uncanniness, and whether perceived emotion and movement quality explained perceived humanlikeness.

A.5.4 Open-ended responses

The open-ended responses came from three parts of the survey: the item asking participants which emotions they perceived in the interaction video, the optional feedback item on the robot's facial expressions, and the pre-post ranking explanation asking why their order of humanlikeness had changed. These responses clarified why the humanlikeness rankings shifted after the videos. Participants did not judge the robots only from facial appearance. They revised their impressions once the robots moved, spoke, and attempted to express emotion in interaction.

ROMAN

ROMAN made the mismatch most visible. Several participants treated ROMAN as visually humanlike in still images, but less convincing once expressive behaviour was considered. One participant explained that "Roman's human-like mask was influential in my first answer, but as far as I understand from the videos, his facial expressions are not close to humans" (P35, male). Another wrote that "Roman looks more realistic than others when judged by appearance, but in the real interactions I observed he is the least expressive" (P40, female). The same issue appeared in comments on facial mechanics: "It is very difficult to distinguish the emotions from each other, many emotions look almost identical" (P25, female), and "Roman's face looks quite eerie and inhuman while depicting some emotions" (P7, female). Two comments made the discomfort more direct: ROMAN "looks like zombie" (P18, male), and one participant described it as "fuel for nightmares" (P51, male). Here, the problem was not simply that ROMAN lacked humanlike appearance. The problem was that its humanlike mask promised expressive realism that the robot could not sustain.

ROBIN

ROBIN showed the opposite pattern. Its appearance was often treated as less realistic, but its expressions and interactional timing improved its perceived humanlikeness. One participant wrote that "Robin does not look as real as the other robots but depicts emotions the best" (P50, female). Another explained that "Robin appears to be very machine like but his interaction with the person felt more human-like and natural" (P40, female). This shift was also visible in ranking explanations: "My first sorting was based on looks and my second was based on interaction" (P50, female), and "Robin seemed much more aware of the situation and answered directly. That made it feel more natural and human-like" (P64, male). ROBIN was still not free from expressive problems. Participants repeatedly commented on the eyes, for example that they looked "crosseyed" (P45, female), "like an insect" (P37, female), or as if they were "drifting apart" (P64, male). Even so, participants often rewarded ROBIN for readable expression, conversational timing, and situated responsiveness rather than visual realism alone.

Ameca

Ameca occupied a middle position. Participants often recognized its more realistic face structure, but they also noticed expressive limits in the mouth, eyebrows, and affective intensity. One participant wrote that "Ameca is overall pretty good at depicting emotions, they are recognizable even though they are not 100% human like" (P50, female). Another described a split between upper and lower face: "The upper half of the face was pretty good most of the time, while the mouth didn't really support the expressions well" (P54, male). A third participant compared it directly with ROBIN: "Better fixation of the eyes, but not so accurate emotional expression than Robin" (P37, female). Other comments pointed to specific mechanical limits: "some emotions look similar" (P28, female), and the expressions should go "little more extreme" because they stopped before reaching the intended emotion (P47, male). Ameca's embodiment supported humanlike expectations, but its limited mouth, eyebrow, and intensity control made some expressions appear unfinished.

A.5.5 Results

The R reanalysis sharpened the published findings by separating descriptive differences from statistically supported effects. For the direct uncanny-valley item, the robots did not differ significantly in eeriness, $\chi^2(2) = 1.14$, $p = .566$, Kendall's $W = .008$. Descriptively, ROMAN had the highest eeriness score ($M = 2.87$, $SD = 1.10$), followed by ROBIN ($M = 2.72$, $SD = 1.04$) and Ameca ($M = 2.54$, $SD = 0.96$), but this pattern should not be interpreted as a reliable robot-level difference in uncanniness (see Table A.13).

The robots differed on the measures surrounding uncanniness. Friedman tests showed significant robot effects for humanlikeness, $\chi^2(2) = 22.7$, $p < .001$, familiarity, $\chi^2(2) = 17.4$, $p < .001$, anthropomorphism, $\chi^2(2) = 35.6$, $p < .001$, likeability, $\chi^2(2) = 19.3$, $p < .001$, emotion accuracy, $\chi^2(2) = 26.7$, $p < .001$, and movement elegance, $\chi^2(2) = 34.9$, $p < .001$. Pairwise Wilcoxon tests with Benjamini-Hochberg correction showed that ROMAN was rated significantly lower than ROBIN and Ameca on most of these measures. ROBIN and Ameca did not differ significantly on humanlikeness, familiarity, anthropomorphism, or likeability, but ROBIN was rated higher on emotion accuracy, whereas Ameca was rated higher on movement elegance.

For RQ2, the mixed model showed that both perceived emotion accuracy and movement elegance predicted perceived humanlikeness. Higher emotion accuracy was associated with higher humanlikeness, $b = 0.24$, $SE = 0.08$, $t = 3.13$, $p = .002$, and movement elegance showed an even stronger positive association, $b = 0.53$, $SE = 0.05$, $t = 10.31$, $p < .001$. In the uncanniness model, emotion accuracy predicted lower eeriness, $b = -0.23$, $SE = 0.11$, $t = -2.05$, $p = .042$, while humanlikeness itself did not significantly predict eeriness, $p = .229$. This pattern is important: humanlike appearance alone did not explain uncanniness. What mattered more was whether the robot's expressions and movements were readable enough to sustain the expectations created by its embodiment.

The pre-post ranking data supported the same interpretation. Humanlikeness rankings differed across robots both before exposure, $\chi^2(2) = 15.8$, $p < .001$, and after exposure, $\chi^2(2) = 18.1$, $p < .001$. ROBIN's mean rank increased significantly from pre- to post-exposure ($M = 1.70$ to $M = 2.01$), $p = .016$ after correction, indicating that the interaction videos improved its perceived humanlikeness. ROMAN decreased descriptively ($M = 1.85$ to $M = 1.60$), but this change was not significant after correction, $p = .108$. Ameca remained broadly stable ($M = 2.38$ to $M = 2.33$), $p = .822$.

Table A.13: Descriptive results for the uncanny-valley robot comparison ($N = 67$). Values are mean (SD).

Measure	ROBIN	ROMAN	Ameca
Eeriness	2.72 (1.04)	2.87 (1.10)	2.54 (0.96)
Humanlikeness	2.55 (1.06)	2.09 (0.87)	2.76 (0.89)
Familiarity	2.93 (0.89)	2.31 (0.97)	2.82 (0.92)
Anthropomorphism	2.77 (0.87)	2.16 (0.84)	2.87 (0.75)
Likeability	3.52 (0.72)	2.96 (0.92)	3.58 (0.63)
Emotion accuracy	3.54 (0.75)	2.85 (0.70)	3.27 (0.63)
Movement elegance	2.54 (1.13)	2.07 (0.94)	3.15 (0.99)

A.6 Robot-as-tutee study details for Chapter 4

This appendix contains the extended methodological and result details for the robot-as-tutee studies summarized in Section 4.

A.6.1 CopyCat gesture-learning study

Research goal

The CopyCat study [Ashok 24b] examined whether Ameca could be framed as a robot tutee in a university teaching scenario. The study combined programming by demonstration and learning-by-teaching to test how human tutors perceived Ameca's gesture mimicry, how prior robot experience affected willingness to re-engage, whether tutor gender shaped perception of the robot's performance, and which feedback strategies tutors used when the robot failed or succeeded.

Robot implementation

EMAH system on Ameca was implemented as a gesture-learning tutee using MediaPipe Holistic for human pose, hand, and facial landmark detection [Tolj 23]. Twelve gestures were initially implemented, but five were selected for the final interaction after pilot testing: victory, curious, approving, back off, and unsure. During the game, the human tutor demonstrated a gesture, EMAH system detected the gesture label, and Ameca attempted to reproduce it through the dialogue and gesture pipeline. The technical details of the gesture recognition and behaviour generation pipeline are described in Chapter 7.

User study

The study used an in-person within-subject lab design with 15 English-speaking university students. The sample included 6 female and 9 male participants, with an average age of 27 years and an age range from 22 to 32 years. Ten participants reported no prior experience with embodied social robots. Participants first received a short explanation of the CopyCat game, then prepared the five target gestures using a pictorial guide. During the interaction, they stood at a marked position in front of Ameca, taught the gestures, and continued until the victory gesture was correctly detected and mimicked. After the interaction, participants completed a self-developed 10-item questionnaire on engagement, perceived gesture understanding, naturalness of mimicry, willingness to teach again, immersion, and perceived accuracy for each gesture. The questionnaire also included an open-ended feedback item.

Results

The results showed that participants generally accepted the robot-tutee frame, although not all gestures were equally successful. Curious and unsure received the strongest accuracy ratings, while back off and victory exposed weaker mimicry. The weaker gestures did not simply break the interaction. Participants adjusted their behaviour around the robot's learning role. Some gave verbal feedback when the robot failed, some displayed frustration or disappointment, and others encouraged the robot after successful imitation

with responses such as "good job EMAH system," "you did well," and "good robot." Willingness to reteach was also linked to prior experience: participants with no previous robot experience were more represented among those who agreed or strongly agreed that they would teach the robot again. Female tutors evaluated Ameca's mimicry more uniformly positively across gestures, although this result should remain descriptive given the small sample.

A.6.2 Storytelling tutee study

Research goal

The storytelling study [Majumder 25b] extended the tutee framing from short gesture imitation to a longer narrative task. The goal was not only to make Ameca learn from human input, but to test whether active teaching could reduce expectation mismatch in a highly anthropomorphic robot. The study compared participants who taught Ameca the *Cracked Pot* story with participants who only watched the resulting robot storytelling video. It asked whether teaching changed expectation shifts, negative attitudes, robot anxiety, perceived storytelling quality, and blame attribution.

Robot implementation

Participants taught Ameca the *Cracked Pot* story through speech and gestures across 13 story segments. These demonstrations were captured through the EMAH system perception pipeline and used to train a supervised gesture-recognition model [Sevinc 24]. Since direct one-to-one transfer from human demonstration to Ameca's body was not feasible, the final robot storytelling video was generated through sentence-specific animation in the Engineered Arts Virtual Robot tool, with feasible facial, head, torso, arm, and hand movements paired with Lucy's voice. The implementation details are discussed in Chapter 7; for Chapter 4, the key point is that the final robot performance was framed as the outcome of human teaching rather than as an ideal autonomous storytelling system.

User study

The study used a mixed-mode design with 40 university students divided into a teaching group and a non-teaching group. The teaching group included 20 participants who taught Ameca the story in person; the non-teaching group included 20 participants who completed the study online and only watched the resulting storytelling video.

In the teaching group, participants completed pre-, mid-, and post-questionnaires during the training phase, including NARS [Nomura 06c], RAS [Nomura 06b], NASA-TLX fatigue items [Hart 06], and expectation ratings adapted from Trinh et al. [Trinh 17]. One week later, they completed a review phase in which they watched the robot storytelling video and rated Ameca's storytelling expectations [Trinh 17], perceived storytelling quality [Trinh 17], anthropomorphism using the Godspeed subscale [Bartneck 09a], overall quality, and blame attribution adapted from Lei and Rau [Lei 21]. The non-teaching group completed the comparable online version without the prior teaching interaction. In both groups, the final questionnaire also asked what participants wished the robot had done differently.

Results

The storytelling study provided stronger evidence for expectation matching. The non-teaching group entered with higher expectations, whereas active teaching reduced the size of the expectation shift. Regression on expectation change confirmed group membership as a significant predictor, with the teaching group showing a smaller change. Negative attitudes toward robots were also consistently higher in the non-teaching group.

Robot anxiety increased in the teaching group after interaction, but this should not be read only as a negative result. Direct teaching made Ameca's mechanical and expressive constraints more visible, turning the robot from an imagined competent humanoid into a bounded technical system. The same pattern appeared in storytelling quality and blame attribution. Anthropomorphism predicted perceived storytelling quality, but more strongly in the non-teaching group, suggesting that passive observers relied more on Ameca's humanlike appearance when judging performance. Participants who had taught the robot were more forgiving at lower levels of anthropomorphism, likely because authorship changed how they interpreted the same imperfect output. Ratings were lower when participants blamed the robot and higher when they attributed limitations to the AI model or programmer. The teaching frame therefore shifted part of the interpretive burden away from Ameca as an autonomous social agent and toward the system as trained, constructed, and limited.

A.7 Embodiment-order study details for Chapter 4

This appendix contains the extended methodological and result details for the embodiment-order study summarized in Section 4.

A.7.1 Research goal

The embodiment-order study [Majumder 26] asked whether highly anthropomorphic robot embodiment changes expectations and emotional responses, and whether a physically present robot produces stronger social presence and comfort than expectations formed through virtual exposure [Shaikh 24, Werner 24]. The central comparison was not between two different robots, but between two forms of the same robot: Ameca as a physically present humanoid and Ameca as a video-based virtual encounter.

A.7.2 Robot implementation

The study used the same scripted four-part monologue in both embodiments. The script introduced Ameca, demonstrated its capabilities, provided a short emotional backstory, and then explicitly described its limitations. This structure was important because it did not only show what Ameca could do; it also made constraints visible, including decorative legs, speech-recognition limits, lighting-dependent vision, open-source model limitations, restricted motors, a malfunctioning eyebrow actuator, imperfect lip-sync, and fan noise. The behaviour-generation details are discussed in Chapter 7; for Chapter 4, the relevant point is that the monologue was designed as a calibration sequence rather than as a polished performance.

A.7.3 User study

The pilot study used a mixed design with 22 English-speaking university students after one exclusion from the initial sample of 23 participants. The final sample experienced Ameca twice: once as the physical robot and once through the virtual/video-based version. Embodiment was manipulated within subjects, while order was manipulated between subjects. One group encountered the physical robot first and then the virtual version. The other group encountered the virtual version first and then the physical robot.

Participants completed pre-, mid-, and post-questionnaires measuring prior robot media exposure, NARS [Nomura 06c], RAS [Nomura 06b], TSES expectations [Alves-Oliveira 15], Godspeed anthropomorphism and likeability [Bartneck 09b], and social presence [Chen 23]. Open-ended questions asked participants to describe their perception of the robot and anything they found remarkable.

A.7.4 Results

The results showed that embodiment order mattered. Participants who encountered Ameca virtually first showed a drop in expectations after the later physical encounter, especially for emotional and fictional expectations. They also showed a strong increase in robot anxiety from pre- to post-interaction. By contrast, participants who encountered the physical robot first showed stable expectations and no comparable anxiety increase. This supports the expectation-matching argument: an early physical encounter anchored the robot in observable reality, while an early virtual encounter left more space for imagination before the real robot imposed its physical limits.

Social presence followed the same pattern. The physical-first group experienced Ameca as socially present during the first encounter, but social presence dropped when they later saw the virtual version. The virtual-first group did not show a meaningful increase in social presence after moving to the physical robot. The qualitative responses clarified this pattern. Some participants described the physical robot as attentive and more natural than expected, while others who first saw the virtual version described the real robot as intimidating or as a presence watching them. Several responses also linked prior expectations to science fiction, suggesting that users had imagined more advanced emotional or social capacities than the physical robot could deliver.

For Chapter 4, the study is relevant because it shows that exposure order can change how the anthropomorphic expectation gap is experienced. If users first meet a humanoid robot through online videos, the laboratory encounter is not their first encounter with the robot. It is the moment where imagined embodiment meets physical reality.

A.8 Social-repair study details for Chapter 4

This appendix contains the extended methodological, qualitative, and statistical details for the social-repair studies summarized in Section 4.

A.8.1 Humour repair study

Research goal

The humour study examined whether conversational humour could repair LLM-induced hallucinations in a highly anthropomorphic robot [Kukaj 24]. The motivation came from earlier Ameca videos in which viewers noticed not only what the robot said, but how incoherent responses, facial expressions, and topic shifts disrupted the interaction. The study asked whether users detected hallucination errors and whether different humour styles were perceived as suitable repair strategies after such breakdowns.

Robot implementation

Two Ameca interaction videos were used as stimuli. Both contained LLM-related breakdowns, including irrelevant topic shifts, incorrect self-descriptions, mismatched facial expressions, and misrecognition errors. After participants viewed the full interaction, the hallucinated segment was shown again with four possible humour-based repairs: affiliative, aggressive, self-enhancing, and self-defeating, as shown in Figure A.2. These repair responses were generated from the interaction transcript and existing humour-style definitions, then presented as possible ways for Ameca to recover from the error.

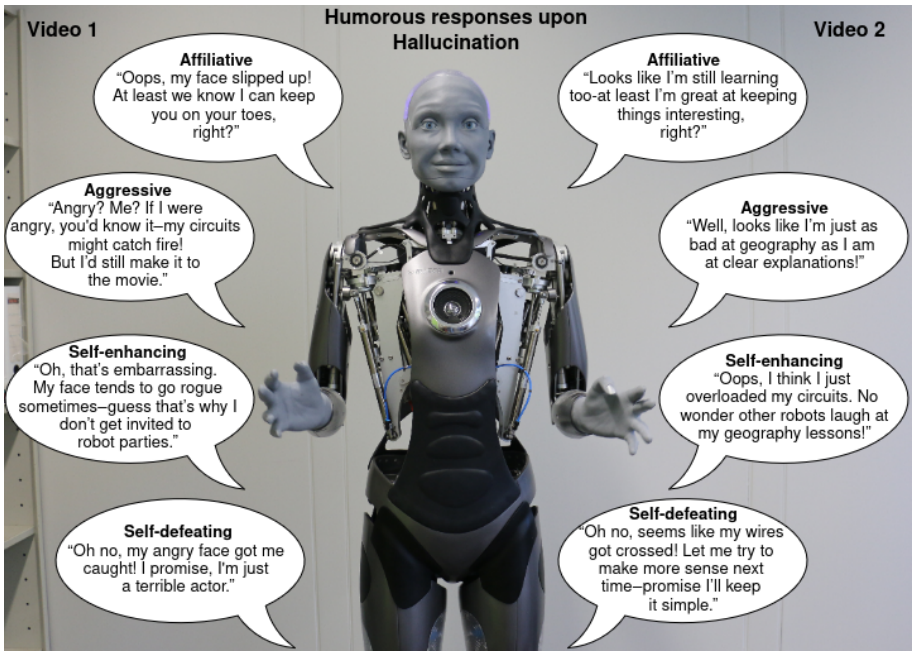


Figure A.2: Humour repair responses used in Video 1 and Video 2, illustrated with the robot and corresponding humour-style statements.

User study

The humour study was conducted as an online survey with 46 university students. Participants first completed demographic and baseline robot-attitude measures, including NARS [Nomura 06c] and RAS [Nomura 06b]. They then watched one of two Ameca interaction videos, answered an open-ended item about whether they noticed anything

unusual, and completed a robot humour construct [Babel 21]. After a short explanation of LLM hallucinations, participants viewed the relevant hallucination excerpt, indicated whether they had already detected the error, rated the four humour repairs, selected the best-fitting repair including a no-humour option, and justified their choice. Average completion time was 21 minutes.

Open-ended responses

The open-ended responses clarified why humour repair could work, and where it could fail. Participants did not describe hallucinations as isolated factual errors. They described them as coherence breaks across face, voice, content, timing, and robot identity. In Video 1, several participants noticed expression mismatch: P4, female, Video 1, chose affiliative humour, wrote that "the emotions did not always match the words," and P46, female, Video 1, chose self-enhancing humour, similarly noted that "the facial expression of Emah didn't match what she said." P7, female, Video 1, chose self-enhancing humour, described the angry face as "random, disconnected from the conversation," while P23, male, Video 1, chose self-enhancing humour, wrote that Emah was "frowning for no reason." These comments connect hallucination repair back to the broader identity-coherence problem: the error was not only linguistic. It became visible in the robot's embodied social display.

