



HAL
open science

How Do Robots Become “Social Robots”? An Empirical Specification of the (Non-)Emergence of Robots as Social Agents

Damien Rudaz

► **To cite this version:**

Damien Rudaz. How Do Robots Become “Social Robots”? An Empirical Specification of the (Non-)Emergence of Robots as Social Agents. Sociology. Institut Polytechnique de Paris, 2024. English. NNT : 2024IPPAT025 . tel-04884570

HAL Id: tel-04884570

<https://theses.hal.science/tel-04884570v1>

Submitted on 13 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

How Do Robots Become “Social Robots”? An Empirical Specification of the (Non-)Emergence of Robots as Social Agents

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 : Ecole Doctorale de l'Institut Polytechnique de
Paris (ED IP Paris)
Spécialité de doctorat : Sciences sociales et Management

Thèse présentée et soutenue à Palaiseau, le 12 septembre 2024, par

Damien Rudaz

Composition du Jury :

Heike Baldauf-Quilliatre Professeure, Université Lyon 2 Lumière (– UMR 5191)	Présidente/Examinatrice
Karola Pitsch Professeure, University of Duisburg-Essen	Rapporteuse
Mathias Broth Professeur, Linköping University	Rapporteur
Nicolas Sabouret Professeur, Université Paris-Saclay (– UMR 9015)	Examineur
Stuart Reeves Maitre de conférences, University of Nottingham	Examineur
Chloé Mondémé Chargée de recherche, CNRS (– UMR5206)	Examinatrice
Christian Licoppe Professeur, Télécom Paris (– UMR 9217)	Directeur de thèse

How Do Robots Become “Social Robots”?
**An Empirical Specification of the (Non-)Emergence
of Robots as Social Agents**

Thesis submitted for the degree of Doctor of Philosophy

by

Damien Rudaz

Télécom Paris

Ecole Doctorale de l'Institut Polytechnique de Paris

Supervised by Christian Licoppe

September 12, 2024

ABSTRACT

As opposed to viewing the inner workings of human-robot interactions (HRI) as black boxes, this work investigates the finely tuned micro-interactive practices through which a robot emerges as a “social agent” in different settings. Using the micro-analytic approach of Ethnomethodological Conversation Analysis (EMCA), it examines several large corpora of encounters between humans and the humanoid robot Pepper. These corpora were collected in various yet relatively typical natural and experimental settings for human-robot interactions: a museum, an office building, and a laboratory. Their exploration allows us to broaden the list of documented interactional processes occurring during human-robot encounters (indexed to specific settings, sequential contexts, spatial configurations, etc.) by which a robot can be said to be, momentarily and locally, treated as an “agent” in a social interaction. In the end, attending to the moment-by-moment production of the robot’s status as a practical accomplishment leads our inquiry to a respecification of the interactional work commonly glossed by the lay use of the terms “social robot”.

However, rather than merely a quietist attempt at clarifying conceptual mix-ups, our approach responds to design, ergonomic, and user experience (UX) concerns regarding “social” robots. That is, by attending to the locally organized practices taking place in human-robot encounters, we attempt to provide a different type of explanation as to “what went wrong” or “what went right” in an interaction with a robot: explanations based on the features made relevant by the participants themselves *as they are practically immersed within the urgency of these ongoing human-robot interactions*. By paying close attention to this moment-by-moment fine-tuning of participants’ conduct, we confront some regularly observable shortcomings in current “social” robots’ situated encounters.

Critically, this work was partially produced from within the company that creates the robot our participants interacted with: it results from an industrial partnership with the technology company Aldebaran (formerly named Softbank Robotics), which designs and manufactures the Pepper robot. Hence, another facet of this research is to detail the use of an Ethnomethodological and Conversational Analytic approach amid collaborations with designers, user experience researchers, and engineers in Aldebaran’s office – to “make the robot work” or to improve its design – over several years. Because the analytical stance of Ethnomethodology and Conversation Analysis is rather uncommon in technology companies, this partnership offers various examples of ineffability or incommensurability with alternative approaches (as well as examples of friction with the design and programming tools pre-adapted to these approaches) connected to underlying representations as to what is a “conversation” or a “social action”.