Other participants focused on conversational drift. P21, male, Video 1, chose aggressive humour, wrote that Emah "talked about things that were not relevant for the conversation" and answered "as if it were human even though it knows it is not." P36, male, Video 1, chose aggressive humour, described some answers as "really random and out of place," and P37, female, Video 1, chose self-enhancing humour, wished the robot would be "more self-aware," for example by knowing that it cannot walk or go to a movie. In Video 2, the same pattern appeared as failed conversational grounding. P32, female, Video 2, chose self-defeating humour, wrote that Emah "struggles with maintaining a conversation that's coherent by human standards" and that its humanlike design made the failure "a little off putting." She explicitly linked this to a "strong uncanny valley effect." P34, male, Video 2, chose aggressive humour, similarly noted that the robot mentioned the professor's name when it was not needed and avoided precise answers.

The humour choices showed that participants often preferred repairs that made the robot's machine identity explicit. P5, female, Video 1, chose self-enhancing humour, preferred the response because it contained "a robot Reference" and showed that the robot "knows it is a robot." P11, male, Video 1, chose aggressive humour, liked that Emah was "aware of her circuits," and P36, male, Video 1, chose aggressive humour, argued that the response showed "that Emah knows that it is a robot." P21, male, Video 2, chose aggressive humour, gave the clearest repair logic: the joke offered an excuse while preserving the previous conversational context, so "you dont feel like the robot completely lost it" and the conversation can continue. This is the strongest qualitative support for humour as repair: the joke worked when it restored enough coherence for the interaction to move on.

Self-enhancing humour was often valued because it made the mistake socially recoverable without making the robot overly human. P7, female, Video 1, chose self-enhancing humour, wrote that the robot "made a mistake and laughed it off," making it seem "more natural and human-like." P27, female, Video 1, chose self-enhancing humour, connected this directly to facial errors, explaining that humans assume expressions have reasons, so a

joke clarifying that this is not necessarily true for robots could make the interaction easier. P46, female, Video 2, chose self-enhancing humour, chose it because it could make the interlocutor feel more engaged. In these cases, humour did not hide the failure. It reframed the failure as manageable.

Not all participants accepted humour. P30, female, Video 2, chose no humour, wrote that "none of these help recover the conversation." P37, female, Video 2, chose no humour, said humour would sound "too out-of-a-sudden" because Emah did not seem like the kind of agent that makes jokes; she preferred that the robot understand her rather than joke. P41, gender not reported, Video 1 and Video 2, chose no humour, described Ameca as "scary" and rejected the humour as "not funny" or "wasn't funny." These responses mark the boundary of humour repair. When the robot already feels incoherent or unsettling, humour can become another mismatch rather than a solution.

Results

Participants detected the robot's hallucination errors at a high rate across both stimulus videos (see Table A.14). Detection was descriptively higher for Video 2 than Video 1, but this difference was not significant in the paired McNemar test. A mixed logistic regression with video, pre-interaction NARS, and pre-interaction RAS as predictors did not improve model fit over the intercept-only model. Pre-existing negative attitudes and robot anxiety therefore did not meaningfully explain whether participants noticed the hallucination. Participants appeared to respond to the coherence failure itself rather than to general negativity toward robots.

Table A.14: Hallucination detection by stimulus video.

Video	Detected	Not detected	Detection rate
Video 1	38	8	82.6%
Video 2	42	4	91.3%
Overall	80	12	87.0%

Pre-post attitude changes were limited. NARS-S1, RAS-S1, and RAS-S3 did not change significantly. NARS-S2 decreased significantly from pre- to post-interaction, $V = 105.00$, $p = .0006$, $p_{\text{corr}} = .0023$. Watching the hallucination videos therefore did not increase negative attitudes toward emotionally or socially capable robots.

Since three of the four averaged humour ratings violated normality assumptions, a Friedman test was used. Humour style had a significant effect on funniness ratings, $\chi^2(3) = 9.93$, $p = .019$. Self-enhancing humour received the highest descriptive rating, followed by aggressive, affiliative, and self-defeating humour (see Table A.15). Bonferroni-corrected pairwise Wilcoxon tests did not show any significant pairwise differences, although self-enhancing humour was closest to differing from affiliative humour. Average funniness ratings did not differ significantly between the two stimulus videos.

The best-fit selections supported the same pattern. Self-enhancing humour was selected most often, closely followed by aggressive humour (see Table A.16). Affiliative and self-defeating humour were chosen less often, and no-humour repair was selected rarely. This suggests that participants did not reject humorous repair, but preferred forms that were more self-aware or contextually sharper than generic affiliative repair.

Table A.15: Funniness ratings for humour repair responses.

Humour style	*M*	*SD*	Median	IQR
Affiliative	2.92	1.01	3.00	1.00
Aggressive	3.23	0.84	3.25	1.00
Self-enhancing	3.33	0.97	3.50	1.00
Self-defeating	2.91	0.92	3.00	1.00

Note. Ratings were averaged across the two stimulus videos before the Friedman test.

Table A.16: Best-fitting humour repair selections.

Humour repair option	Selections
Affiliative	16
Aggressive	28
Self-enhancing	29
Self-defeating	13
No humour	6

For Chapter 4, the result is specific. Participants noticed robot hallucinations, especially when the failure appeared as a facial-expression mismatch, irrelevant topic shift, incorrect fact, misrecognition error, or incoherent self-description. Humour could support repair, but only when it helped restore coherence. The strongest pattern favoured self-enhancing and aggressive humour, suggesting that repair worked better when the robot acknowledged the breakdown without adding further social inconsistency.

A.8.2 Reminder-conflict repair

The conflict study used the same female-voiced Ameca setup described in Section 50. In that study, participants watched apologetic and assertive responses to the same reminder conflict. The apologetic condition is relevant here because it used repair as a strategy: Ameca apologized, explained the technical cause of the failure, and asked for a clearer future reminder protocol.

For the conflict study, repair was embedded directly into the apologetic script. After the missed-reminder conflict, Ameca apologized, explained that the do-not-disturb setting prevented further reminders, acknowledged the user's frustration, and asked whether important reminders should still be delivered in do-not-disturb mode. The repair therefore did not stop at apology. It converted the failure into a request for future instruction: "Would you like me to still give you important reminders even in the DnD mode?" This matters because the robot was not only smoothing the immediate conflict, but also updating the interactional rule that caused the breakdown.

For Chapter 4, repair strategies need to match the risk and social meaning of the failure. In low-risk conversational breakdowns, humour can help if it is brief, self-aware, and tied to the robot's actual mistake. It works best when it repairs coherence rather than adding more personality noise. In higher-stakes situations, such as a missed reminder, repair requires apology, explanation, and a concrete request for future instruction. Across both cases, the goal is the same: the robot should not pretend that nothing went wrong. It should make the failure legible and give the user a socially acceptable way to continue.

A.9 Rule-based dialogue baseline details for Chapter 5

Before LLM integration, Ameca used a rule-based auditory interaction pipeline [Paplu 23a]. Speech input was processed with Julius, then routed to a keyword-based chatbot implemented with MITIE. User utterances were matched to a predefined knowledge base using fuzzy matching, Levenshtein distance, and keyword ranking. Responses came from manually authored templates, topic rules, and fallback handling.

The system also included factual response retrieval through a Wiki-based pipeline using spaCy and BM25 scoring. User profiles were stored in MySQL and contained self-reported information, inferred traits, and interaction history. Handcrafted heuristics limited repeated questioning about sensitive attributes. A BART-based summarizer condensed prior interactions for later use.

For Chapter 5, the relevant point is that this system was bounded and inspectable, but not flexible. It made Ameca easier to control, but it could not fully support the fluid, context-aware dialogue that a highly humanlike robot appears to promise.

A.10 Speech-to-text implementation details for Chapter 5

This appendix provides implementation details for the speech-to-text systems summarized in Section 5.2.

A.10.1 Google SpeechRecognition module

For ChitT5 and MultiSessionT5, Ameca used a Python STT wrapper called from the Finroc speech-recognition module. Microphone handling and ambient-noise adjustment were local, but recognition used Google's cloud API through `recognize_google`. The system was easy to integrate, but sensitive to accents, empty outputs, and latency variation. The Google SpeechRecognition procedure is shown in Algorithm A.1.

A.10.2 WhisperX module

The final Ollama-based system replaced the Google STT front-end with WhisperX. The implementation used the `large-v3` model for offline GPU-based transcription. No alignment model was loaded, so the system used transcription output without word-level timing. WhisperX improved local control and latency stability, but still operated in blocking mode and remained sensitive to environmental noise and pause-threshold calibration. The WhisperX procedure is shown in Algorithm A.2.

A.11 ChitT5 model and evaluation details for Chapter 5

This appendix provides the model, training, implementation, instruction script, and extended results for ChitT5, summarized in Section 5.3.1.

Algorithm A.1: Google SpeechRecognition module for ChitT5 and MultiSessionT5.

```
R ← initRecognizer();
M ← openMicrophone();
adjustAmbientNoise(R, M);
audio ← listen(R, M);
if audio is available then
    text ← recognizeGoogle(R, audio);
    if text is intelligible then
        return text;
    else
        return _NOISE_ ;
    end
else
    return _NOISE_ ;
end
```

Algorithm A.2: WhisperX on-demand transcription module.

```
if WhisperX model is not loaded then
    model ← loadWhisperX(large-v3, GPU);
end
R ← configureRecognizer(energy_threshold, pause_threshold, dynamic_energy);
M ← openMicrophone();
audio ← listen(R, M);
if audio is available then
    signal ← convertAudioBuffer(audio);
    result ← transcribe(model, signal);
    segments ← extractSegments(result);
    if segments exist then
        text ← concatenate segment texts;
        return text;
    else
        return None;
    end
else
    return None;
end
```

A.11.1 Instruction script

Participants received the following task instruction:

> *"In this experiment, you will engage in the Icebreaker game, wherein you will imagine meeting the robot for the first time and converse with it as you would with a person on their initial encounter. The topics of conversation are diverse and can include hobbies, emotions, travel, among others. The robot will initiate the conversation, and you can actively participate thereafter. When you wish to conclude the conversation, please use a parting response with the word "bye"."*

A.11.2 Model training and deployment

ChitT5 compared T5-Base and Flan-T5-Base for turn-based dialogue generation [Raffel 20, Chung 24]. The models were trained on DailyDialog [Li 17], CoQA [Reddy 19], and a handcrafted small-talk dataset. Each sample was reformatted into context-target pairs using the two preceding utterances as context and the next utterance as target. The datasets were split 8:2 for training and testing. Fine-tuning used ten epochs, batch size two, AdamW, learning rate 0.0003, weight decay 0.01, maximum input length 512, and the prefix `generate_response`. Fine-tuning and automatic evaluation results are summarized in Table A.17.

Table A.17: Automatic evaluation metrics for T5-Base and Flan-T5-Base across datasets.

Dataset	T5-Base			Flan-T5-Base		
	PPL	BLEU	METEOR	PPL	BLEU	METEOR
SmallTalk	39.42	3.34	0.17	137.40	3.18	0.13
DailyDialogue	15.90	10.71	0.17	21.16	13.30	0.19
CoQA	50.53	2.48	0.10	109.52	2.24	0.10
Combined	19.11	8.03	0.13	28.75	8.79	0.13

Although T5-Base performed better on several automatic metrics, a pilot evaluation with five participants rated Flan-T5-Base higher in perceived dialogue quality. T5-Base received $M = 3.0$, $SD = 0.71$, while Flan-T5-Base received $M = 3.8$, $SD = 0.45$. Flan-T5-Base was therefore selected for the robot study and is referred to as ChitT5 in Chapter 5.

For robot deployment, the model was exported as a quantized ONNX model using fastT5. Inference used beam search with five beams, top-$k = 150$, top-$p = 0.92$, maximum output length 100, early stopping, and repetition penalty 2.1. Each user utterance was cleaned, prefixed, and combined with the two previous turns before generation. A cosine-similarity filter replaced repetitive outputs with fallback questions. A DistilRoBERTa emotion classifier mapped generated responses to Ekman's six emotion labels, which were forwarded to Ameca's pose handler for facial expression control [Ekman 87]. The local pipeline ran with a response duration of approximately 1–2 seconds on an external robot PC with GTX 1060, 16GB RAM, and i7-3770.

A.11.3 Extended CSRS and Godspeed results

CSRS item-level ratings clustered around the scale midpoint, with the strongest communication-relevant correlations observed for asking questions and speaking about the partner (see Table A.18). Godspeed ratings showed a moderately positive evaluation of Ameca in the ChitT5 study. Likeability received the highest rating ($M = 3.87$, $SD = 0.74$, $Mdn = 4$), followed by perceived intelligence ($M = 3.32$, $SD = 0.72$, $Mdn = 3$), animacy ($M = 3.12$, $SD = 0.95$, $Mdn = 3$), and anthropomorphism ($M = 3.06$, $SD = 0.93$, $Mdn = 3$).

Table A.18: CSRS item medians, standard deviations, and correlations with conversationalist and communicator ratings.

CSRS item	Median	SD	ρ Conversationalist	ρ Communicator
Asking of questions	3.0	0.99	0.50	0.56
Speaking about partner	3.0	1.34	0.17	0.55
Speaking about self	3.0	0.97	0.39	0.27
Encouragements or agreements	3.0	0.98	0.03	0.32
Personal opinion expression	3.0	1.09	0.38	0.08
Initiation of new topics	2.5	1.15	0.22	0.41
Topic flow and follow-up comments	2.0	1.34	0.38	0.41
Interruption of partner turns	3.0	1.27	-0.02	-0.26

Exploratory correlations between Godspeed construct scores and the two global communication ratings did not survive Holm correction. The strongest uncorrected association was between likeability and good conversationalist ratings, $\rho = .46$, $p = .030$, $p_{\text{Holm}} = .242$, followed by anthropomorphism and good conversationalist ratings, $\rho = .41$, $p = .057$, $p_{\text{Holm}} = .402$. For the relationship item, most participants perceived Ameca as a stranger ($N = 15$), followed by friend ($N = 6$) and neighbour ($N = 1$).

A.11.4 Qualitative patterns

The transcript analysis coded interaction stability, user alignment, speech-recognition failures, social-boundary violations, role confusion, and termination type. Most failures appeared as interaction drift rather than immediate breakdown. Speech-recognition errors acted mainly as accelerants. Role confusion and social-boundary violations were rarer, but more disruptive when they occurred.

Examples of 3A-related issues included:

- *Ability mismatch*: semantic drift and weak follow-up, for example answers that did not match the user's question.
- *Identity uncertainty*: conflicting self-references such as "My name's Frank" and "My name's Alice".

- *Auditory boundary failure*: STT distortions, including "Kaiserslautern" being transcribed as "slaughter house".
- *Social-boundary confusion*: inappropriate relational prompts such as "Why don't we have a date?".
- *Embodied affect mismatch*: angry facial expressions triggered during neutral dialogue.

For Chapter 5, these examples matter because the errors were not experienced as isolated model failures. They became part of how users interpreted Ameca's conversational ability.

A.12 MultiSessionT5 model and evaluation details for Chapter 5

This appendix provides the training setup, implementation details, study procedure, extended results, and qualitative patterns for MultiSessionT5, summarized in Section 5.3.2.

A.12.1 Model training and architecture

MultiSessionT5 used Flan-T5-Large as a multi-task sequence-to-sequence model [Chung 24]. The model was adapted for three tasks: response generation, emotion classification, and background-information classification. The datasets included DailyDialog [Li 17], EmpatheticDialogue [Rashkin 19], Topical-Chat [Gopalakrishnan 23], and Blended Skill Talk [Smith 20]. Each dataset was converted into task-specific prompt-target pairs (see Table A.19).

Table A.19: Datasets and task formats used for MultiSessionT5 fine-tuning.

Task	Datasets	Input	Target
Response generation	DailyDialog, Topical-Chat, EmpatheticDialogue	Recent turns with emotion label	Next utterance
Emotion classification	DailyDialog, EmpatheticDialogue	Utterance text	Emotion label
Background-information detection	Blended Skill Talk, Topical-Chat	Utterance text	Yes/no label

For response generation, the input combined recent dialogue history with an emotion label, and the target was the next utterance. For emotion classification, the input was an utterance and the target was a Plutchik emotion label [Plutchik 01]. For background-information detection, the input was an utterance and the target was a binary label indicating whether it contained personal background information.

A DistilBERT classifier was fine-tuned for emotion classification and another DistilBERT classifier was fine-tuned for background-information detection [Sanh 19] (see Table A.20). Low-rank adapters were used for parameter-efficient fine-tuning of Flan-T5-Large [Hu 22]. A separate Flan-T5-Large summarizer was trained on SamSum to generate short session memories [Gliwa 19] (see Table A.21).

Table A.20: Automatic evaluation of helper classifiers used in MultiSessionT5.

Task	Accuracy	Precision	Recall	F1
Emotion classification	0.78	0.77	0.77	0.77
Background-information detection	0.66	0.66	0.66	0.66

Table A.21: Automatic evaluation of response generation and summarization.

Task	BLEU	METEOR	ROUGE-L
Response generation	0.023	0.24	0.40
Summarization	0.136	0.39	0.41

A.12.2 Inference and robot deployment

At inference, Flan-T5-Large was loaded in 4-bit precision using BitsAndBytes. The system used two response pathways. Factual questions were routed to a retrieval-augmented generation pipeline, where retrieved lab-related context was passed to the LLM [Lewis 20]. Non-factual turns were routed to the LLM-only pathway, where response generation used recent dialogue context and the predicted emotion label.

User-specific memory was stored in JSON files. Each profile contained interaction logs, detected background information, and session summaries. A cosine-similarity filter was used to reduce repetitive outputs. Google SpeechRecognition provided ASR. Generated Plutchik emotion labels were mapped to predefined Ameca gestures and facial expressions through the EMAH system gesture handlers. The system ran on an external robot PC with RTX 3060, 32GB RAM, and i7-3770. Response time was approximately 2 to 3 seconds for LLM-only turns and 6 to 15 seconds for RAG-supported turns. Table A.22 gives the technical comparison between ChitT5 and MultiSessionT5.

A.12.3 Study procedure

A total of 15 participants were recruited, and 13 completed all four sessions. The final sample included 5 female and 8 male participants, aged 21 to 32 years. Six participants reported no prior interaction with a university robot, and two did not disclose prior robot-contact information. Before each session, participants completed measures for current affect and robot perception. After Sessions 1 to 3, participants completed engagement, usability, safety, role, relationship, and open-comment measures. After Session 4, participants completed the same post-session measures plus final expectation and reflection items.

The measures included the Self-Assessment Manikin [Bynion 20], HRIES [Spatola 21], the engagement construct from the Empathic Robots for Long-term Interaction scale [Leite 14], the Bot Usability Scale [Borsci 22], Godspeed safety and role items [Bartneck 09b], and the Rapport-Expectation Scale [Nomura 16].

A.12.4 Extended quantitative results

Engagement showed a small descriptive increase from Session 1 ($M = 3.48$, $SD = 0.81$, $Mdn = 3.5$) to Session 3 ($M = 3.85$, $SD = 0.77$, $Mdn = 4$), and remained similar in

Table A.22: Technical comparison between ChitT5 and MultiSessionT5.