RÉSUMÉ

Plutôt que d'aborder les interactions humain-robot (HRI) comme des boîtes noires, ce travail étudie les subtiles pratiques micro-interactionnelles à travers lesquelles un robot émerge comme "agent social" dans différents environnements. A partir de la perspective micro-analytique de l'Ethnométhodologie et de l'Analyse Conversationnelle (EMCA), nous examinons plusieurs larges corpus d'interactions impliquant le robot humanoïde Pepper. Ces corpus ont été produits dans différents environnements, naturels et expérimentaux, typiques des interactions "humain-robot humanoïde" telles qu'elles existent à l'heure actuelle : un musée, des bureaux, et un laboratoire. Leur exploration nous permet d'étendre la liste des processus interactionnels documentés prenant place durant la rencontre entre un humain et un robot (indexés à des environnements, des contextes séquentiels et des configurations spatiales spécifiques) au travers desquels un robot est traité, localement et momentanément, comme un "agent" dans une interaction sociale. Etudier la production du statut du robot, instant par instant, comme un accomplissement pratique nous mène à une re-spécification du travail interactionnel habituellement obscurci par l'usage commun du terme de robot "social".

Ainsi, l'objectif de cette recherche n'est pas de produire un modèle abstrait d'étapes anhistoriques et trans-situationnelles à travers lesquelles un robot émerge comme agent, indépendamment de son design, de sa programmation, ou même des attentes, dispositions et objectifs des humains dans différents environnements. Au contraire, parmi la liste potentiellement infinie des propriétés d'un environnement et d'une interaction qui peuvent être perçues, discrétisées, catégorisées et verbalisées comme telles, nous mettons en lumière ce qui est, localement, traité comme pertinent par les participants (humains ou artificiels) pour faire progresser les activités qu'ils mènent en commun. Ces propriétés d'un environnement ou d'une interaction ne sont donc définies comme pertinentes qu'en référence à un contexte local. En ce sens, ce travail s'oppose à toute définition "ex ante" (c'est-à-dire avant qu'une interaction n'advienne) d'un robot – ou de tout artefact technologique – comme social. Une telle perspective évite de naturaliser, comme stable et allant de soi, le résultat de processus micro sociologiques toujours contingents : un robot n'émerge comme social qu'à travers la conduite des participants durant une interaction.

Cependant, en lieu et place d'une perspective quiétiste visant exclusivement à clarifier des incohérences conceptuelles, notre approche répond à des problématiques de design, d'ergonomie et d'expérience utilisateur (UX) vis à vis des robots "sociaux". En somme, en faisant jour sur les pratiques localement organisées prenant place dans des contextes impliquant des robots, nous tentons de produire un autre niveau d'explication quant à "ce qui a fonctionné" et "ce qui n'a pas fonctionné" dans une interaction avec un robot : des explications basées sur les propriétés de l'interaction rendues pertinentes par les participants eux-mêmes lorsqu'ils sont immergés dans l'urgence pratique de ces interactions "humain-robot". En scrutant, moment par moment, les ajustements granulaires et réflexifs des conduites des participants, nous mettons en lumière différentes catégories de problèmes régulièrement observables dans les rencontres avec des robots dits "sociaux".

Significativement, ce travail a été partiellement produit de l'intérieur de l'entreprise qui produit les robots avec lesquels nos participants ont interagi. En effet, cette étude résulte d'un partenariat industriel avec l'entreprise de technologie Aldebaran (précédemment nommée

Softbank Robotics), qui design et manufacture le robot Pepper. En conséquence, une autre facette de cette recherche est de détailler l'usage d'une perspective tirée de l'Ethnométhodologie et de l'Analyse Conversationnelle lors de projets menés en collaboration avec des designers, des chercheurs en expérience utilisateur et des ingénieurs, dans les bureaux d'Aldebaran – pour "faire marcher le robot" ou pour améliorer son design – durant plusieurs années. Parce que la position analytique de l'Ethnométhodologie et de l'Analyse Conversationnelle est relativement rare dans les entreprises de technologie, ce partenariat offre des exemples variés d'ineffabilité ou d'incommensurabilité avec des approches alternatives (ainsi que des exemples de friction avec les outils de programmation ou de design pré-adaptés à ces approches). Par le contraste qu'elles offrent, ces situations mettent en lumière des représentations sous-jacentes, d'un côté comme de l'autre, quant à ce qu'est une "conversation" (avec un robot), une "action sociale", ou même une "interaction".