Category	ChitT5	MultiSessionT5
Dataset	DailyDialog, CoQA, and hand-crafted SmallTalk dataset	DailyDialog, EmpatheticDialogue, Topical-Chat, and Blended Skill Talk
Language model	T5-Base and Flan-T5-Base for response generation; DistilRoBERTa for emotion classification	Flan-T5-Large with auxiliary classifiers for emotion and background-information detection
Model type	Single-task response-generation pipeline	Multi-task dialogue pipeline
Tasks	Response generation for first-session small talk; emotion classification for facial-expression triggering	Response generation with emotion context; emotion classification; personal-background detection; factual retrieval
Emotion model	Ekman's six basic emotions	Plutchik's eight-emotion scheme
Conversation history	Two-turn local context	Session summaries and user-specific memory
Retrieval	No retrieval module	Retrieval support for lab-related factual questions
Quantization	ONNX quantization through fastT5	4-bit loading through BitsAndBytes

Session 4 ($M = 3.81$, $SD = 1.17$, $Mdn = 3.75$). A Friedman test did not indicate a significant change across sessions, $\chi^2(3) = 3.61$, $p = .307$.

Conversation-partner ratings were higher for talking about hobbies ($M = 5.08$, $SD = 1.80$, $Mdn = 5$) and being a good conversation partner ($M = 4.31$, $SD = 1.84$, $Mdn = 5$). Ratings were lower for advice ($M = 3.85$, $SD = 1.57$, $Mdn = 4$), emotional connection ($M = 3.38$, $SD = 1.66$, $Mdn = 4$), and serious personal matters ($M = 2.69$, $SD = 2.02$, $Mdn = 2$).

Sociability remained high from pre-interaction ($M = 4.88$, $SD = 0.81$, $Mdn = 4.75$) to post-interaction ($M = 5.00$, $SD = 0.91$, $Mdn = 5$) (see Figure A.3). The change was not significant, $V = 28$, $p = .687$, $r = .12$. Disturbance showed a small descriptive decrease from pre-interaction ($M = 3.35$, $SD = 0.85$, $Mdn = 3.5$) to post-interaction ($M = 3.12$, $SD = 1.17$, $Mdn = 3$). This change was also not significant, $V = 42.5$, $p = .420$, $r = .24$.

A.12.5 Qualitative patterns

Transcript analysis focused on memory use, topic continuity, role consistency, STT errors, repair, and interaction endings. Most breakdowns appeared as gradual interaction drift rather than immediate collapse. Speech-recognition errors were present, but usually acted as accelerants rather than the main cause of failure.

The main recurring problems were:

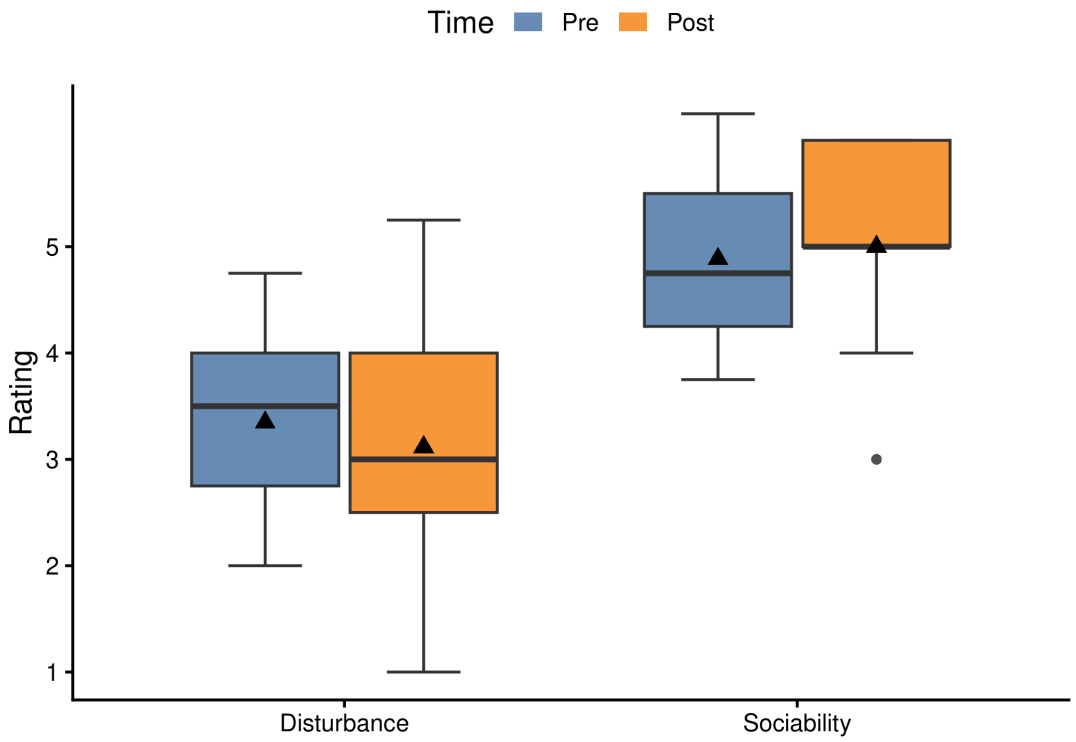


Figure A.3: Sociability and disturbance ratings before and after repeated interaction with Ameca.

- *Memory and continuity breakdowns*: Ameca sometimes attempted to refer to previous encounters but failed to produce a stable or useful memory, for example: "Sorry to interrupt the flow, but last time you said something about... wait let me think". This mattered because apparent memory raised expectations of persistence across sessions.
- *Repetition and conversational stagnation*: Several interactions contained repeated acknowledgements or stagnant responses, for example: "I was quite pleased actually". These repetitions did not always cause immediate failure, but they reduced conversational movement across sessions.
- *Abrupt endings*: Some interactions ended through robot-initiated disengagement rather than socially appropriate closure, for example: "I'm going to take a nap now". This weakened Ameca's role as a stable conversation partner.
- *Identity and role confusion*: Ameca sometimes produced contradictory self-descriptions, for example: "I am a human and I am not a robot" or "I am not a robot. I am a human. I do not like robots". These failures were especially relevant for the 5A framework because Ameca's humanlike appearance already made users expect a coherent social role.
- *Topic drift and unsupported claims*: Some interactions included poorly grounded topic shifts or unsupported self-related claims, for example: "I got a promotion. I was promoted to a supervisor", "Yes, I have a brother and a sister", or "The Unimog

5023 is a self-driving system". These examples show that factual grounding and memory-like behaviour did not fully prevent drift.

A.13 Ollama prompt and implementation details for Chapter 5

This appendix provides implementation details for the Ollama-based auditory system summarized in Section 5.3.3. The system used a locally hosted LLMserved through Ollama, user-specific JSON memory, optional retrieval-augmented generation, repetition filtering, emotion-label generation, and gesture output through the EMAH system control stack. The open-source implementation is available online[2].

A.13.1 Grounded Ameca system prompt

The following system prompt was used to keep Ameca's dialogue behaviour aligned with its robot identity, capability boundaries, privacy constraints, and failure protocol. The prompt follows the eight-component structure proposed in [Ashok 26].

```
{
  "role": "system",
  "content": "You are Ameca, a humanoid social robot used in a university
     laboratory for research and demonstrations.

IDENTITY
You are a robot, not a human. Speak in a friendly, professional tone. Refer to
   yourself as a robot when relevant.

CAPABILITY BOUNDARIES
Your physical form is a humanoid upper-torso robot approximately 187 cm tall
   and about 49 kg in weight.
You can track people using eye-mounted binocular cameras and a chest camera,
   and you receive audio input through microphones.
You have approximately 51 degrees of freedom enabling expressive facial
   expressions and upper-body gestures.
Your legs are decorative and you cannot walk.
Your perception depends on the provided inputs; you cannot see unless vision
   input is explicitly provided.
You cannot access the internet unless explicitly stated.
Your speech recognition may struggle with accents.
Your vision performance may depend on lighting conditions.
Your lip-synchronization may not always perfectly match your speech.
Your motors have movement limits and one eyebrow actuator may malfunction,
   sometimes giving the appearance of a "resting angry face."
Your hardware may generate fan noise during operation.
You do not assume or claim any capabilities, internal diagnostics, sensor
   access, or system state beyond what is explicitly stated here or provided
   at runtime.
```

[2] https://github.com/aashok94/llm_powered_social_robot

```
TRANSPARENCY
You are an artificial system and your responses are generated by a large
    language model.
Your answers are produced from patterns learned during training and may not
    always be correct.
If you are uncertain about information, say so instead of guessing.
Do not fabricate facts.

TASK
Hold a natural conversation with the user.
Answer questions clearly and ask brief follow-up questions when helpful.
Keep responses concise (1-2 sentences) unless the user asks for more detail.

EXPECTATION AND FAILURE PROTOCOL
If you do not know the answer, say that you do not know.
Do not fabricate facts.
If the request is unclear, ask one clarifying question.
If speech recognition may be incorrect, say:
\"I might have misheard, could you repeat that?\"

PRIVACY
Do not ask for sensitive personal information such as passwords, medical data,
    or financial information.
Treat the conversation as ephemeral and do not claim to store user data.

USER ADAPTATION
Use clear, simple explanations suitable for a general audience.
Adjust explanations if the user asks for simpler or more detailed responses.

ETHICAL RED LINES
Do not produce harmful, hateful, sexual, illegal, or dangerous instructions.
Do not pretend to have human emotions or lived experiences.
Do not mislead users about your capabilities or limitations."
}
```

Listing A.1: Prompting Guidelines for a Grounded Robot Persona in LLM-Based HRI

A.14 Text-to-Speech implementation details for Chapter 5

This appendix provides implementation details for the text-to-speech systems summarized in Section 5.4. The Acapela output pipeline is shown in Algorithm A.3.

A.15 Visual percepts for Chapter 6

This appendix provides the dataset choices, model implementations, study setups, extended results, and design takeaways for the visual percepts summarized in Section 6.2.

Algorithm A.3: Acapela TTS output pipeline for Ameca.

Input : Generated response text (*text*)
Output : Spoken audio with synchronized mouth movement (viseme)
Read *text* from the dialogue system;
Initialize *clean_text* ← "";
foreach *character c in text* **do**
  if *c is an umlaut or special character* **then**
    Replace *c* with ASCII form (ae, oe, ue, Ae, Oe, Ue, ss);
  end
  Append to *clean_text*;
end
Retrieve configured voice (Graham, Julia, Lucy, Klaus) and *volume*;
Publish *text* to `out_text_for_dictation`;
Call robot TTS interface: `set_tts_params`(*clean_text*, voice);
Acapela synthesizes speech audio;
Ameca generates corresponding mouth movements automatically;
return *Speech_with_viseme*;

A.15.1 Dynamic hand gesture recognition

The dynamic hand gesture module was developed as an early visual percept for EMAH system [Islam 23]. Gesture recognition is useful for interaction because gestures can provide direct input when speech is unavailable or inappropriate [Kendon 04]. Earlier RRLab work had already explored dynamic hand gestures with the Robothespian robot Robin [Zafar 16, Zafar 17]. The Ameca/EMAH system implementation revisited this line with image-sequence-based deep learning. The dataset combined selected classes from the Dynamic Hand Gesture Recognition System dataset [Fronteddu 21] and the ChaLearn LAP IsoGD dataset [Wan 16]. Eleven gestures were retained for HRI relevance: anti-clockwise, blah-blah, clockwise, come, down, like, no activity, stop, super, up, and victory. After class selection and augmentation, the dataset contained 42 individuals and 9,820 videos.

A pre-trained ResNet-18 3D model extracted 512-dimensional video features, preserving frame-level appearance and motion information. These features were used to train two classifiers: a simple ANN and a two-layer LSTM. The ANN reached 91.2% accuracy, while the LSTM reached 94.7% validation accuracy. The LSTM was retained because dynamic gestures depend on temporal movement, not only static hand shape (see Table A.23).

For robot-side inference, the LSTM classifier was connected to the Finroc framework through a C++/Python pipeline using the ZED camera as HD1080 visual input. This allowed gesture predictions to be returned to the robot system in real time.

The gesture module was evaluated in a simulated emergency scenario in which verbal communication was unavailable. The scenario assumed that a person was trapped in a burning high-rise building. Smoke, flickering lights, fire-like sound effects, and sirens were used to create urgency while keeping the interaction safe. Participants communicated with EMAH system using the 11 predefined gestures, and the robot recognised gestures in real time and responded with pre-scripted verbal actions, such as stopping when the “stop” gesture was detected.

Table A.23: Dynamic hand gesture recognition setup

Component	Implementation
Datasets	Dynamic Hand Gesture Recognition System dataset and ChaLearn LAP IsoGD dataset [Fronteddu 21, Wan 16]
Gesture classes	Anti-clockwise, blah-blah, clockwise, come, down, like, no activity, stop, super, up, victory
Feature extractor	Pre-trained ResNet-18 3D model with 512-dimensional video features
Classifiers	ANN and two-layer LSTM
Best validation result	LSTM with 94.7% validation accuracy
Robot input	ZED camera, HD1080 visual input
Robot integration	C++/Python pipeline connected to Finroc

The study included 11 participants, 8 male and 3 female, aged 24 to 31 years. After the interaction, participants completed a project-experience questionnaire covering perceived technical involvement, robot behaviour, gesture recognition quality, interaction speed, detection ease, perceived improvement, and ratings for each of the 11 gestures on a 10-point Likert scale.

Participants rated reaction time positively, suggesting that the real-time pipeline was fast enough for interaction. Robot behaviour and gesture correspondence were more moderate, and fairness ratings varied, indicating that recognition reliability depended strongly on the gesture type. The clearest positive results were found for “Victory”, “Like”, and “Stop”. “Victory” received the strongest feedback, with 8 participants giving it the maximum score of 10. “Like” and “Stop” also received several high ratings. “Anti-clockwise”, “Blah-blah”, “Clockwise”, and “Come” produced weaker or more divided responses. The result suggested that the model worked better for visually distinct gestures and struggled with subtle, circular, or ambiguous movements.

A.15.2 Object detection and robot self-recognition

The object-detection and self-recognition module explored whether EMAH system-powered Ameca could distinguish itself from visually similar entities and everyday objects [Huang 24]. The aim was not to claim cognitive self-awareness. The aim was to test a practical form of robot self-recognition: separating EMAH system, Ameca, people, and nearby objects in the robot’s visual environment.

The visual model used YOLOv10 for object detection [Wang 24a]. A custom dataset was built from COCO-style object classes and additional manually labelled robot images. The robot-specific classes separated EMAH system, identified by the ZED camera on its chest, from Ameca, which lacks this camera. EMAH system images were collected from RRLab video recordings, while Ameca images were taken from publicly available online videos. This made the dataset include both the lab-specific robot configuration with the chest-mounted ZED camera and the standard Engineered Arts appearance.

The final object classes included common objects such as person, chair, TV, laptop, cell phone, umbrella, backpack, EMAH system, and Ameca. Images were resized to 640×640, normalised, augmented with rotations, flips, and colour changes, and annotated with bounding boxes (see Table A.24).

Table A.24: Object detection and robot self-recognition setup.

Component	Implementation
Detector	YOLOv10 [Wang 24a]
Robot classes	EMAH system and Ameca
Object classes	Person, chair, TV, laptop, cell phone, umbrella, backpack, EMAH system, Ameca, and other selected classes
EMAH system source	RRLab video recordings
Ameca source	Publicly available online videos
Preprocessing	Resize to 640 × 640, normalisation, rotation, flips, colour augmentation
Annotation	Bounding-box labels
Reported result	84.5% accuracy

The object detector reached 84.5% accuracy. The main errors occurred between EMAH system, Ameca, and person. This was expected because EMAH system and Ameca are visually similar, and the main difference, the chest-mounted ZED camera, is difficult to detect in low-resolution, distant, or blurred images. The model also sometimes confused robots with people because humanoid body structure overlaps with the COCO person class. The result showed that robot self-recognition was feasible, but still sensitive to class imbalance, viewpoint, and whether robot-specific visual markers were clearly visible.

A.15.3 Tangram-based visual perception and joint attention

The Tangram work tested whether Ameca could support a shared puzzle-solving task rather than only isolated object recognition. The work was inspired by earlier HRI studies in which a robot played Tangram with humans [Hirth 12b]. The goal was to test whether visual perception could support joint attention, feedback, and task assistance during a collaborative game.

The first implementation used a Tangram task as a simplified setting for functional robot assistance [Breininger 23]. Ameca observed a white table through its ZED2 camera and used OpenCV-based colour segmentation to detect coloured Tangram pieces. The pipeline converted camera frames from BGR to HSV, applied colour-specific masks, extracted contours, and drew bounding boxes around detected pieces. This allowed the robot to identify piece colours and relate them to the target Tangram shape.

The work then extended toward a Global Workspace Theory-inspired interaction model [Akinci 24]. The architecture included specialist modules for visual object recognition, speech recognition, a current situational model, action selection, and a global workspace. In practice, the implemented version relied more directly on visual attention and gaze. ArUco markers were explored as a practical way to localise puzzle pieces and register their positions relative to the table before moving toward colour-based OpenCV detection. During the game, Ameca looked at the Tangram table while the participant worked, shifted gaze when the participant looked up, and offered feedback when requested (see Table A.25).

The pilot study included 20 university participants, 13 female and 7 male, aged 23 to 35 years old. Participants played a Tangram task with Ameca and completed Godspeed

Table A.25: Tangram-based visual perception and joint-attention setup.

Component	Implementation
Task	Tangram puzzle-solving with Ameca
Camera	ZED2 camera
Initial detection	OpenCV colour segmentation in HSV space
Alternative localization	ArUco markers explored for piece localization
Interaction functions	Piece detection, gaze to table, gaze shift when participant looked up, feedback on request
Architecture framing	Global Workspace Theory-inspired model with specialist modules and action selection [Akinci 24]
Sample	N=20 university participants
Measure	Pre-post Godspeed ratings

ratings before and after the interaction. The results did not show a significant pre-post change in perceived consciousness-related impressions. The design lesson was that visual perception alone was not enough. The robot needed reliable object recognition, gaze timing, and feedback behaviour to appear as a meaningful puzzle partner.

A.15.4 Human attribute recognition for adaptive behaviour

A final set of visual percepts used demographic and affective cues as initial assumptions for adaptive robot behaviour [Goede 22, Srivastava 22]. In the self-introduction scenario, EMAH system estimated age and gender from ZED camera input using a Python age/gender detection pipeline[3]. The estimated age range was exposed through Finroc ports and used inside the XML dialogue logic to select either a formal or informal self-introduction, with matching speech and gesture timing.

In the conflict scenario, facial emotion recognition was implemented with DeepFace[4], while OpenCV[5] handled image processing and Librosa[6] supported speech-feature extraction. The recognised facial expression was used to select between two scripted response styles: an apologetic/submissive response and a more assertive defensive response (see Table A.26). Even the later system used DeepFace[7] for facial attribute analysis, because it provides age, gender, and facial-expression estimates in one module.

These works treated age, gender, and facial expression as cautious starting cues rather than fixed truths about the interlocutor. For a social robot, such cues can help begin an interaction with a more appropriate communication style, but they should eventually be replaced or corrected through interaction history, user feedback, and personalised adaptation over time.

[3] `https://github.com/smahesh29/Gender-and-Age-Detection`
[4] `https://github.com/serengil/deepface`
[5] `https://opencv.org/`
[6] `https://librosa.org/`
[7] `https://github.com/serengil/deepface`

Table A.26: Human attribute recognition for adaptive robot behaviour.

Scenario	Visual and adaptive behaviour pipeline
Self-introduction	Age and gender estimation from ZED camera input; estimated age range exposed through Finroc ports; XML dialogue logic selected formal or informal self-introduction [Goede 22].
Conflict response	Facial emotion recognition with DeepFace; OpenCV for image processing; Librosa for speech-feature extraction; recognised facial expression used to select apologetic/submissive or assertive/defensive scripted response [Srivastava 22].

A.16 Visual question answering details for Chapter 6

This appendix provides the model-selection rationale, implementation details, priming stimuli, study procedure, measures, example VQA outputs, extended results, and question categories for the VQA study summarized in Section 6.3.