CONTENTS

ACKNOWLEDGEMENTS.....	xi
TRANSCRIPTION CONVENTIONS.....	xiv
1. INTRODUCTION	1
1.1. Social robots or talking bundles of wires and plastic?	1
1.1.1. Wishful representations and real-life interactions	1
1.1.2. Human-robot interactions as black boxes.....	2
1.2. Topic of study, data, and plan.....	4
1.2.1. Topic of study.....	4
1.2.2. Corpora.....	5
1.2.3. Plan.....	6
1.3. Technological and organizational context for this work	10
1.3.1. The Pepper robot	10
1.3.2. Pepper sensors (what can be leveraged to perceive human actions)	11
1.3.3. Pepper multimodal abilities (what can be leveraged to produce social actions) ..	12
1.3.4. Programming Pepper	13
1.3.5. Organizational context: EMCA in a technology company	13
2. METHOD	15
2.1. Ethnomethodological Conversation Analysis and Artificial Agents	15
2.1.1. Ethnomethodology and Conversation Analysis	15
2.1.2. EMCA for Human-Robot Interactions	17
2.1.3. Studying the “sociality” of robots besides mental representations	17
2.2. Next-turn-proof procedure in human-robot-interaction as a “conceptual loosening” ..	18
2.2.1. Next-turn-proof procedure and membership knowledge in human-human interactions	18
2.2.2. The limited applicability of the next turn proof procedure in human-robot interactions	19
2.2.3. (The absence of) membership knowledge as a working hypothesis	22
2.2.4. Human-robot interactions and loose hermeneutics.....	25
2.2.5. The next-turn proof procedure applied to inter or intraspecies interactions.....	26
2.2.6. Remaining grounds to produce demonstrable interpretations of human-robot interactions	27
2.3. “Applied CA” and Good Old-Fashioned AI: Turning rule-based robots into expert conversationalists, a (long-documented) hopeless endeavor?.....	30
2.3.1. “Actualizing rules” in situ	30
2.3.2. Following rules and orienting to rules	32

2.3.3. What can we salvage from practical “coping” when doing rule-based design? ...	33
2.3.4. Expert conversationalists and rules	34
2.3.5. Some examples of EMCA-inspired interaction flows or conversational guidelines	35
2.3.6. “Applied CA”: Making rule-based robots into better novices?	37
3. AN EMERGENTIST DEFINITION OF SOCIAL AGENCY	39
3.1. The “social” of “social robots” as an emergent property	39
3.1.1. Social facts as accomplishments.....	39
3.1.2. Identities and roles as accomplishments	40
3.1.3. Consequences for local settings which involve “robots”	41
3.1.4. “Human”-“Robot” “Interactions”. A toned-down approach to emergence	42
3.2. The polysemy of “social agency”	43
3.2.1. Participant, member, conversational partner, social actor or social agent?.....	43
3.2.2. A minimal and empirically viable definition	44
3.3. Social agency as accountability and contingency	45
3.3.1. (A) Social agents as morally accountable (accountability as normativity)	45
3.3.2. (B) Social agents as competent participants (accountability as intelligibility)	48
3.3.3. (C) Social agents as producing contingent conducts	49
3.3.4. The last piece of our definition: contingency as a third criterion to define a “social agent”	53
3.4. Result: A summarized definition of “social agent”	55
3.4.1. An interactionist formulation	55
3.4.2. The analytical status of the category “social agents”	56
4. AN EMCA APPROACH IN A TECHNOLOGY COMPANY: INEFFABILITY AND FRICTIONS	57
4.1. Ethnomethodological conversation analysis and the corporate world	57
4.1.1. Ineffability and ethnomethodology: bug or feature?	57
4.1.2. Disjunctions with the practical constraints and theoretical backgrounds of roboticists	59
4.2. Do we need EMCA if we have common-sense? Practical mastery <i>versus</i> analytical understanding of conversation.....	59
4.3. Why use precision tools for the analysis if we must use a sledgehammer for the design?.....	61
4.4. “Losing the phenomenon”: Micro-analytic approaches and synthetic reports	64
4.4.1. EMCA and the pace of a technology company	64
4.4.2. Examples from the construction of a report	65
4.4.3. The “path of least resistance”: discarding the mutually responsive character of participants’ conduct	68
4.4.4. Losing the phenomenon.....	69