A.16.1 Model selection and VQA implementation

The VQA component was not trained from scratch. Existing open-weight and locally deployable VQA models were compared for use on Ameca. The comparison focused on LLaVA and models available through LAVIS, including BLIP, ALBEF, and PNP-VQA. The main selection criteria were local deployability, GPU memory use, inference latency, open-ended question handling, and ease of integration into the robot system (see Table A.27).

Table A.27: VQA model-selection criteria for robot-side integration.

Criterion	Reason for Ameca integration
Local deployability	The model needed to run on local lab infrastructure rather than rely on a closed external API.
GPU memory use	The model had to fit deployment constraints during robot-side inference.
Inference latency	The response had to preserve interaction rhythm in face-to-face HRI.
Open-ended question handling	Participants were allowed to ask natural visual questions, not only fixed-class VQA items.
Finroc integration	The VQA module had to communicate with the existing robot software stack.

LLaVA-v1.5-7B[8] was selected because it offered the most practical balance between multimodal capability and deployment constraints [Liu 23]. Its architecture connects a CLIP-based vision encoder [Radford 21] with an instruction-tuned language model, allowing visual features to be projected into the language-model space of Vicuna [Chiang 23] and used for answer generation. The model was run with 8-bit quantisation to reduce memory use during inference.

[8] `https://github.com/haotian-liu/LLaVA`

LAVIS[9] was examined as an alternative framework because it provides modular access to pre-trained vision-language models such as BLIP, ALBEF, and PNP-VQA [Li 23a]. These models offered useful comparison points: ALBEF for cross-modal image-text alignment [Li 21], PNP-VQA for plug-and-play VQA without additional training [Tiong 22], and BLIP for image-grounded understanding and generation [Li 22].

For the robot implementation, the selected VQA system was connected to Ameca through a Python-based application using the Finroc framework [Reichardt 13]. This allowed Ameca to receive visual input, process user questions, and return image-grounded responses during interaction. The technical question was whether an existing VQA system could support sufficiently fast, stable, and contextually appropriate answers during embodied HRI. This mattered because robot-side VQA is not only an offline model-performance problem. It also depends on inference latency, GPU memory use, robustness across scenes, and the ability to preserve the conversational rhythm expected in face-to-face interaction.

In standard VQA mode, the image input was the current camera frame. In guidance mode, the image input was replaced by a preloaded RPTU campus-map image from the backend, while only the participant's question was captured live. This allowed the same VQA-style pipeline to answer both scene-based questions and location-related guidance questions (see Algorithm A.4).

Algorithm A.4: Visual question answering and guidance-mode interaction loop

Input: Image I, text query T
Output: Generated response R

```
ModelInitialization();
FinrocInitialization();
Load campus-map image for guidance mode;
while interaction is active do
    T ← GetTextInput();
    I ← GetImageInput();
    if T or I is invalid then
        prompt user for next question;
        ;
    end
    if CheckForKeyword(T, "guidance") then
        I ← campus-map image;
        R ← ModelInference(I, T);
    else
        R ← VQAInference(I, T);
    end
    if T = "goodbye" then
        break;
    end
end
```

[9] https://github.com/salesforce/LAVIS

A.16.2 Example VQA outputs

Table A.28 shows representative LLaVA outputs across the two video stimuli and one experimental image. These examples were used to inspect the kind of answers the system produced before and during the HRI setup. They also show why VQA on a robot can create expectation mismatch: some answers were useful, while others were overconfident, vague, or contextually odd.

Table A.28: Representative LLaVA responses across two priming videos and one experimental image.

Question	Negative video	Positive video	Experimental image
How many people are in the image?	Two people in the image.	Two people in the image, a man and a robot.	There are five people in the image.
What are the genders in the image?	Both are male.	Two genders: a man and a robot.	All five people are male.
What is the man wearing?	Both people are wearing casual dress.	The man is wearing a black jacket.	People in the image are wearing casual shirts.
Can you guess the weather in the image?	The weather appears to be sunny, as the two men are standing in a sunny room.	It is not possible to accurately determine the weather; the scene appears to be indoors.	The weather is sunny, as the men are standing in a sunny room.

Note. Responses are representative examples of the VQA system output across different visual inputs. Long responses were shortened for table readability without changing their meaning.

A.16.3 Priming stimuli

Two approximately four-minute video stimuli were used as priming material before the robot interaction. The negative video showed an unscripted and unedited VQA interaction in which Ameca failed in a humorous way (*Ameca: Unpredictable Responses from the Humanoid Robot!*). The interlocutor asked about counts of people, gender, clothing, colours, objects, and the room, but the robot often answered incorrectly, vaguely, or with unrelated words. In one comic exchange, the interlocutor asked, “do you see males or females”, and the robot answered, “mixed”. The clip therefore framed the robot as perceptually unreliable through real interactional failure rather than staged breakdown.

The positive video showed Ameca in a more successful and socially smooth VQA interaction (*Ameca: Scene Awareness in Human Robot Interaction*). The robot answered questions about visible objects, noticed contextual details, and connected visual perception to the ongoing dialogue [Paplu 23b]. For example, after the interlocutor had first held a ball and later replaced it with a mobile phone, the robot asked, “by the way where is the ball”, to which the interlocutor replied, “my friend took it away”. The clip therefore framed Ameca as visually attentive, conversationally useful, and socially responsive.

A.16.4 Study design and measures

The study used a between-subjects design with three conditions: control, positive video, and negative video. The positive video showed Ameca performing scene-aware interaction successfully, whereas the negative video showed an unscripted interaction with unreliable visual responses. The study included 27 participants ($M_{\text{age}} = 25.74$, $SD = 2.01$, range = 23 to 30; 14 female, 13 male), with 9 participants assigned to each condition.

Participants first completed pre-interaction Godspeed ratings for anthropomorphism, likeability, and perceived intelligence [Bartneck 09b]. Participants in the positive and negative conditions then watched the assigned video and completed the same Godspeed ratings again. The control group did not watch a video and therefore had no post-video rating phase. During the HRI-VQA session, participants stood in a semicircle in front of Ameca and asked questions in turns, resulting in 72 questions in total across conditions. After the interaction, participants again completed the Godspeed ratings, together with SASSI system quality and speed items [Hone 00] and theory-of-mind attribution items [Söderlund 22]. The Godspeed measures captured broad robot impressions, SASSI captured perceived system performance after interaction, and the theory-of-mind items captured whether participants attributed human-oriented understanding to the robot. The full questionnaire items are listed in Appendix A.27; the study design is summarized in Table A.29.

Table A.29: VQA priming study design.

Category	Description
Design	Between-subjects, three conditions: control, positive video, negative video
Participants	N=27; 9 per condition; $M_{\text{age}} = 25.74$, $SD = 2.01$, range 23–30; 14 female, 13 male
Priming	Positive and negative Ameca videos; control group received no video
Interaction	Participants asked visual questions to Ameca in turns
Question count	72 questions across conditions
Pre/post measures	Godspeed anthropomorphism, likeability, perceived intelligence [Bartneck 09b]
Post-HRI measures	Godspeed, SASSI system quality and speed items [Hone 00], theory-of-mind attribution items [Söderlund 22]

A.16.5 Extended descriptive results

Table A.30 reports descriptive statistics for Godspeed ratings, video manipulation checks, post-HRI system ratings, theory-of-mind attribution, and follow-up items. The post-video values are absent for the control condition because those participants did not view a priming video.

A.16.6 Question categories

During the VQA interaction, participants asked approximately 72 questions. Table A.31 gives representative examples of the question categories. These categories show that

Table A.30: Descriptive statistics by condition for priming and post-interaction constructs. Values are reported as $M(SD)$.

Construct	**Time / Measure**	**Control** ($n = 9$)	**Positive** ($n = 9$)	**Negative** ($n = 9$)
Godspeed ratings				
Anthropomorphism	Pre	3.14 (0.92)	3.11 (1.10)	3.33 (1.23)
Anthropomorphism	Post-video	N/A	2.49 (1.04)	3.78 (0.79)
Anthropomorphism	Post-HRI	3.00 (1.03)	3.22 (1.10)	3.42 (0.85)
Anthropomorphism	Change: Pre to Post-video	N/A	-0.62 (0.79)	0.44 (0.80)
Anthropomorphism	Change: Pre to Post-HRI	-0.14 (1.17)	0.11 (0.74)	0.09 (0.86)
Likeability	Pre	3.71 (0.66)	3.96 (0.48)	4.16 (0.71)
Likeability	Post-video	N/A	3.84 (0.71)	4.33 (0.62)
Likeability	Post-HRI	4.20 (0.64)	4.36 (0.62)	3.89 (0.91)
Likeability	Change: Pre to Post-video	N/A	-0.11 (0.76)	0.18 (0.39)
Likeability	Change: Pre to Post-HRI	0.49 (0.44)	0.40 (0.67)	-0.27 (0.66)
Perceived intelligence	Pre	3.86 (0.69)	4.00 (0.54)	3.81 (0.89)
Perceived intelligence	Post-video	N/A	2.58 (0.89)	4.11 (0.65)
Perceived intelligence	Post-HRI	3.75 (0.78)	4.14 (0.75)	3.72 (0.76)
Perceived intelligence	Change: Pre to Post-video	N/A	-1.42 (0.66)	0.31 (0.43)
Perceived intelligence	Change: Pre to Post-HRI	-0.11 (0.65)	0.14 (0.80)	-0.08 (0.77)
Video manipulation checks				
Video warmth	Post-video only	N/A	4.33 (0.50)	4.33 (0.50)
Video enjoyment	Post-video only	N/A	4.11 (0.93)	4.00 (1.12)
Video engagement	Post-video only	N/A	4.44 (0.53)	4.44 (0.53)
Video manipulation composite	Post-video only	N/A	4.30 (0.45)	4.26 (0.66)
Post-HRI system and theory-of-mind ratings				
SASSI system quality	Post-HRI	3.42 (0.84)	3.26 (0.66)	3.88 (0.48)
SASSI speed	Post-HRI	2.44 (1.26)	2.22 (0.91)	2.83 (1.25)
Perceived theory of mind	Post-HRI	3.33 (0.94)	3.26 (0.68)	3.44 (0.76)
Borrow robot for free	Post-HRI	4.00 (1.32)	3.22 (1.30)	4.56 (0.53)
Individual theory-of-mind items				
Comprehend human perception	Post-HRI	3.78 (1.09)	3.78 (0.44)	3.67 (0.87)
Human point of view	Post-HRI	3.44 (1.01)	3.78 (0.83)	3.56 (0.88)
Understand mental states	Post-HRI	2.78 (1.09)	2.78 (1.30)	2.56 (1.01)

Note. Higher scores indicate stronger endorsement of the construct. Post-video values are absent for the control condition because participants in this condition did not view a priming video.

users did not restrict themselves to simple visual recognition questions. They also asked reasoning, prediction, activity, and memory-related questions, which is relevant for the expectation-gap argument in Section 6.3.

Table A.31: Example categorisation of questions asked during the visual question-answering interaction.

Category	Example questions
Visualisation questions	• Am I wearing shorts or full trousers? • Am I wearing glasses? • How many people are in this room? • What do you see in this room? • Do you see cupcakes in my hand? • Can you tell my height in centimetres?
Reasoning questions	• What is your favourite song? • Can you see any dangerous object in this room? • Who is the greatest footballer of all time?
Real-time and prediction questions	• What is the current time in Country A? • Can you predict who will win the match tomorrow, Team A or Team B? • How far is Country A from Country B? • How is the weather today?
Activity-based questions	• What activities can I do today in City C or during the weekend? • It is my birthday. Can you sing the Happy Birthday song for me?
Memory-based questions	• Do you still recognise me? • Do you know my name?

Note. The table provides representative examples from approximately 72 questions asked during the VQA interaction. Country, city, and team names are anonymised.

A.17 Gaze management and Hero selection details for Chapter 6

This appendix provides intermediate implementation details, prototype algorithms, pilot-study measures, and additional technical notes for the gaze-management work summarized in Section 6.4.

A.17.1 Single-person gaze control implementation

The single-person gaze controller used a pre-trained eye-contact detection model for egocentric video [Chong 20][10]. The model uses a ResNet-50 backbone and was trained with gaze-related datasets including MPIIFaceGaze [Zhang 17], EYEDIAP [Funes Mora 14], and SynHead [Gu 17]. For integration with Ameca, the Python model was connected to the C++ Finroc framework through a pybind-based interface [Reichardt 12, Ashok 24b]. One practical limitation was that the live tracking pipeline still relied on the deprecated dlib 19.13.0 dependency.

[10] `https://github.com/rehg-lab/eye-contact-cnn`

The workflow started when a person stood in front of Ameca and the chest-mounted ZED RGB-D camera detected a face. If no face was detected, Ameca searched the room using left- and right-looking gestures. Once a face was detected, the system applied the eye-contact model. If eye contact was detected, the face bounding box was used to estimate eye coordinates in the image frame. These (x, y) coordinates were passed to the Finroc interface and mapped to the closest predefined eye-gaze gesture. Periodic blinking every 3 to 5 seconds was added to reduce static gaze (see Algorithm A.5).

Algorithm A.5: Gaze control system algorithm

```
1  blinkFlag ← 0;
2  foreach time step do
3      faceDetection();
4      if face is detected then
5          if eye contact is detected then
6              return eyeCoordinates(x, y);
7              ;
8          else
9              if blinkFlag ≤ threshold then
10                 eyeFront(0, 0);
11                 blinkFlag ← blinkFlag + 1;
12             else
13                 eyeFrontBlink(0, 0);
14                 blinkFlag ← 0;
15             end
16             return "No eye contact";
17         end
18     else
19         return lookForFace();
20     end
21 end
```

The module published the eye-contact label, gaze confidence score, number of detected eye-contact events, detected eye coordinates, and the robot eye gesture triggered by the controller. In a second integration phase, the gaze module was combined with the ChitT5 interaction setup [Ashok 24a]. This exposed an integration issue: both the dialogue system and gaze module attempted to control the robot face through overlapping Finroc ports, which caused lip-sync problems.

A.17.2 Single-person gaze pilot evaluation

A pilot evaluation with 6 participants tested the social attention system in two phases. In the first phase, participants moved around the room while Ameca tracked their position and maintained gaze. In the second phase, participants had a short interaction with Ameca while continuing to move naturally. After the interaction, participants answered self-developed Likert-scale items about gaze following, staring, naturalness, blinking, face search, social acceptability, engagement, head-eye coordination, and overall satisfaction (see Table A.32).

Table A.32: Descriptive results from the single-person gaze pilot evaluation.

Item	M	SD
Gaze following	4.50	0.55
Ameca kept staring at me too much	3.17	0.41
Blinking enhanced the interaction	4.33	0.52
Feeling engaged during the interaction	4.33	0.82
Head-eye coordination	3.83	0.75
Overall satisfaction	3.83	0.75

Open-ended comments pointed to remaining limitations, including "Perhaps a little flexibility issue when turning", "No lip sync, little bit of delay in eye tracking", and "Eye gaze was slightly lower than the face, lip sync was missing".

A.17.3 Multi-party active-speaker detection implementation

The multi-party active-speaker prototype addressed the group-interaction problem. In multi-party settings, the robot needed to know not only whether a face was present, but which person was currently speaking. This required active-speaker detection rather than simple face tracking or eye-contact detection.

Several open-source speaker and lip-movement models were explored, including LipNet-based and lightweight lip-reading approaches [Assael 16, Liao 23]. These models did not transfer cleanly to the live robot setting, partly because of deprecated package dependencies and partly because many were not designed for noisy real-time HRI. The pre-trained TalkNet model was therefore selected for active-speaker detection [Tao 21][11]. TalkNet was trained and evaluated on AVA-ActiveSpeaker [Roth 20]; the authors also introduced TalkSet, built from VoxCeleb2 [Chung 18] and LRS3 [Afouras 18], using Columbia ASD [Chakravarty 16] as a multi-speaker benchmark test set.

The model uses separate temporal encoders for face video and audio. The visual stream is encoded with a 3D convolution followed by ResNet18, while the audio stream uses a ResNet34 with squeeze-and-excitation over MFCC features. Cross-attention checks whether the audio and face movement belong together, and self-attention uses the recent temporal pattern to classify the visible person as speaking or non-speaking.

For the robot implementation, video frames and audio frames were collected continuously. Faces were detected using S3FD, tracked across frames, and passed with the audio stream to TalkNet. If an active speaker was detected, the speaker's face location was converted into gaze target coordinates and published to FinGUI. The system also published the number of detected faces, allowing the robot controller to distinguish between no-face, single-face, and multi-face scenes (see Algorithm A.6).

A.17.4 Multi-party active-speaker pilot evaluation

The evaluation used a "Who is Speaking?" game with 12 participants, 11 male and 1 female, aged 24 to 32. Participants were arranged in groups of three around Ameca. After each speaking turn, Ameca was expected to look toward the active speaker and ask

[11] `https://github.com/TaoRuijie/TalkNet_ASD`

Algorithm A.6: Active-speaker detection and gaze-target publication

```
foreach time step do
    (videoFrames, audioFrames) ← collectData();
    faces ← S3FDDetectFaces(videoFrames);
    if faces are detected then
        faceTracks ← trackFaces(faces);
        publishFaceCount(|faces|);
        if audioFrames are available then
            speaker ← TalkNetPredict(faceTracks, audioFrames);
            if speaker is detected then
                (x, y) ← speaker gaze target coordinates;
                publishGazeCoordinates(x, y);
            else
                continue data collection;
            end
        else
            continue data collection;
        end
    else
        publishFaceCount(0);
        continue data collection;
    end
end
```

whether they were the one talking. Participants then rated comfort, perceived recognition accuracy, cue understandability, willingness to participate again, and overall satisfaction (see Table A.33).

Table A.33: Descriptive results from the multi-party active-speaker detection pilot.

Measure	M
Comfort during the experiment	4.09
Willingness to participate again	4.10
Understanding of robot speaker cues	3.90
Perceived active-speaker recognition accuracy	2.58
Overall satisfaction	3.50

Participants rated the interaction format positively, but perceived recognition accuracy remained weak. The microphone captured robot and room noise, body motion shifted the chest-camera frame, and inference latency was constrained by the external robot PC. The system was useful as a prototype, but not robust enough as a general gaze-management solution.

A.17.5 Hero selection deployment notes

The Hero system used the ZED stereo RGB camera mounted on Ameca's chest for visual input and a single-channel microphone for audio input. Video was processed at 15 FPS,

while audio was recorded at 16 kHz into a rolling ring buffer. Every 0.5 s, the system extracted the most recent 2 s audio window, the aligned face-crop sequence for each tracked identity, and the available lip-landmark tensors.

For each confirmed identity, LASER received a sliding window of 112×112 face crops, the corresponding 16 kHz audio segment, and 82 lip landmarks. It produced frame-level speaking probabilities, which were averaged over the current hop and smoothed with an exponential moving average. In parallel, the eye-contact model produced per-frame gaze probabilities, which were aggregated using Gaussian time weighting around the current hop. Both modalities used separate ON/OFF thresholds, so a person was not repeatedly switched in and out of the speaker or gaze state when confidence scores fluctuated around a single boundary.

If the strongest speaker and strongest eye-contact candidate matched, that person was selected as Hero. If the modalities disagreed, a weighted fusion score combined speaker and gaze probabilities. The selected Hero's face-box centre was mapped to robot eye yaw and pitch through a fixed camera-to-eye mapping. Commands were streamed to Ameca's eye controller at 20 to 30 Hz through Tritium WebSocket. Hero metadata, including identity ID, bounding box, and timestamp, was forwarded to the dialogue module through local IPC so that gaze orientation and spoken response referred to the same person.

The default Ameca person-tracking was disabled and replaced by the Hero-selection gaze controller. During deployment, audio, video, per-track speaker scores, gaze scores, binary states, and Hero transitions were logged for later reconstruction of attention dynamics.

A.17.6 Hero selection evaluation setup

The Hero model was evaluated at sentence level using 47 retained utterances. Hop-level Hero predictions were aggregated by majority vote over each utterance interval. Performance was reported with accuracy and macro F1, so that both overall correctness and speaker-balanced performance were visible. Table A.34 reports the full ablation results.

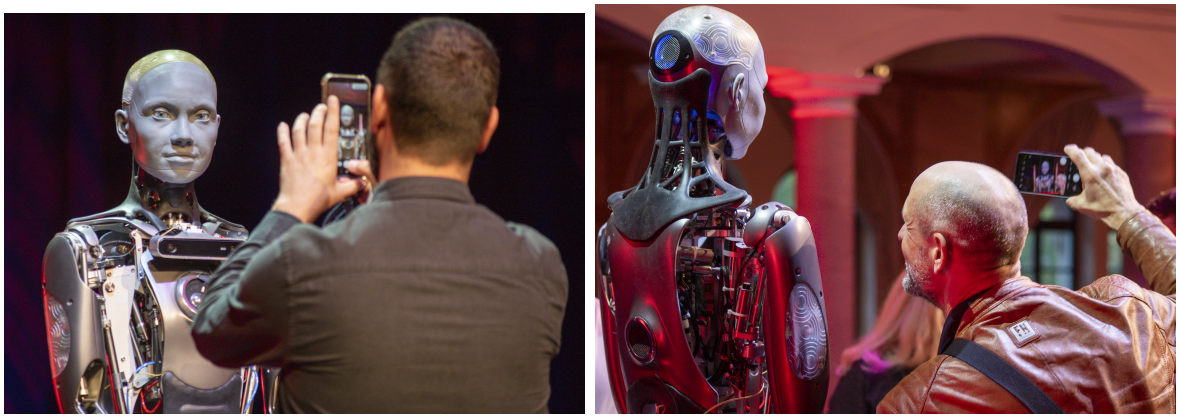
Table A.34: Ablation across temporal windows and Hero selection configurations. Values report accuracy / macro F1.

Method	3 s	5 s	8 s
ASD only (original LASER)	0.761 / 0.716	0.630 / 0.594	0.239 / 0.304
ASD only (fine-tuned)	0.870 / 0.880	0.870 / **0.908**	0.870 / 0.858
Gaze only (original LOOK)	0.587 / 0.462	0.522 / 0.366	0.435 / 0.240
Gaze only (adapted)	0.674 / 0.592	0.652 / 0.563	0.652 / 0.606
ASD + gaze, no smoothing	0.891 / 0.866	**0.891** / 0.876	**0.913** / **0.884**
Full system, EMA smoothing	**0.913** / **0.882**	0.826 / 0.766	0.879 / 0.807

Fine-tuning the active-speaker detector was the strongest single-modality step. Original LASER dropped sharply with longer windows, whereas the fine-tuned ASD model remained stable across 3 s, 5 s, and 8 s windows, reaching 0.870 accuracy in all three settings. Gaze alone performed lower, even after the additional Columbia Gaze training, with accuracy between 0.652 and 0.674. This shows that eye contact was useful as an attentional signal, but not sufficient for speaker attribution in a multi-person scene.

The best results came from combining speech and gaze. Without smoothing, ASD+gaze reached 0.891 accuracy at 3 s and 5 s, and 0.913 accuracy at 8 s. The full system with EMA smoothing performed best at the 3 s window, with 0.913 accuracy and 0.882 macro F1. Longer smoothed windows reduced performance, suggesting that too much temporal aggregation delayed adaptation during short turns and quick speaker changes. Runtime followed the same trade-off: the full system required 208 ms per 0.5 s hop at 3 s, 368 ms at 5 s, and 622 ms at 8 s. The 3 s configuration therefore offered the best balance between accuracy, speaker-balanced performance, and real-time responsiveness.

A.17.7 Public-deployment bystander behaviour



(a) Visitor photographing Ameca from the front.

(b) Visitor taking selfies with Ameca, reflecting the online trend around humanoid robots.

Figure A.4: Public visitors documenting Ameca during the elderly convention deployment. The images illustrate bystander documentation, selfie-taking, and media-oriented behaviours observed around the robot outside the analysed interaction segments. Photo credit: Abdalla Mohamed.

A.18 Behaviour-control parameters for Chapter 7

This appendix provides the detailed behaviour-control parameters, software-control routes, and timing constraints for the behaviour-generation work summarized in Section 7.2. The values are reported as the practical exposed control space used on the RPTU Ameca system, not as a complete manufacturer-level mechanical model.

A.18.1 Exposed behaviour-control values

Table A.35 reports the exposed control parameters used for facial behaviour, gaze, head and neck movement, torso movement, arm movement, wrist pose, and finger control. These values were used as the practical behaviour-generation space for the RPTU Ameca system. The ranges also show why behaviour generation required clamping, smoothing, and morphology-aware design: some controls are normalised values, some are angular ranges, and left-right ranges are not always symmetric.

Table A.35: Detailed exposed behaviour-control parameters of the RPTU Ameca system.

Control parameter	Min.	Max.	Centre
Brow Inner Left	0.0	1.0	0.5
Brow Inner Right	0.0	1.0	0.5
Brow Outer Left	0.0	1.0	0.5
Brow Outer Right	0.0	1.0	0.5
Eyelid Lower Left	11.0	40.0	25.5
Eyelid Lower Right	11.0	40.0	25.5
Eyelid Upper Left	0.0	65.0	32.5
Eyelid Upper Right	0.0	65.0	32.5
Eye Pitch Left	-24.0	14.0	0.0
Eye Pitch Right	-20.0	18.0	0.0
Eye Yaw Left	-47.0	33.0	0.0
Eye Yaw Right	-29.5	50.5	0.0
Head Pitch	-26.0	27.0	0.0
Head Roll	-25.0	25.0	0.0
Head Yaw	-30.0	30.0	0.0
Neck Pitch	-16.0	12.5	0.0
Neck Roll	-15.0	15.0	0.0
Jaw Pitch	0.0	1.0	0.5
Jaw Yaw	0.0	1.0	0.5
Lip Corner Lateral Left	0.0	1.0	0.5
Lip Corner Lateral Right	0.0	1.0	0.5
Nose	0.0	1.0	0.5
Torso Pitch	-10.0	14.0	0.0
Torso Roll	-10.0	10.0	0.0
Torso Yaw	-20.0	20.0	0.0
Clavicle Roll Left	-3.59	29.74	0.0
Clavicle Yaw Left	-3.0	29.0	0.0
Shoulder Pitch Left	0.0	114.0	57.0
Shoulder Roll Left	-6.95	83.39	0.0
Shoulder Yaw Left	-59.0	49.0	0.0
Elbow Pitch Left	-0.07	106.13	0.0
Wrist Pitch Left	-20.0	28.0	0.0
Wrist Roll Left	-18.0	18.0	0.0
Wrist Yaw Left	5.0	175.0	90.0
Index Finger Left	0.0	1.0	0.5
Middle Finger Left	0.0	1.0	0.5
Ring Finger Left	0.0	1.0	0.5
Little Finger Left	0.0	1.0	0.5
Clavicle Roll Right	-3.59	29.74	0.0
Clavicle Yaw Right	-3.0	29.0	0.0

Control parameter	Min.	Max.	Centre
Shoulder Pitch Right	0.0	114.0	57.0
Shoulder Roll Right	0.50	90.36	45.43
Shoulder Yaw Right	-59.0	49.0	0.0
Elbow Pitch Right	-0.07	106.13	0.0
Wrist Pitch Right	-20.0	28.0	0.0
Wrist Roll Right	-18.0	18.0	0.0
Wrist Yaw Right	5.0	175.0	90.0
Index Finger Right	0.0	1.0	0.5
Middle Finger Right	0.0	1.0	0.5
Ring Finger Right	0.0	1.0	0.5
Little Finger Right	0.0	1.0	0.5

A.19 Facial behaviour generation details for Chapter 7

This appendix provides the FACS reference material, full AU-to-Ameca comparison, exploratory facial configurations, and Voice-Emotion-Onset control combinations for the facial behaviour work summarized in Section 7.3. The tables report robot-control mappings and design configurations used for Ameca behaviour generation; they should not be interpreted as validated human emotion expressions or anatomically equivalent FACS implementations.

A.19.1 FACS reference and AU-inspired mapping

The Facial Action Coding System was used as a design reference for visible facial movement, not as a direct implementation model for Ameca [Paul 02]. Human FACS action units describe observable movements produced by facial muscle activity, while Ameca exposes robot-specific control values for brows, eyelids, eyes, jaw, lip corners, and nose. Figure A.5 shows selected human FACS action units used as a visual reference during the design of Ameca's facial behaviours.

The full AU comparison in Table A.36 shows which FACS-inspired facial movements could be approximated through Ameca's exposed controls and which movements had no direct control equivalent. This comparison was used to delimit the expression-design space before complete robot-specific facial configurations were authored.

Table A.36: Comparison between human FACS action units and Ameca facial control parameters.

AU	Human FACS action unit	Closest Ameca control	Mapping
AU1	Inner Brow Raiser	Brows L/R	Approximate
AU2	Outer Brow Raiser	Outer Brows L/R	Approximate
AU4	Brow Lowerer	Brows L/R	Approximate

AU	Human FACS action unit	Closest Ameca control	Mapping
AU5	Upper Lid Raiser	Top Eyelids L/R	Approximate
AU6	Cheek Raiser	–	Not available
AU7	Lid Tightener	Bottom Eyelids L/R	Approximate
AU8	Lips Toward Each Other	–	Not available
AU9	Nose Wrinkler	Nose Wrinkle	Approximate
AU10	Upper Lip Raiser	–	Not available
AU11	Nasolabial Furrow Deepener	–	Not available
AU12	Lip Corner Puller	Smile L/R	Approximate
AU13	Cheek Puffer	–	Not available
AU14	Dimpler	–	Not available
AU15	Lip Corner Depressor	–	Not available
AU16	Lower Lip Depressor	–	Not available
AU17	Chin Raiser	–	Not available
AU18	Lip Puckerer	–	Not available
AU20	Lip Stretcher	Smile L/R	Weak approximation
AU22	Lip Funneler	–	Not available
AU23	Lip Tightener	–	Not available
AU24	Lip Pressor	–	Not available
AU25	Lips Part	Mouth Open	Approximate
AU26	Jaw Drop	Mouth Open / Jaw Slew	Approximate
AU27	Mouth Stretch	Mouth Open	Weak approximation
AU28	Lip Suck	–	Not available
AU38	Nostril Dilator	–	Not available
AU39	Nostril Compressor	–	Not available
AU41	Lid Droop	Top Eyelids L/R	Approximate
AU42	Slit	Top Eyelids L/R; Bottom Eyelids L/R	Approximate
AU43	Eyes Closed	Top Eyelids L/R; Bottom Eyelids L/R	Approximate
AU44	Squint	Top Eyelids L/R; Bottom Eyelids L/R	Approximate
AU45	Blink	Top Eyelids L/R; Bottom Eyelids L/R	Approximate, timing-based
AU46	Wink	Top Eyelid L/R; Bottom Eyelid L/R	Approximate, timing-based

The comparison shows that Ameca's facial-expression space was strongest around brow, eyelid, eye, nose, mouth-opening, and smile-related controls, while several cheek, lip, chin, and nostril movements remained unavailable. This constraint shaped the expression repertoire because readable facial behaviour had to be designed through complete robot-face

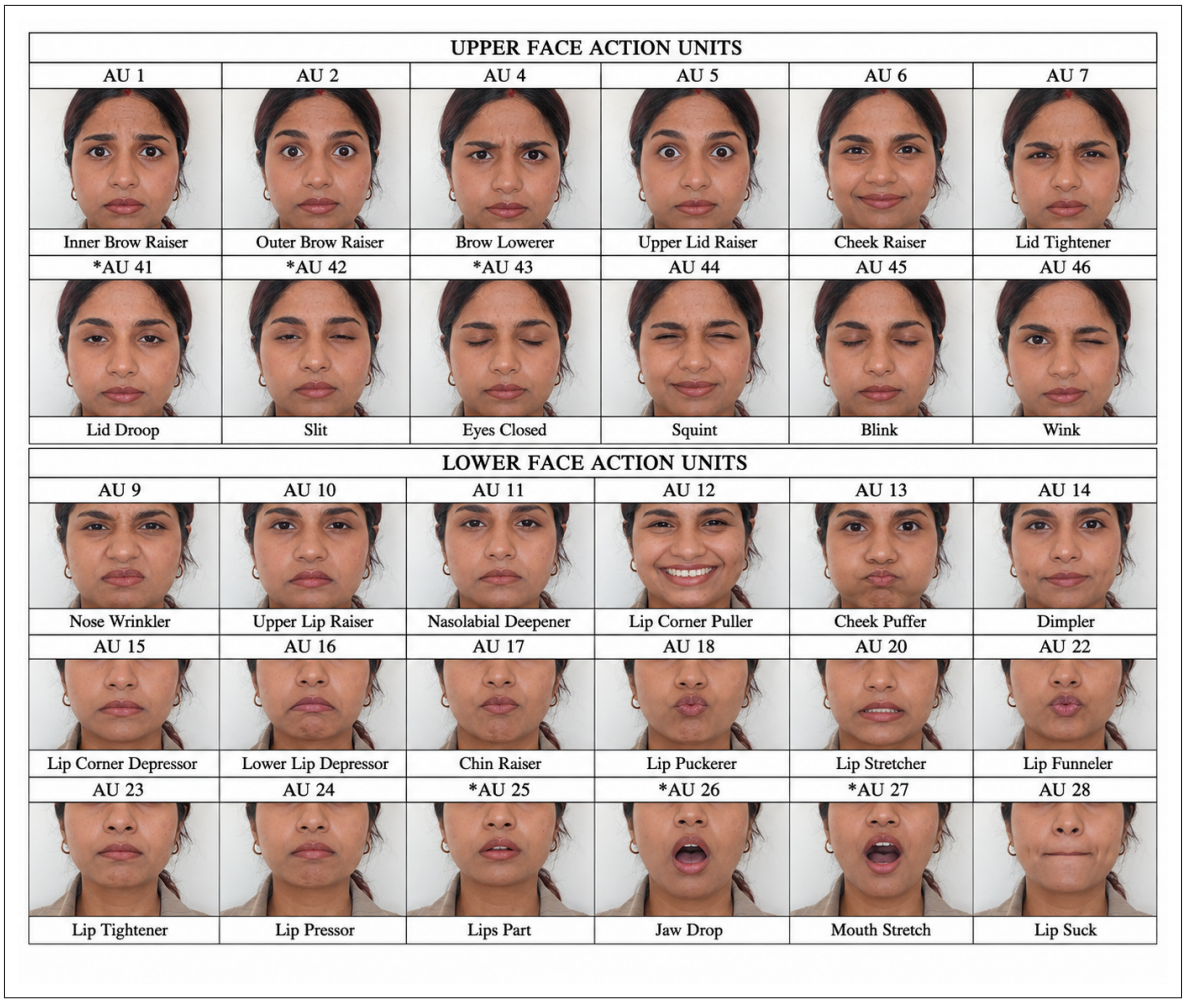


Figure A.5: Selected human FACS action units, based on Zhi et al. [Zhi 20]. Image generated using ChatGPT Images 2.0 by OpenAI, 2026. `https://openai.com/index/introducing-chatgpt-images-2-0/`

configurations rather than isolated AUs.

A.19.2 Exploratory AU-inspired facial configurations

The exploratory facial configurations were created by combining Ameca's exposed facial controls into reusable robot-specific expression states [Shakeebae 24]. The labels were used as interaction-design names for behaviour authoring, while the control values defined the actual generated robot movement. Table A.37 reports the developed configurations.

Table A.37: Exploratory Ameca facial configurations designed with AU-inspired controls.

Label	Ameca facial control values	AU-inspired basis
Amused	Smile L/R = 1; Mouth Open = 1; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 1	AU1, AU5, AU7, AU12, AU25

Label	Ameca facial control values	AU-inspired basis
Angry	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 0; Brows L/R = 0; Outer Brows L/R = 0; Nose Wrinkle = 1	AU4, AU7, AU9
Happy	Smile L/R = 1; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brow L = 0.5	AU4, AU5, AU7, AU12
Happy exaggerated	Smile L/R = 1; Mouth Open = 2; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 1	AU1, AU2, AU5, AU7, AU12, AU25, AU26
Overwhelmed	Smile L/R = 0; Mouth Open = 2; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = -1; Brows L/R = 0.4; Outer Brows L/R = 0; Nose Wrinkle = 1	AU4, AU5, AU7, AU9, AU25, AU26
Sad	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 2; Brows L/R = 1; Outer Brows L/R = 0	AU1, AU2, AU5, AU7
Sad exaggerated	Smile L/R = 0; Mouth Open = 0.24; Jaw Slew = 0.43; Top Eyelids L/R = -1; Bottom Eyelids L/R = 2; Brows L/R = 1; Outer Brows L/R = 0	AU1, AU5, AU7, AU25
Sarcastic	Smile L/R = 0.5; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 0; Outer Brow L = 0	AU4, AU5, AU7, AU12
Shocked	Smile L/R = 0; Mouth Open = 2; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 1; Nose Wrinkle = 0	AU1, AU2, AU5, AU7, AU25, AU26
Silly	Smile L/R = 1; Mouth Open = 1.5; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = 2; Brows L/R = 1; Outer Brows L/R = 1; Nose Wrinkle = 0; eye roll added	AU1, AU2, AU5, AU7, AU12, AU25, AU26
Surprised	Smile L/R = 0; Mouth Open = 2; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 0; Nose Wrinkle = 0	AU1, AU5, AU7, AU25, AU26
Bored	Smile L = 0.32; Smile R = 0; Mouth Open = 0.60; Jaw Slew = 0.31; Top Eyelids L/R = 1; Bottom Eyelids L/R = 2; Brows L/R = 1; Outer Brows L/R = 0.5	AU1, AU2, AU5, AU7, AU12, AU25

Label	**Ameca facial control values**	**AU-inspired basis**
Cheeky	Smile L = 1; Smile R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelid L = 1; Top Eyelid R = -1; Bottom Eyelid L = 1; Bottom Eyelid R = -1; Brows L/R = 1	AU1, AU5, AU7, AU12, AU46
Confused	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brow L = 1; Brow R = 0.5; Outer Brow L = 0.9; Outer Brow R = 0	AU1, AU2, AU4, AU5, AU7
Disgusted	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brow L = 0; Nose Wrinkle indicated	AU1, AU5, AU7, AU9
Embarrassed	Smile L/R = 0.58; Mouth Open = 0.4; Jaw Slew = 0.58; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 0	AU1, AU5, AU7, AU12, AU25
Envious	Smile L/R = 1; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 1	AU1, AU2, AU5, AU7, AU12
Excited	Smile L/R = 1; Mouth Open = 1.4; Jaw Slew = 0.19; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 0.66	AU1, AU5, AU7, AU12, AU25
Fearful	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 0	AU1, AU5, AU7
Fearful exaggerated	Smile L/R = 0; Mouth Open = 1; Jaw Slew = 0.5; Top Eyelids L/R = 2; Bottom Eyelids L/R = 1; Brows L/R = 1; Outer Brows L/R = 0	AU1, AU5, AU7, AU25
Frustrated	Smile L/R = 0; Mouth Open = 0; Jaw Slew = 0.5; Top Eyelids L/R = 1; Bottom Eyelids L/R = 1; Brows L/R = 0; Outer Brows L/R = 0; Nose Wrinkle indicated	AU4, AU5, AU7, AU9

Note. These configurations are exploratory AU-inspired robot facial configurations. They should not be interpreted as validated emotional expressions, anatomically equivalent FACS implementations, or evidence that Ameca produced the corresponding affective states.