4.4.5. Slides, reports, and practical knowledge	71
4.5. “Technologies of the intellect”: Frictions with usual modes of representation of human-robot interactions	71
4.6. The agency of programming and scripting tools for creating “conversations”	75
4.6.1. Crafting disfluent speech and filled pauses (“uhh”).....	76
4.6.2. Making the robot “continue speaking” after a silence.....	79
4.6.3. The agency of technical tools	87
4.7. Divergent assumptions as to what an “interaction” is	87
4.7.1. Degree of granularity and interactional consequences	87
4.7.2. From sequentiality to “successiveness”	88
4.7.3. Composition over position: “interaction” as questions and answers.....	89
4.7.4. Are these issues relevant for the design of conversational chatbots relying on generative AIs?	90
5. FROM OBJECT TO AGENT: THE EMERGENCE OF A ROBOT AS A (GREETABLE) SOCIAL PARTICIPANT	91
5.1. Introduction.....	91
5.2. Previous work.....	92
5.2.1. The significance of greetings for human-robot interactions.....	92
5.2.2. Conditional relevance as a breaking point in the robot’s status as an interactant.....	93
5.2.3. Pre-beginning designs in Human-Robot interactions: “coming into sight” vs “coming into existence”	94
5.3. Method	96
5.3.1. Participants	96
5.3.2. Experimental setup	96
5.3.3. Instructions.....	97
5.3.4. Scenario.....	97
5.3.5. “Activation steps” achieved by the robot during the first seconds of the interaction	98
5.4. Distribution of first greeting occurrences during the experiment.....	98
5.5. Five paths to the production of a first greeting	100
5.5.1. Orienting to physical co-presence as an adequate framework for the initiation of a greeting sequence	100
5.5.2. Orienting to mutual gaze as projecting an imminent next action from the robot	102
5.5.3. Multiple greetings – Orienting to the robot’s wave as the confirmation of an ongoing greeting sequence	105
5.5.4. Orienting to the robot’s waving gesture as an upgrade of its first greeting	107
5.5.5. Absence of greetings – Orienting to the robot as an autonomous, machine-based, script	109
5.6. Discussion	112

5.6.1. Sequential ambiguity	112
5.6.2. The waving gesture as a threshold.....	113
5.6.3. Should a robot be designed to harness conditional relevance?	114
5.7. Entry into physical co-presence as a blind spot in HRI.....	115
6. “PLAYING THE ROBOT’S ADVOCATE”: THE CONTINUOUS ACCOMPLISHMENT OF A ROBOT AS A SOCIAL AGENT IN A PUBLIC SETTING.....	117
6.1. Introduction.....	117
6.2. Data and method.....	118
6.2.1. Data collection	118
6.2.2. “Playing the robot’s advocate”: Delimitation of the phenomenon	118
6.3. Fragments	121
6.3.1. Glossing the robot’s (non-)response as an intentional act	121
6.3.2. Treating the robot’s conduct as indexing the absence of a sequentially adequate response from the main speaker	124
6.3.3. Describing the robot’s conduct as intentionally breaching the relevancies of the talk.....	128
6.3.4. Formulating the sequential implications of the robot’s previous utterance	131
6.4. Discussion	136
6.4.1. Collaboratively making sense of a robot’s conduct in a public space.....	136
6.4.2. Participants’ practical concerns when interacting with a robot in a public space	139
6.4.3. The robot as a tool to advance local human activities	142
6.4.4. Were bystanders doing “not-doing repair”?	143
6.5. The inner workings of the robot’s “sociality”	145
6.5.1. The interactional “good will” of participants.....	145
6.5.2. The burden of interactional work	146
6.5.3. Respecifying the “sociality” of a robot.....	147
7. “DISPLAYS OF HEARING” AND SOCIAL AGENCY: THE RELEVANCE OF AUTOMATIC SPEECH RECOGNITION TRANSCRIPTS IN INTERPRETING A ROBOT’S CONDUCT	148
7.1. Introduction.....	148
7.2. Previous work.....	151
7.3. Method	153
7.3.1. Natural data collection.....	153
7.3.2. Experimental data collection	154
7.3.3. Robot’s Conversational Design	155
7.3.4. Automatic Speech Recognition Transcript design	157
7.4. ASR transcripts and informational configurations.....	158
7.4.1. Similar informational configurations <i>outside of</i> social robotics	158
7.4.2. Similar informational configurations <i>with</i> conversational agents	159