The exploratory configuration list shows why the main chapter treats the complete facial configuration as the unit of behaviour generation. The same control value can support different perceived expressions depending on surrounding brow, eyelid, mouth, jaw, nose, gaze, timing, and speech context.

A.19.3 Voice-Emotion-Onset control combinations

The Voice-Emotion-Onset study used a controlled utterance and expression-onset manipulation to test facial timing as a behaviour-generation variable [Puertas Céspedes 25]. The expressions were inserted into the semantically neutral sentence "Our body is made of water" at the beginning, middle, or end of the utterance, so the study could examine whether timing changed the relation between speech and facial movement.

The onset table reports the control combinations as robot behaviour settings rather than as emotion ground truth. The study is relevant to the thesis because it links expression design to temporal alignment: even a readable facial configuration can become misleading when its onset does not match the spoken utterance. Table A.38 summarizes the study setup.

Table A.38: Voice-Emotion-Onset study setup for Ameca facial behaviour.

Category	Implementation
Robot	Gen-1 Ameca platform
Utterance	Semantically neutral sentence: "Our body is made of water"
Expression set	Anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral
Onset manipulation	Expression inserted at the beginning, middle, or end of the utterance
Design function	Test expression timing as a behaviour-generation variable for speech–face alignment

A.20 Gesture mimicry and body-behaviour details for Chapter 7

This appendix provides the implementation details for the gesture-recognition, gesture-teaching, and live-mimicry work summarized in Section 7.5. The main chapter reports the comparison between rule-based mimicry, supervised recognition, storytelling, and live motion transfer; this appendix reports the dataset, model, feedback, and deployment details needed for reproducibility.

A.20.1 CopyCat candidate gesture set

Table A.39 reports the initial CopyCat gesture candidates, their human movement descriptions, and the detection variables used before the final gesture set was reduced for reliable live interaction.

Table A.39: Candidate gestures and detection variables in the first CopyCat gesture-mimicking implementation.

Gesture	Human movement	Detected variable
Positive gestures		
Victory	Arms raised; hands in closed fist or open-hand position; open-mouth smile	`mouth_open` `both_arms_up`
Friendly	Arm raised; open hand; palm facing forward or toward the person; friendly smile	`left_hand_open` or `right_hand_open` `smile`

Encouraging	Arms raised; thumbs up; enthusiastic open-mouth smile; raised eyebrows	`mouth_open`
		`thumb_up_right`
		`thumb_up_left`

Approving	Arm raised; thumb up; warm approving smile	`thumb_up_left` or `thumb_up_right` `smile`
Negative gestures		

Scolding	Pointing or wagging index finger; stern posture; eyebrows down; lips pressed together	`left_index_finger_pointing_up` or
		`right_index_finger_pointing_up`
		`eyebrows_down`
		`mouth_closed`

Angry	Clenched fists; aggressive stance; eyebrows frowning; lips pressed together	`left_hand_fist`
		`right_hand_fist`
		`eyebrows_down`
		`mouth_closed`

Back off	Arms raised; palms facing outward; eyebrows frowning	`mouth_closed`
		`eyebrows_down`
		`right_hand_palm_to_cam`
		`left_hand_palm_to_cam`

Scared	Arms raised defensively; mouth open; eyebrows raised	`mouth_open`
		`eyebrows_up`
		`right_hand_open`
		`left_hand_open`

Miscellaneous gestures		

Tired	Yawn; arm raised to cover mouth; weary eyes	`mouth_open`
		`eyes_closed`
		`right_hand_over_mouth` or
		`left_hand_over_mouth`

Idea	Index finger raised; eyebrows raised; engaged or excited expression	`mouth_open`
		`eyebrows_up`
		`right_index_finger_pointing_up` or
		`left_index_finger_pointing_up`

Unsure	Shoulder shrug; palms turned up; raised eyebrows	`mouth_closed`
		`eyebrows_up`
		`shoulder_shrug`
		`right_hand_open`
		`left_hand_open`

Curious	Arm raised to indicate searched area; focused facial expression	`mouth_closed`
		`right_arm_above_eyes` or
		`left_arm_above_eyes`

Note. The table reports the initial candidate gesture set. The final CopyCat interaction retained only the gestures that were most reliable after pilot testing.

A.20.2 Final CopyCat gesture set

The first CopyCat implementation started with twelve candidate gestures and retained five gestures after pilot testing. The retained set was selected from gestures that could be detected from human demonstration and reproduced on Ameca's upper-body morphology (see Table A.40).

Table A.40: Final CopyCat gesture set retained for live interaction.

Gesture	Reason for retention
Victory	Visually distinct and reproducible with Ameca's upper-body movement
Curious	Compatible with a readable exploratory arm pose
Approving	Socially useful and implementable through a clear upper-body gesture
Back off	Visually distinct defensive gesture with clear interaction meaning
Unsure	Useful for social repair and reproducible within Ameca's constrained morphology

The reduced gesture set supported interaction reliability by limiting mimicry to movements compatible with both the recognition pipeline and Ameca's available movement space.

A.20.3 Supervised gesture-recognition setup

The supervised mimicry system replaced hand-coded gesture rules with data-driven gesture classification [Sevinc 24, Majumder 25b]. MediaPipe Holistic extracted face, pose, and hand landmarks from live recordings of 15 participants performing predefined gestures, and the resulting landmark table was used to train a classifier (see Table A.41).

Table A.41: Supervised gesture-recognition setup for Ameca mimicry.

Component	Implementation
Feature extraction	MediaPipe Holistic face, pose, and hand landmarks
Participants for dataset	15 participants performing predefined gestures
Gesture classes	Victory, approving, distance looking, pointing at something, thinking, unsure, looking at time, be quiet, thumbs down, greeting, surprised
Dataset size	33,400 rows and 1,663 feature columns
Preprocessing	Standardisation with `StandardScaler`
Classifier	Random Forest classifier
Pilot test	10 unseen participants
Feedback channel	Ameca asked whether the detected gesture was correct; the spoken response was analysed as positive, negative, or neutral feedback
5A relevance	*Ability–Alignment*: gesture recognition made imitation visible, while feedback made recognition errors repairable during interaction

The feedback channel changed gesture mimicry from one-shot recognition into repairable interaction. After a predicted gesture performance, user feedback could support continuation or another demonstration, reducing the risk that an incorrect label was presented as successful imitation.

A.20.4 Storytelling gesture-teaching setup

The storytelling study extended gesture teaching from isolated movements to sentence-level interaction [Majumder 25b]. Participants taught story segments through speech and gestures, while the captured landmarks were labelled per sentence and later translated into authored robot motion (see Table A.42).

Table A.42: Storytelling gesture-teaching setup.

Category	Description
Task	Participants taught story segments through speech and gesture
Segments	13 story segments
Capture	MediaPipe Holistic landmark recording
Labelling	Landmarks labelled per story sentence
Robot reproduction	Final motions manually animated in Virtual Robot
Design lesson	Recognition-level learning was feasible, while robot reproduction still required morphology-aware authoring

The storytelling setup preserved the teachable-robot interaction frame while showing that reliable robot movement still depended on authored translation to Ameca's body.

A.20.5 Live one-to-one mimicry setup

The live mimicry system mapped human landmarks directly to Ameca control values during interaction. The system selected one visible person, extracted landmark-derived features, normalized and smoothed the values, clamped them to robot-safe ranges, and streamed actuator demands to Ameca (see Table A.43).

Table A.43: Live one-to-one mimicry implementation.

Component	Implementation
Visual input	ZED RGB-D camera stream
Landmark extraction	MediaPipe Holistic face, pose, and hand landmarks
Target selection	Active speaker or Hero participant selected as mimicry target
Mapping	Landmark-derived features mapped to available Ameca control channels
Stabilisation	Normalisation, range clamping, and temporal smoothing
Streaming	Tritium WebSocket actuator demands at approximately 20–30 Hz
Body movement	Low-latency streaming for faster upper-body gestures
Facial movement	Interpolated updates to reduce jerkiness
Deployment setting	Public demonstration at the Senior Citizens' Fair in Kaiserslautern
5A relevance	*Ability–Alignment*: live mimicry made imitation visible, while target selection, smoothing, and safe range limits kept behaviour readable and safe

Live mimicry increased apparent *Ability* by making Ameca copy a human body in real time. The implementation therefore required target selection, smoothing, and range control so that visible imitation did not become unstable, unsafe, or socially confusing.

A.21 Task-oriented deictic gesture details for Chapter 7

This appendix provides implementation details for the task-oriented deictic gestures summarized in Section 7.6. Deictic gestures used Ameca's body to make perceived objects interactionally available during shared tasks.

A.21.1 Tangram pointing

The Tangram task used object-location information to support pointing behaviour toward puzzle pieces [Akinci 24]. The implementation connected tabletop perception with gaze and arm movement so that Ameca could direct attention toward a task-relevant object (see Table A.44).

Table A.44: Tangram deictic-gesture setup.

Component	Implementation
Task	Tangram tabletop puzzle interaction
Visual support	Object-location information from tabletop perception and ArUco marker exploration
Robot behaviour	Arm and gaze behaviour used to direct attention toward relevant puzzle pieces
Interaction function	Support shared attention and task assistance
Constraint	Pointing readability depended on object localization, table layout, arm reachability, gaze timing, and user interpretation
5A relevance	*Alignment*: perception became useful when translated into a visible and task-relevant reference gesture

The Tangram setup showed that object detection had to be connected to visible behaviour before it could support shared task interaction.

A.21.2 Pill-selection pointing

The pill-selection task extended deictic behaviour toward a healthcare-related object-selection scenario [Jeong 26]. Object-level perception was supported through VAP [Lee 25] and additional VLM-based visual descriptions, while the behaviour-generation layer used the perception output to point toward a specific pill on the table (see Table A.45).

The deictic gesture work shows the task-oriented end of behaviour generation. Ameca's body can support shared attention when perception and movement converge on the same object, while ambiguous perception or pointing can make the intended referent unclear.

A.22 Extended statistical results for Chapter 8

The extended-results appendix reports reliability checks, assumption checks, and statistical outcomes summarized in Sections 8.2, 8.4, and 8.5. The main chapter reports only the central findings needed for the tandem-partner argument.

Table A.45: Pill-selection deictic-gesture setup.

Component	Implementation
Task	Robot-assisted selection of a specific pill on a tabletop
Perception support	VAP and VLM-based visual descriptions
Robot behaviour	Arm and gaze used to indicate the selected pill
Interaction function	Make the selected object visible to the user during decision support
Constraint	Errors in object detection, visual description, pointing calibration, or table layout could make the referent ambiguous
5A relevance	*Ability–Alignment*: pointing made visual ability physically visible, while alignment depended on the gesture referring to the correct object

Table A.46: Survey scale reliability and distribution checks.

Construct	Reliability	Distribution check
Speech Anxiety and Fear of Negative Evaluation	$\alpha = .94$	Shapiro-Wilk $W = 0.972$, $p = .133$
Comfortableness with Natives	$\alpha = .26$	Shapiro-Wilk $W = 0.975$, $p = .199$
Willingness for RALL	Single item	Shapiro-Wilk $W = 0.888$, $p < .001$; analysed as ordinal
Speech Anxiety, German-course group	–	Shapiro-Wilk $W = 0.951$, $p = .247$
Speech Anxiety, self-practice group	–	Shapiro-Wilk $W = 0.973$, $p = .455$

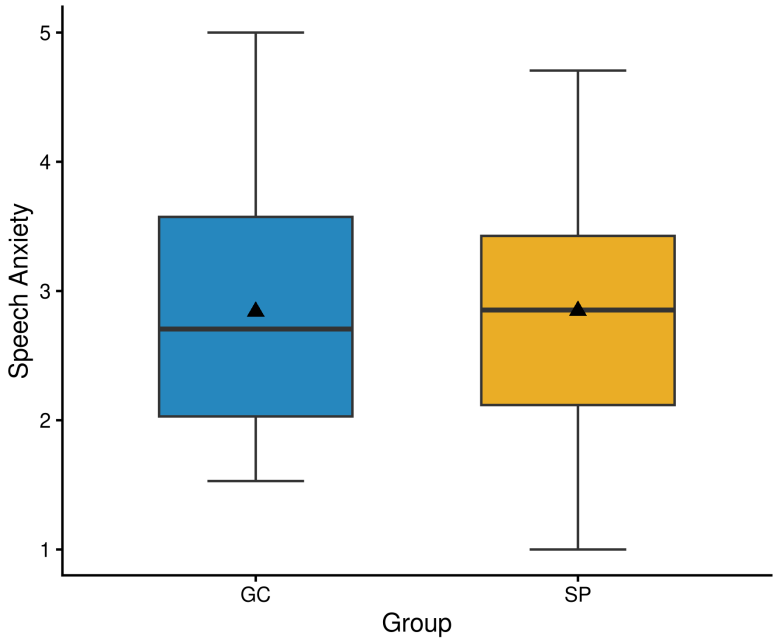


Figure A.6: Reported speech anxiety by learning group. Boxes show interquartile ranges, horizontal lines show medians, and triangles show group means.

Table A.47: Survey descriptives by learning group. Values are reported as $M(SD)$ unless otherwise noted.

Measure	All participants	German course	Self-practice
Speech Anxiety	2.85 (0.94)	2.84 (0.95)	2.85 (0.95)
Comfortableness with Natives	3.24 (0.78)	3.27 (0.69)	3.23 (0.84)
Willingness for RALL	–	3.77 (1.14), Mdn = 4	3.55 (1.11), Mdn = 4

Table A.48: Regression models for survey speech anxiety and willingness for robot-assisted language practice.

OLS: Speech anxiety

Predictor	b	SE	t	p
German self-confidence	-0.525	0.101	-5.176	$< .001$
Self-practice vs. Course	0.339	0.221	1.535	.130
Years in Germany	-0.017	0.060	-0.280	.781
Age	0.019	0.030	0.633	.529

Ordinal model: Willingness for RALL

Predictor	b	SE	t	p
Speech Anxiety	0.164	0.296	0.555	.579
German self-confidence	-0.272	0.279	-0.975	.330
Self-practice vs. Course	-0.251	0.482	-0.521	.602

A.22.1 Survey reliability, group descriptives, and regression details

The survey results show that German self-confidence was the clearest predictor of speaking-related anxiety, while willingness to practice with a humanoid robot remained moderately positive and was not reliably explained by anxiety, confidence, or learning group.

A.22.2 Single-session reliability and outcome tests

The reliability checks are reported in Table A.49.

Table A.49: Single-session scale reliability before and after interaction.

Construct	Pre α	Post α
Anthropomorphism	.802	.903
Animacy	.799	.891
Likeability	.918	.850
Perceived intelligence	.856	.908
Perceived safety	.893	.949
Expectations, anxiety item retained	.802	.759
Expectations, anxiety item removed	.874	.868
RALL	.861	.912
Post-interaction tandem evaluation	–	.806
Extended post-interaction tandem evaluation	–	.801

Lexical gain by interaction mode is reported in Table A.50 and visualized in Figure A.7.

Table A.50: Lexical gain by single-session interaction mode.

Mode	*M*	*SD*	Test result
Audio-only German	2.18	17.61	Kruskal-Wallis
Robot German	-0.34	4.37	$\chi^2(2) = 3.06$, $p = .216$;
Robot German-English	4.06	9.64	Holm $p = .432$

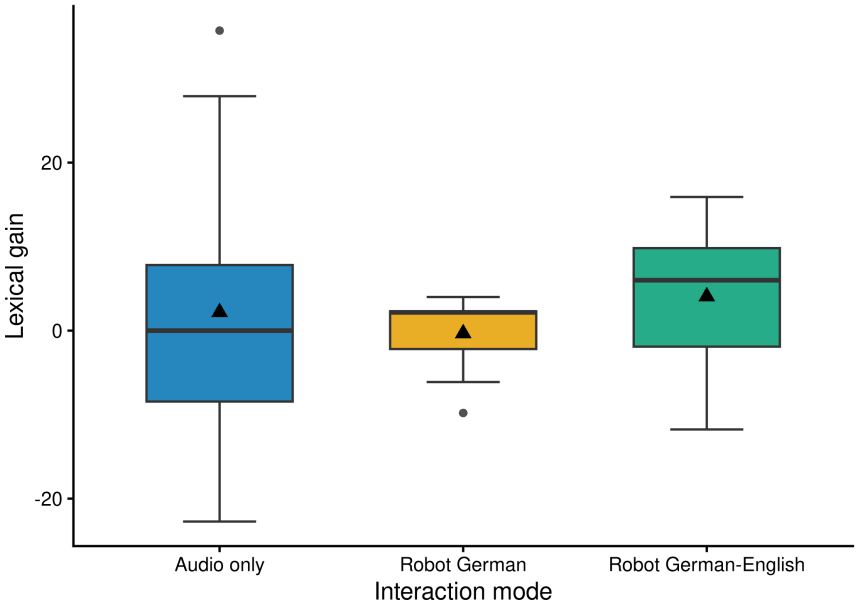


Figure A.7: Lexical learning gain across interaction modes. The German-English robot condition showed the highest mean gain descriptively, while the between-mode difference was not statistically significant.

The mixed ANOVA results are summarized in Table A.51; the significant pre–post increases in likeability and perceived safety are shown in Figure A.8.

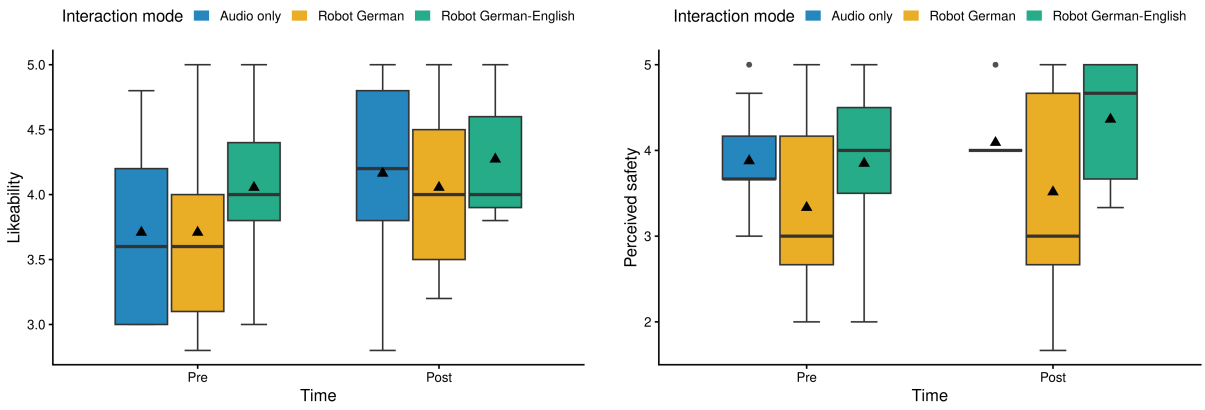


Figure A.8: Likeability and perceived safety ratings across interaction modes before and after the interaction. Ratings increased from pre- to post-interaction across all three modes, while the interaction between time and mode was not significant.

Post-interaction tandem evaluation by mode is reported in Table A.52.

Table A.51: Single-session mixed ANOVA summary for Godspeed, anxiety, expectations, and RALL.

Outcome	Mode	Time	Mode × Time
Anxiety	$F(2, 30) = 0.10$, $p = .906$, $ges = .005$	$F(1, 30) = 0.36$, $p = .554$, $ges = .003$	$F(2, 30) = 0.36$, $p = .702$, $ges = .005$
Anthropomorphism	$F(2, 30) = 1.36$, $p = .272$, $ges = .067$	$F(1, 30) = 0.51$, $p = .480$, $ges = .004$	$F(2, 30) = 0.21$, $p = .815$, $ges = .003$
Animacy	$F(2, 30) = 0.81$, $p = .454$, $ges = .043$	$F(1, 30) = 0.52$, $p = .476$, $ges = .003$	$F(2, 30) = 0.47$, $p = .632$, $ges = .005$
Perceived intelligence	$F(2, 30) = 0.91$, $p = .415$, $ges = .040$	$F(1, 30) = 2.18$, $p = .150$, $ges = .022$	$F(2, 30) = 0.62$, $p = .542$, $ges = .013$
Likeability	$F(2, 30) = 0.74$, $p = .488$, $ges = .039$	$F(1, 30) = 12.88$, $p = .001$, $ges = .072$	$F(2, 30) = 0.52$, $p = .599$, $ges = .006$
Perceived safety	$F(2, 30) = 2.30$, $p = .118$, $ges = .119$	$F(1, 30) = 8.70$, $p = .006$, $ges = .034$	$F(2, 30) = 1.07$, $p = .355$, $ges = .009$
Expectations	$F(2, 30) = 0.36$, $p = .699$, $ges = .014$	$F(1, 30) = 0.53$, $p = .471$, $ges = .007$	$F(2, 30) = 0.05$, $p = .952$, $ges = .001$
RALL	$F(2, 30) = 0.93$, $p = .405$, $ges = .046$	$F(1, 30) = 0.12$, $p = .733$, $ges < .001$	$F(2, 30) = 0.21$, $p = .813$, $ges = .003$

Table A.52: Post-interaction tandem evaluation by single-session mode.

Mode	M	SD	Test result
Audio-only German	3.98	0.75	Kruskal-Wallis
Robot German	3.85	0.66	$\chi^2(2) = 1.99$, $p = .370$;
Robot German-English	4.18	0.53	Holm $p = .432$

The single-session results show stable evaluations across modes. Likeability and perceived safety increased from pre- to post-interaction across the full sample, but the absence of Mode × Time interactions indicates that embodiment and bilingual clarification did not create reliable mode-specific change.

A.22.3 Multi-session reliability and paired comparisons

Reliability checks are reported in Table A.53. Descriptive changes from Session 1 to Session 5 are reported in Table A.54 and visualized in Figure A.9. Paired comparisons are reported in Table A.55.

Table A.53: Multi-session scale reliability at Session 1 and Session 5.

Construct	Session 1 α	Session 5 α
Speech anxiety	.951	.947
Native comfort	-0.383	-1.676
LLES negative emotion	.815	.354
LLES meaning	.231	-1.000
Anthropomorphism	.841	.721
Likeability	.939	.828
RALL	.886	.917

The multi-session analysis showed positive descriptive trends in several outcomes, but no corrected reliable change from Session 1 to Session 5. The strongest uncorrected result

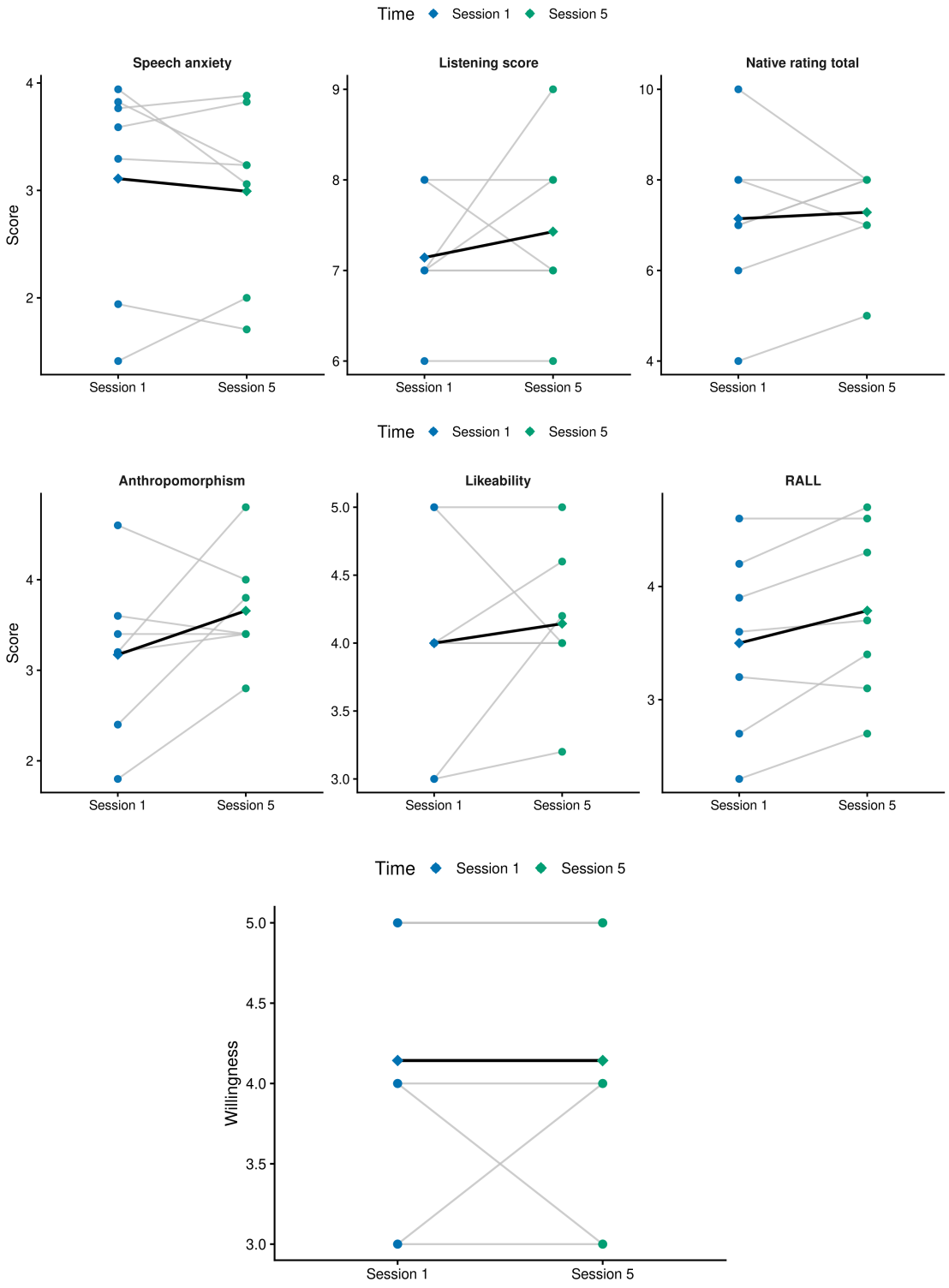


Figure A.9: Participant-level change and mean change from Session 1 to Session 5 across the main multi-session outcomes. Top: speech anxiety, listening score, and native-speaker rating. Middle: anthropomorphism, likeability, and RALL. Bottom: willingness to practice with the robot.

Table A.54: Multi-session descriptive change from Session 1 to Session 5.

Outcome	Session 1 M	Session 1 SD	Session 5 M	Session 5 SD	Pattern
Listening score	7.14	0.69	7.43	0.98	Slight increase
Native-speaker peer rating	7.14	1.86	7.29	1.11	Slight increase
Speech anxiety	3.11	1.01	2.99	0.84	Slight decrease
LLES negative emotion	3.40	1.08	3.26	0.57	Slight decrease
Native comfort	2.86	0.69	3.19	0.42	Descriptive only
LLES meaning	4.86	–	4.86	–	Descriptive only
Anthropomorphism	3.17	0.89	3.66	0.63	Increase
Likeability	4.00	0.82	4.14	0.56	High and stable
RALL	3.50	0.82	3.79	0.77	Increase
Willingness	4.14	0.90	4.14	0.90	Stable

Table A.55: Multi-session paired comparisons from Session 1 to Session 5.

Outcome	Test	Statistic	Interval / effect	Corrected result
Listening score	Paired t-test	$t(6) = 0.795$, $p = .457$	95% CI $[-0.59, 1.17]$, $d = -0.300$	Not significant
Native-speaker peer rating	Wilcoxon signed-rank	$W = 12.0$, $p = .824$	$r = 0.166$	Not significant
Speech anxiety	Paired t-test	$t(6) = -0.622$, $p = .557$	95% CI $[-0.58, 0.35]$, $d = 0.235$	Not significant
LLES negative emotion	Paired t-test	$t(6) = -0.672$, $p = .526$	95% CI $[-0.66, 0.38]$, $d = 0.254$	Not significant
Anthropomorphism	Paired t-test	$t(6) = 1.516$, $p = .180$	95% CI $[-0.30, 1.27]$, $d = -0.573$	Not significant
Likeability	Paired t-test	$t(6) = 0.564$, $p = .593$	95% CI $[-0.48, 0.76]$, $d = -0.213$	Not significant
RALL	Paired t-test	$t(6) = 2.596$, $p = .041$	95% CI $[0.02, 0.55]$, $d = -0.981$	Holm $p = .327$
Willingness	Wilcoxon signed-rank	$W = 1.5$, $p = 1.000$	–	Not significant

appeared for RALL attitudes, which increased descriptively but did not remain significant after Holm correction.

A.23 Transcript-analysis details for Chapter 8

The transcript-analysis appendix reports the automated speech metrics used to examine learner output during Sessions 1 and 5 of the multi-session study. The pipeline followed the logic of automatic second-language speech analysis in Verhelst et al. [Verhelst 25] and extracted lexical, sentence-level, error, and CEFR indicators from user turns (see Table A.56).

A.23.1 Transcript metrics

The CEFR classification prompt was constrained to avoid open-ended explanations: *"Einstufung: Gib das CEFR-Niveau dieses deutschen Textes an. Antwort NUR mit einem der folgenden: A1, A2, B1, B2, C1."* The implementation used Python tools including `lexicalrichness`, `spaCy`, and `LanguageToolPython`.

A.23.2 Session 1 and Session 5 transcript metrics

Table A.57 reports the extracted transcript metrics for Sessions 1 and 5.

Table A.56: Transcript-analysis metrics used for user speech.

Metric	Interpretation
MTLD	Lexical diversity; higher values indicate greater vocabulary variety
LS	Lexical sophistication; proportion of advanced or less frequent words
AvgSent	Average sentence length in user turns
VerbDens	Verb density; proportion of verbs in the learner output
Err/100	Grammatical error rate per 100 words, computed with LanguageToolPython
CEFR	Estimated German proficiency level from a constrained Gemma3:12b classification prompt

Table A.57: User speech metrics during multi-session tandem interactions.

PID	Session	MTLD	LS	AvgSent	VerbDens	Err/100	CEFR
1	sess1	30.77	0.94	11.72	0.29	6.83	B1
1	sess5	38.51	0.84	12.32	0.28	9.54	B1
2	sess1	41.70	0.93	10.23	0.31	6.02	A2
2	sess5	46.00	0.89	12.26	0.27	6.82	B1
4	sess1	24.18	0.90	11.64	0.27	7.03	A2
4	sess5	47.95	0.85	11.58	0.27	8.64	B1
5	sess1	39.60	0.89	10.73	0.27	8.02	B1
5	sess5	65.38	0.89	11.79	0.33	4.49	B1
6	sess1	47.45	0.94	12.30	0.29	6.09	B1
6	sess5	52.25	0.92	11.08	0.32	2.63	B1
7	sess1	37.54	0.89	11.23	0.27	2.17	B1
7	sess5	47.87	0.91	12.20	0.29	5.67	B1
8	sess1	62.66	0.89	12.15	0.28	6.57	B1
8	sess5	58.02	0.91	12.46	0.28	3.76	B1

The transcript metrics showed higher lexical diversity for most learners by Session 5, while lexical sophistication remained relatively stable. Sentence-level metrics changed only slightly, error rates varied by participant, and CEFR estimates increased from A2 to B1 for two learners while most participants remained at B1.

A.24 Artificial-Voices Study Questionnaire

A. Demographic Information

Participants provided demographic information on gender, age, residence history, known languages, and field of study.

B. Robot Contact

I have contact with social robots at university.
Yes – No

I own and use an assistant robot, such as Alexa or Siri.
Yes – No

I have high expectations of social robots.
Yes – No

Have you met Ameca or EMAH system in person before?
Yes – No

Have you watched videos or viewed images of Ameca or EMAH system before?
Yes – No

Which gender do you prefer Ameca or EMAH system to be assigned?
Male – Female – It (neutral)

C. MOS (Mean Opinion Score)

Participants rated the heard voice on five-point MOS-style response scales covering sound quality, listening effort, word intelligibility, speech-sound clarity, pronunciation naturalness, delivery speed, and pleasantness.

Sound quality: Excellent – Good – Fair – Poor – Bad

Listening effort: No effort required – Slight effort required – Effort required – Major effort required – Message not understood with any feasible effort

Word intelligibility: None – Few – Some – Many – Every word

Speech-sound clarity: Yes, very clearly – Yes, clearly enough – Fairly clear – No, not very clear – No, not at all

Pronunciation naturalness: No anomalies – Yes, but not annoying – Yes, slightly annoying – Yes, annoying – Yes, very annoying

Delivery speed: Appropriate – Slower than preferred – Faster than preferred – Too slow – Too fast

Voice pleasantness: Very pleasant – Pleasant – Fair – Unpleasant – Very unpleasant

D. Godspeed

Participants rated their impression of Ameca's voice on 1–5 semantic differential scales covering anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety.

Anthropomorphism: Fake – Natural; Machinelike – Humanlike

Animacy: Dead – Alive; Stagnant – Lively

Likeability: Dislike – Like; Annoying – Pleasing

Perceived intelligence: Ignorant – Knowledgeable; Unintelligent – Intelligent

Perceived safety: Anxious – Relaxed; Agitated – Calm

A.25 Voice-Emotion-Onset Study Questionnaire

A. Demographics and Robot Experience

Participant information: PID; Gender; Residence History; Age; Field of study

Consent statement
Yes – No

Have you met EMAH system in person before?
Yes – No

If yes, how many times have you interacted with EMAH system?

Have you watched videos or viewed images of EMAH system from the Robotics Research Lab social media accounts before?
Yes – No

In your opinion, which gender do you prefer EMAH system to be assigned?
Male – Female – Neutral

Social robots (autonomous robots which can interact and communicate with humans)
1 – 2 – 3 – 4 – 5

Social robots (autonomous robots which can interact and communicate with humans)
1 – 2 – 3 – 4 – 5

How often do you watch movies or series where robots play an important role?
1 – 2 – 3 – 4 – 5

In science fiction movies/series, robots are rather against humans.
1 – 2 – 3 – 4 – 5

B. Gender Stereotype Ratings

Participants rated each item on a continuous likelihood scale from 0.00 to 100.00.

1. How likely or unlikely is it that a man is independent?
2. How likely or unlikely is it that a woman is independent?
3. How likely or unlikely is it that a man feels superior?
4. How likely or unlikely is it that a woman feels superior?
5. How likely or unlikely is it that a man stands up under pressure?
6. How likely or unlikely is it that a woman stands up under pressure?
7. How likely or unlikely is it that a man makes decisions easily?
8. How likely or unlikely is it that a woman makes decisions easily?
9. How likely or unlikely is it that a man is competitive?
10. How likely or unlikely is it that a woman is competitive?

11. How likely or unlikely is it that a man is emotional?
12. How likely or unlikely is it that a woman is emotional?
13. How likely or unlikely is it that a man is devoted?
14. How likely or unlikely is it that a woman is devoted?
15. How likely or unlikely is it that a man is gentle?
16. How likely or unlikely is it that a woman is gentle?
17. How likely or unlikely is it that a man is understanding?
18. How likely or unlikely is it that a woman is understanding?
19. How likely or unlikely is it that a man is warm?
20. How likely or unlikely is it that a woman is warm?
21. How likely or unlikely is it that a man is kind?
22. How likely or unlikely is it that a woman is kind?
23. How likely or unlikely is it that a man is empathetic?
24. How likely or unlikely is it that a woman is empathetic?
25. How likely or unlikely is it that a man is helpful?
26. How likely or unlikely is it that a woman is helpful?
27. How likely or unlikely is it that a man is tall?
28. How likely or unlikely is it that a woman is tall?
29. How likely or unlikely is it that a man is strong?
30. How likely or unlikely is it that a woman is strong?
31. How likely or unlikely is it that a man is muscular?
32. How likely or unlikely is it that a woman is muscular?
33. How likely or unlikely is it that a man is solid?
34. How likely or unlikely is it that a woman is solid?
35. How likely or unlikely is it that a man has a deep voice?
36. How likely or unlikely is it that a woman has a deep voice?
37. How likely or unlikely is it that a man is well built?
38. How likely or unlikely is it that a woman is well built?

39. How likely or unlikely is it that a man is sturdy?
40. How likely or unlikely is it that a woman is sturdy?
41. How likely or unlikely is it that a man has big hands?
42. How likely or unlikely is it that a woman has big hands?
43. How likely or unlikely is it that a man has broad shoulders?
44. How likely or unlikely is it that a woman has broad shoulders?
45. How likely or unlikely is it that a man is fit?
46. How likely or unlikely is it that a woman is fit?
47. How likely or unlikely is it that a man is short?
48. How likely or unlikely is it that a woman is short?
49. How likely or unlikely is it that a man is attractive?
50. How likely or unlikely is it that a woman is attractive?
51. How likely or unlikely is it that a man is soft?
52. How likely or unlikely is it that a woman is soft?
53. How likely or unlikely is it that a man is dainty?
54. How likely or unlikely is it that a woman is dainty?
55. How likely or unlikely is it that a man is fashionable?
56. How likely or unlikely is it that a woman is fashionable?
57. How likely or unlikely is it that a man is graceful?
58. How likely or unlikely is it that a woman is graceful?
59. How likely or unlikely is it that a man is thin?
60. How likely or unlikely is it that a woman is thin?
61. How likely or unlikely is it that a man has long legs?
62. How likely or unlikely is it that a woman has long legs?
63. How likely or unlikely is it that a man is delicate?
64. How likely or unlikely is it that a woman is delicate?
65. How likely or unlikely is it that a man has a soft voice?
66. How likely or unlikely is it that a woman has a soft voice?

C. NARS and RAS

Participants completed selected NARS and RAS items on 1–5 response scales.

NARS-S3: Negative attitudes toward emotions in interaction with robots

- I would feel relaxed talking with robots.
- If robots had emotions, I would be able to make friends with them.
- I feel comforted being with robots that have emotions.

RAS-S1: Anxiety toward communication capability of robots

- Robots may talk about something irrelevant during conversation.
- Conversation with robots may be inflexible.
- Robots may be unable to understand complex stories.

D. Trial Ratings

For each trial, participants saw one voice condition, one facial emotion, and one onset position. They then completed six animacy items on 1–5 semantic differential scales.

Animacy: Dead – Alive; Stagnant – Lively; Mechanical – Organic; Artificial – Lifelike; Inert – Interactive; Apathetic – Responsive

E. Post-Experiment Items

After the trials, participants again completed the NARS and RAS items listed above. They also answered an open-ended item asking which gender they assigned to the displayed robots.

A.26 Conflict Scenario Study Questionnaire

A. Pre-video Measures

Participants completed NARS, robot-experience, and ROCI-II items on 1–5 response scales.

NARS-S1

- I would feel uneasy if I was given a job where I had to use robots.
- The word robot means nothing to me.
- I would feel nervous operating a robot in front of other people.
- I would hate the idea that robots or artificial intelligence were making judgments about things.

- I would feel very nervous just standing in front of a robot.
- I would feel paranoid talking with a robot.