7.5. Eye-tracking as a complement to an ethnomethodological and conversation analytic approach: some theoretical considerations.....	159
7.6. Gaze data analysis	160
7.6.1. Conditions	160
7.6.2. Data Preparation	161
7.6.3. Post-utterance gaze behaviors between conditions.....	164
7.6.4. Evolution of post-utterance gaze fixations over the course of the interaction....	165
7.6.5. Discussion: Gaze analysis as a preliminary overview.....	167
7.7. Qualitative fragments.....	168
7.7.1. Responding to the ASR transcript as a noticeable source of trouble before the robot's verbal and gestural answer	168
7.7.2. Responding to the ASR transcript as a noticeable source of trouble after the robot's verbal and gestural answer	175
7.7.3. Reading the transcript as a full-fledged speaking turn	178
7.7.4. The ASR transcript as a resource for third parties in contesting the sequential relevance of (embodied or verbal) responses from the robot	187
7.8. A wide range of situated uses for the ASR transcript	194
7.8.1. Troublesome ASR transcripts could override the verbal and gestural response of the robot	194
7.8.2. The specific situated functions of the ASR transcript.....	195
7.8.3. The ASR transcript as a publicly noticeable phenomenon.....	196
7.8.4. The crucial relevance of the ASR transcript for the robot's own perception of the situation	196
7.9. Some theoretical implications for mutual understanding	198
7.9.1. Sense-making and sequential adjacency	198
7.9.2. What is the minimum "interpretative latitude" required in human-robot interactions?	199
7.9.3. Impact on the "sequential plasticity" of the robot's turns	200
7.9.4. Does the ASR transcript enforce a cognitivist definition of mutual understanding?	201
7.10. "Interactional work" and the ASR transcript.....	203
7.10.1. Does the ASR transcript increase the "interactional work" required from human participants?	203
7.10.2. A new parameter to "work with" during tensions between progressivity and intersubjectivity	204
7.10.3. ASR transcripts as a systematic threat to (assumed) common understanding	205
7.11. Design recommendations	206
7.11.1. Preventing from using ASR transcripts in leisurely conversational interactions	206
7.11.2. Displaying a non-systematic written transcript.....	207

7.11.3. The specificity of leisurely interactions with a robot in a public space	208
7.12. Methodological takeaways.....	209
7.12.1. What is the human responding to? The tablet or the robot?	209
7.12.2. The limits of eye-tracking to capture participants' involved experience.....	210
8. CONCLUSION AND OPEN QUESTIONS	212
8.1. The practices glossed by “social agency”	212
8.1.1. Empirical results.....	213
8.1.2. Theoretical and practical frictions	214
8.1.3. A (fleeting?) specification of the emergence of robots as social agents?	215
8.2. Do human-robot interactions shed light on the inner-workings of “human-human” interactions?	216
8.2.1. Human-robot encounters as breaching experiments	216
8.2.2. Should we use “placeholder” analytical categories?	217
8.2.3. Analogically similar or sui generis practices?	219
9. REFERENCES	222
10. APPENDIX.....	253
10.1. List of figures and tables.....	253
10.1.1. List of figures.....	253
10.1.2. List of tables.....	256
10.2. Ethical approval and consent forms for the INSEAD experiment.....	257
10.3. Ethical approval and consent forms for the Cité des Sciences experiment	260

To Florence Rudaz

ACKNOWLEDGEMENTS

Acknowledgements are especially intricate when working with ethnomethodologists or conversation analysts. Their institutionalized habit of offering both their time and expertise during “datasessions” renders the intellectual lineage of some results nearly impossible to trace. Such datasessions allowed me to profit from the ideas, habits, raw material and support from more researchers than this page could display – several analyses in this manuscript are but clumsy attempts to fan the spark lit by casual conversations that arose during these events. In particular, I owe a lot to the participants of the EMCA-AI meetings, organized by Saul Albert, Hannah Pelikan and Lynn de Rijk. They offered a benevolent environment in which to slowly develop mental habits and an “eye” for the study of micro-sociological phenomena. As I was pursuing this PhD, Saul, Hannah and Lynn displayed an uncanny ability to balance their own rigorous research and the undoubtedly burdensome work of maintaining a scientific community afloat. Overall, whether our discussions took place during online datasessions or in person, this research has benefited immensely from the insights of Sylvaine Tuncer, Lucien Tisserand, Hendrik Buschmeier, Heike Baldauf, Marc Relieu, Nicolas Rollet, Julien Morel, Jakub Mlynář, Wyke Stommel, Chloé Clavel, Christian Heath, Mathias Broth, Brian L. Due, Paul Luff, Kerstin Fischer, Le Song, Andreas Bischof and many others.