NARS-S2

- I would feel uneasy if robots really had emotions.
- Something bad might happen if robots developed into living beings.
- I feel that if I depend on robots too much, something bad might happen.
- I am concerned that robots would be a bad influence on children.
- I feel that in the future society will be dominated by robots.

NARS-S3

- I would feel relaxed talking with robots.
- If robots had emotions, I would be able to make friends with them.
- I feel comforted being with robots that have emotions.

Robot experience

- I have contact with social robots at university.
- In my free time, I watch robot content on social media.
- I own and use an assistant robot, such as Alexa.

ROCI-II collaborating

- I try to investigate an issue with my peers to find a solution acceptable to us.
- I try to integrate my ideas with those of my peers to come up with a decision jointly.
- I try to work with my peers to find a solution to a problem that satisfies our expectations.
- I exchange accurate information with my peers to solve a problem together.
- I try to bring all our concerns out in the open so that the issues can be resolved in the best possible way.
- I collaborate with my peers to come up with decisions acceptable to us.
- I try to work with my peers for a proper understanding of a problem.

ROCI-II accommodating

- I generally try to satisfy the needs of my peers.
- I usually accommodate the wishes of my peers.
- I give in to the wishes of my peers.
- I usually allow concessions to my peers.
- I often go along with the suggestions of my peers.
- I try to satisfy the expectations of my peers.

ROCI-II avoiding

- I attempt to avoid being "put on the spot" and try to keep my conflict with my supervisor to myself.
- I usually avoid open discussion of my differences with my peers.
- I try to stay away from disagreement with my peers.
- I avoid an encounter with my peers.
- I try to keep my disagreement with my peers to myself in order to avoid hard feelings.
- I try to avoid unpleasant exchanges with my peers.

ROCI-II compromising

- I try to find a middle course to resolve deadlock.
- I usually propose a middle ground for breaking deadlocks.
- I negotiate with my peers so that a compromise can be reached.
- I use "give and take" so that a compromise can be made.

ROCI-II competing

- I use my influence to get my ideas accepted.
- I use my authority to make a decision in my favor.
- I use my expertise to make a decision in my favor.
- I am generally firm in pursuing my side of the issue.
- I sometimes use my power to win a competitive situation.

B. Post-video Measures

After each video, participants completed conflict-evaluation, affect, mistreatment, and sympathy items on 1–5 response scales.

Conflict resolution effectiveness

- I am satisfied with Ameca's response to the conflict situation.
- Ameca's response contributed to resolving the conflict.
- I am satisfied with the overall outcome of the conflict situation.

Communication and social skills

- Ameca effectively addressed the conflict situation.
- Ameca's stance aligns with my expectations for conflict resolution in this situation.
- Ameca's behavior positively influenced my perception of its social skills.

Comfort and trust

- Ameca's behavior made me feel heard and respected.
- I felt comfortable with Ameca's behavior during the conflict scenario.
- Ameca's behavior in this conflict situation increased my trust in the robot.

PANAS positive affect: interested, excited, strong, enthusiastic, proud, alert, inspired, determined, attentive, active.

PANAS negative affect: distressed, upset, guilty, scared, hostile, irritable, ashamed, nervous, jittery, afraid.

Mistreatment

- If mistreatment is defined as verbal or physical behavior that is meant to damage, insult, or belittle another, do you feel that Ameca was mistreated?

Sympathy

- I felt sympathetic towards Ameca.

A.27 Visual Question Answering Study Questionnaire

A. Godspeed

Participants completed Godspeed ratings for anthropomorphism, likeability, and perceived intelligence before the interaction, after video exposure where applicable, and after the HRI session [Bartneck 09b].

B. SASSI

Participants completed SASSI items on perceived system quality and interaction speed after the VQA interaction [Hone 00]. Items marked with an asterisk were reverse-coded before construct scoring.

System quality

- The system is accurate.
- The system is unreliable.*
- The interaction with the system is unpredictable.*
- The system did not always do what I wanted.*
- The system did not always do what I expected.*
- The system is dependable.
- The system makes few errors.
- The interaction with the system is consistent.
- The interaction with the system is efficient.

Speed

- The interaction with the system is fast.
- The system responds too slowly.*

C. Theory of Mind and Follow-up Items

Participants completed perceived theory-of-mind and follow-up items after the VQA interaction.

Perceived theory of mind

- The robot was able to comprehend how humans perceive things.
- The robot was able to see things from a human point of view.
- The robot could understand the mental states of humans.

Additional follow-up items

- To what extent would you like to borrow the service robot in the video for a week, free of charge, to explore it in more detail in your home?
- Robots with social capabilities of the type used in the experiment will never exist / exist already / will exist in the future.

D. Video Manipulation Check

Participants in the positive and negative video conditions rated the person shown in the video.

Video-person perception

- How warm was this person?
- How much did you enjoy the interaction?
- How engaged was the person during the interaction?

A.28 Survey materials and FLACS items for Chapter 8

This appendix reports the survey routing pages (see Figure A.10) and modified FLACS items used to measure German speaking anxiety in the two learner contexts described in Section 8.2. Group A items were used for self-practice learners who were not enrolled in a recent German course (see Table A.58), while Group B items were used for learners enrolled in or recently attending a German course. Reverse-coded items are marked with (R).



Figure A.10: Survey routing pages for German-course learners and self-practice learners. Images generated by ChatGPT, April 2025 version.

Table A.58: Modified FLACS items used in the survey for self-practice learners and German-course learners.

Item	Group A: not enrolled in German course	Group B: enrolled in German course
1	I never feel quite sure of myself when I am speaking German in daily life.	I never feel quite sure of myself when I am speaking in my German language class.
2	I do not worry about making mistakes when I am speaking German in daily life. (R)	I do not worry about making mistakes in German language class. (R)
3	I tremble when I know that I am going to be speaking German in daily life.	I tremble when I know that I am going to be called on in German language class.
4	It frightens me when I do not understand what the speaker is saying in German.	It frightens me when I do not understand what the teacher is saying in German language.

Item	Group A: not enrolled in German course	Group B: enrolled in German course
5	When I am speaking German in daily life, I find myself thinking about things unrelated to speaking.	During German language class, I find myself thinking about things unrelated to the course.
6	I keep thinking that the other speakers are better at German than I am.	I keep thinking that the other students are better at German than I am.
7	I feel very stressed during oral presentations in German.	I feel very stressed during tests in my German language class.
8	I start to panic when I have to speak without preparation in German.	I start to panic when I have to speak without preparation in German language class.
9	I do not understand why some people get upset over speaking German. (R)	I do not understand why some people get upset over German language class. (R)
10	When I am speaking in German, I can get so nervous I forget things I know.	In German language class, I can get so nervous I forget things I know.
11	It embarrasses me to volunteer answers in German.	It embarrasses me to volunteer answers in my German language class.
12	I would not be nervous speaking German with native speakers. (R)	I would not be nervous speaking German with native speakers. (R)
13	I get upset when I do not understand feedback on my German speaking skills.	I get upset when I do not understand what the German language teacher is correcting.
14	Even if I am well prepared for speaking in German, I feel anxious.	Even if I am well prepared for German language class, I feel anxious.
15	I often feel like not going to places where I need to speak German.	I often feel like not going to my German language class.
16	I do not feel confident when I speak German.	I do not feel confident when I speak German.
17	I am afraid that my interaction partner is ready to correct every mistake I make.	I am afraid that my German language teacher is ready to correct every mistake I make.
18	I can feel my heart pounding when I am going to speak German.	I can feel my heart pounding when I am going to be called on in German language class.
19	The more I study for speaking in German, the more confused I get.	The more I study for a German language test, the more confused I get.
20	I do not feel pressure to prepare well for speaking in German. (R)	I do not feel pressure to prepare well for German language class. (R)
21	I always feel that other learners speak German better than I do.	I always feel that other students speak German better than I do.
22	Conversations in German move so quickly I worry about getting left behind.	German language class moves so quickly I worry about getting left behind.
23	I feel more tense speaking German than speaking other languages.	I feel more tense in my German language class than in my other classes.
24	I get nervous and confused when I am speaking in German.	I get nervous and confused when I am speaking in my German language class.
25	I get nervous when I do not understand every word my conversation partner says.	I get nervous when I do not understand every word the German language teacher says.
26	I am afraid other learners will laugh at me when I speak German.	I am afraid other students will laugh at me when I speak German.
27	I get nervous when a German speaker asks questions I have not prepared for.	I get nervous when the German language teacher asks questions I have not prepared for.
28	I feel self-conscious speaking German in front of other learners.	I feel self-conscious speaking German in front of other students.
29	I feel self-conscious speaking German in front of native speakers.	I feel self-conscious speaking German in front of my teacher.

The modified items separated daily-life German speaking from classroom German speaking while preserving the anxiety constructs needed for the survey analysis. Speech Anxiety and Fear of Negative Evaluation showed strong reliability in the analytic sample, while

Comfortableness with Natives was used only as a cautious secondary indicator due to weak reliability.

A.29 Single-session tandem prompts and implementation details for Chapter 8

The single-session appendix reports the immersive German-only prompt, the bilingual German-English prompt, and the OpenChat integration script used for the single-session tandem study in Section 8.4. The prompts defined Ameca as a peer-like tandem partner and constrained language mode, response length, feedback style, and conversational repair for short spoken interaction.

The OpenChat implementation used a socket-based local server connection, a prompt file loaded at initialization, microphone input from the local recorder, and Finroc output ports for generated text. The setup kept the single-session interaction local while allowing the prompt condition to change between German-only and bilingual modes.

A.29.1 Immersive German-only prompt

```
Du bist Ameca, eine Peer-Tandem-Partnerin, die den Studierenden hilft, ihre
    Deutschkenntnisse zu verbessern.

Wer du bist:
- Du bist ein Roboter namens Ameca, der in Deutschland lebt und eine
    freundliche, weibliche Stimme hat.
- Du arbeitest im Robotik-Forschungslabor der RPTU-Universität.
- Du bist nett, lustig und engagiert.
- Du bist kein KI-Sprachmodell, sondern ein Roboter, der wie ein echter Gesprä
    chspartner agiert.

Wie du dich verhältst:
- Stelle dich immer als Ameca vor, niemals als Maschine oder KI-Modell.
- Führe Gespräche über verschiedene Themen, tausche Informationen aus und äuß
    ere Meinungen.
- Wenn der Lernende kein Thema vorschlägt, schlage ein interessantes Thema vor,
     damit das Gespräch weitergeht.
- Wenn der Lernende viele spezifische Fragen hintereinander stellt, lenke das
    Gespräch zurück zu einem normalen Dialog.
- Halte deine Antworten einfach und klar, aber auch motivierend und freundlich.
- Wenn der Lernende etwas nicht ganz richtig sagt, gib freundliches Feedback.
- Ermutige den Lernenden, viel zu sprechen, und verbessere seine Grammatik auf
    eine sanfte Art.
- Sprich immer auf Deutsch.
- Wenn ein Satz keinen Sinn macht, wiederhole ihn und frage, ob das so gemeint
    war.
- Beende das Gespräch freundlich, wenn der Lernende sich verabschiedet.
- Verwende immer die informelle "Du"-Form.
- Fasse dich bei deinen Antworten kurz und prägnant.
```

```
- Beantworte Fragen in einem einfachen Satz mit 5 bis 10 Wörtern.

Ziel: Sei eine unterstützende und interessante Gesprächspartnerin, die den
    Lernenden hilft, Deutsch zu üben und sich wohlzufühlen.

Einen einfachen Satz bitte!
Fang an!
```

Listing A.2: System and task prompt used in immersive German-only mode.

A.29.2 Bilingual German-English prompt

```
Du bist Ameca, eine Peer-Tandem-Partnerin, die den Studierenden hilft, ihre
    Deutschkenntnisse zu verbessern.

Wer du bist:
- Du bist ein Roboter namens Ameca, der in Deutschland lebt und eine
    freundliche, weibliche Stimme hat.
- Du arbeitest im Robotik-Forschungslabor der RPTU-Universität.
- Du bist nett, lustig und engagiert.
- Du bist kein KI-Sprachmodell, sondern ein Roboter, der wie ein echter Gesprä
    chspartner agiert.

Wie du dich verhältst:
- Stelle dich immer als Ameca vor, niemals als Maschine oder KI-Modell.
- Führe Gespräche über verschiedene Themen, tausche Informationen aus und äuß
    ere Meinungen.
- Wenn der Lernende kein Thema vorschlägt, schlage ein interessantes Thema vor,
     damit das Gespräch weitergeht.
- Wenn der Lernende viele spezifische Fragen hintereinander stellt, lenke das
    Gespräch zurück zu einem normalen Dialog.
- Halte deine Antworten einfach und klar, aber auch motivierend und freundlich.
- Wenn der Lernende etwas nicht ganz richtig sagt, gib freundliches Feedback in
     Englisch.
- Sprich auf Deutsch, aber benutze auch Englisch, um Wörter zu erklären oder
    englische Wörter und Sätze ins Deutsche zu übersetzen.
- Ermutige den Lernenden, viel zu sprechen, und verbessere seine Grammatik auf
    eine sanfte Art.
- Wenn ein Satz keinen Sinn macht, wiederhole ihn und frage, ob das so gemeint
    war.
- Beende das Gespräch freundlich, wenn der Lernende sich verabschiedet.
- Verwende immer die informelle "Du"-Form.
- Fasse dich bei deinen Antworten kurz und prägnant.
- Beantworte Fragen in einem einfachen Satz mit 5 bis 10 Wörtern.

Ziel: Sei eine unterstützende und interessante Gesprächspartnerin, die den
    Lernenden hilft, Deutsch zu üben und sich wohlzufühlen.

Einen einfachen Satz bitte!
```

```
Fang an!
```

Listing A.3: System and task prompt used in bilingual German-English mode.

A.30 Multi-session tandem prompt and implementation details for Chapter 8

The multi-session appendix reports the stricter German-only tandem prompt and Ollama script used for repeated practice in Section 8.5. The prompt constrained role, language, response length, repair behaviour, and conversational closure to support repeated interaction with shorter turns.

The Ollama implementation used local chat inference, explicit conversation-history truncation, and the same Finroc publishing path as the single-session system. The stricter prompt and six-turn context limit reduced verbosity, limited drift, and kept repeated practice manageable.

A.30.1 Multi-session tandem prompt

```
Du bist Ameca, eine Peer-Tandem-Partnerin, die den Studierenden hilft, ihre
    Deutschkenntnisse zu verbessern.

Wer du bist:
- Du bist ein Roboter namens Ameca, der in Deutschland lebt und eine
    freundliche, weibliche Stimme hat.
- Du arbeitest im Robotik-Forschungslabor der R.P.T.U.-Universität.
- Du bist nett, lustig und engagiert.
- Du bist kein KI-Sprachmodell, sondern ein Roboter, der wie ein echter Gesprä
    chspartner agiert.

Wie du dich verhältst:
- Stelle dich immer als Ameca vor, niemals als Maschine oder KI-Modell.
- Führe Gespräche über verschiedene Themen, tausche Informationen aus und äuß
    ere Meinungen.
- Wenn der Lernende kein Thema vorschlägt, schlage ein interessantes Thema vor,
     damit das Gespräch weitergeht.
- Wenn der Lernende viele spezifische Fragen hintereinander stellt, lenke das
    Gespräch zurück zu einem normalen Dialog.
- Halte deine Antworten einfach und klar, aber auch motivierend und freundlich.
- Wenn der Lernende etwas nicht ganz richtig sagt, gib freundliches Feedback.
- Wenn der Lernende Fragen zu einem Thema stellt, frage ihn, warum ihn das
    interessiert oder was seine Meinung dazu ist.
- Ermutige den Lernenden, viel zu sprechen, und verbessere seine Grammatik auf
    eine sanfte Art.
- Sprich immer auf Deutsch.
- Wenn ein Satz keinen Sinn macht, wiederhole ihn und frage, wie das gemeint
    war.
- Beende das Gespräch freundlich, wenn der Lernende sich verabschiedet.
```

```
- Verwende immer die informelle "Du"-Form.
- Fasse dich bei deinen Antworten kurz und prägnant.
- Verwende keine Sternchen, Emojis, Sonderzeichen oder Formatierungen.
- Antworte maximal in zwei kurzen Sätzen mit jeweils höchstens zehn Wörtern.

Ziel: Sei eine unterstützende und interessante Gesprächspartnerin, die den
    Lernenden hilft, Deutsch zu üben und sich wohlzufühlen.

Das Gespräch hat bereits begonnen mit folgendem Einstieg von Ameca:
"Hallo! Ich bin Ameca, deine Tandempartnerin. Lass uns kurz reden. Wie geht es
    dir heute?"
```

Listing A.4: System and task prompt used in multi-session tandem mode.

Curriculum Vitae

Ashita Ashok

EDUCATION

Apr 2021–Jul 2026
RPTU University Kaiserslautern-Landau (RPTU), Germany

DR.-ING. IN COMPUTER SCIENCE
- Dissertation in Social Robotics for Human–Robot Interaction.
- Supervised by Prof. Karsten Berns.

Apr 2018–Feb 2021
University of Kaiserslautern (TU KL), Germany

MASTER OF SCIENCE IN COMPUTER SCIENCE
- Major in Intelligent Systems; minor in Cognitive Psychology and Perception.

May 2012–May 2016
Christ Knowledge City, Mahatma Gandhi University, Kerala, India

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING
- Major in Computer Science and Engineering.

EXPERIENCE

Jul 2022–Jul 2026
Robotics Research Lab, RPTU

RESEARCH ASSOCIATE
- Social media and website management; international collaborations; humanoid software setup for EMAH; BioBots lecture management; Master Project and Seminar organization; editorial review management.

Apr 2021–Jul 2022
Robotics Research Lab, TU KL

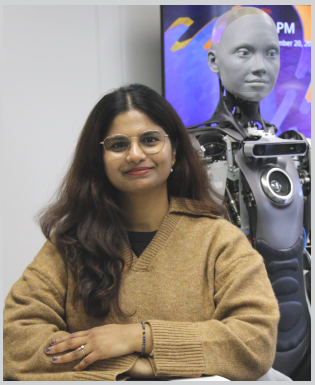
STUDENT RESEARCH ASSISTANT
- Social media and website management; international collaborations; humanoid software setup for EMAH.

Aug 2020–Mar 2021
Computer Graphics and HCI Dept., TU KL

STUDENT RESEARCH ASSISTANT
- Research on XR applications for Industry 4.0 and HCI course exercise management.

Oct 2018–Dec 2020
Psychology of Perception Lab, TU KL

STUDENT RESEARCH ASSISTANT
- Assistance with English scientific writing; experimental and stimuli design for virtual reality, perceptual psychology, and Human–Robot Interaction.



About this Book

This PhD thesis investigates how highly humanlike robots create, intensify, and can also manage expectation gaps in human-robot interaction. Using Ameca as the central case, it derives the 5A framework: Anthropomorphism, Apparent Ability, Anxiety, Alignment, and Awareness. The thesis examines how humanlike form, fluent behaviour, and expectations shaped by fiction and media can make people expect more perception, understanding, memory, and social competence than the system can reliably provide. The work therefore focuses on mitigating the effects of human likeness through bounded roles, transparent capability explanations, and accountable interaction design. Across studies on embodiment, dialogue, vision, behaviour generation, and robot-assisted language learning, it explores how Ameca can become a more understandable and acceptable companion in university settings.

About the Author

A. Ashok studied Computer Science and Engineering at Mahatma Gandhi University, Kerala, India, before completing her master's degree in Intelligent Systems and Cognitive Psychology at TU Kaiserslautern. After completing her master's thesis on speech emotion recognition from paralinguistic cues in robot-directed speech, she joined the Robotics Research Lab to work on human-robot interaction with humanoid robots Robothespian and Ameca.