I do not know if science is about standing on the shoulders of giants, but indubitably, this thesis rests on the goodwill and free time of engineers, designers and friends. Some of the most acute emergencies I encountered – a programming issue to address in the middle of an experiment, navigating through Paris carrying a humanoid robot, searching for a fully equipped laboratory for an eye-tracking study, etc. – were solved by people with little to no direct involvement in my project. To say it bluntly, this manuscript would not exist if no one had gone out of their way to help me. In particular, a huge part of the technical foundations of my experiments would have collapsed if many Aldebaran engineers had not struggled for hours on end to solve one or another of the obstacles I faced; often completely outside their official role and missions. Among many others, I am especially indebted to Marine Chamoux, Yufo Fukuda, Claire Rivoire, Miriam Bilac, Florian Guyet, Sera Buyukgoz, Susana Sánchez, Kyohei Ota, Robin Beilvert, Timothée Rey-Vibet, Thibaut Betbeder, Jessica Marthe-Rose, Adrien Kvaternik, Brian Giannini, Ravi Boolaky and Victor Paléologue.

Starting from the beginning of this project, I owe a lot to Sandrine Tourcher. First, for recruiting me as an inexperienced User Experience researcher back in 2019; then, after she heard of my interest in pursuing a PhD, for pushing this plan through its early stages. Her dedication transformed what started as a short mission at Aldebaran into an adventure spanning several years. Going even further back, I owe a still unpaid debt to Victor Vincent and his unshakeable belief that even the most arid of technical environments offer something to discover about the social world. As illusory as it may be to identify an exact start for this thesis, a major turning point occurred in the poorly heated Parisian café in which we discussed various connections between sociology and ergonomics, around March 2016. The patient input provided by Victor this evening ultimately led to my discovery of Marc Relieu’s excellent courses on conversation analysis and ethnomethodology. Marc himself would turn out to be pivotal in the genesis of this research; it is on his advice that Christian Licoppe and I first discussed the possibility of a PhD.

Looking back at the convoluted experiments I suggested in the early days of this PhD, I can only praise Christian's patience. I am extremely grateful for his trust and for his availability when I needed guidance. Throughout this thesis, I have shamelessly taken advantage of his readiness to scrutinize any video corpus I would throw his way, as tedious as this corpus may appear at first sight. Significantly, the data supplied in this work have been presented in numerous datasessions: Christian organized or initiated a major part of those. His rigor and swiftness in the minute analysis of strips of interaction was as impressive as it was helpful; after discussing his analytical finesse with other researchers, I now realize that I wasn't alone in thinking, "how does he do it?"

The temporality of research does not always match the rhythm of a technology company. Be they scientific, logistical, administrative, or financial, many of the challenges I encountered during this thesis seemed insurmountable: this or that experiment had to be scrapped, this or that chapter had to be trimmed. Marine Chamoux's sharp problem-solving skills and enthusiasm always proved me wrong. In spite of everything, she managed to create surprisingly stable working conditions in the middle of a fast-paced environment. I owe her a great deal. Rebecca Stower and Karen Tatarian's scientific and moral support were also unhelped for during the early days of this PhD – starting with their invitation to be involved in an expertly designed experiment, when my PhD had not even started yet. Now that I have a better grasp of the rigorous work required to conduct large-scale experiments, I realize how lucky I was to be served this one on a silver platter. I am truly thankful for their help. I am similarly grateful to the Cité des Sciences for offering me the opportunity to install and incrementally improve a rather cumbersome human-robot setup. Their commitment to support scientific research (even when it intruded upon their own facilities) enabled the collection of precious data. I am especially indebted to Hugues Pinson for his eagerness to try different human-robot settings and for providing me with a fully equipped laboratory overnight. Without his assistance, several sections of this PhD would still be mere drafts.

Naturally, as some distributed cognition theorists may suggest, this work would not exist if Aldebaran had not provided me with near-ideal writing conditions. The manuscript preparation phase of this PhD was immensely facilitated by the peaceful corner of Aldebaran's offices where I spent the past few months. If, by unfortunate chance, the reader were to come across imperfectly formulated ideas in the course of this work, may they rest assured that these passages would be in a much more regrettable state without the calm and benevolent atmosphere of Aldebaran's second floor. In particular, as one of the key "external artifacts" that determined the final form of this research, I am genuinely indebted to whoever installed, paid for and maintained Aldebaran's free coffee machines. This coffee was the strongest I have ever tasted. It realistically shortened the writing phase of this PhD by several months, as well as, probably, my life expectancy.

My most heartfelt thanks go to Agathe for putting up with me during the early stages of this PhD and during the exhilarating years that preceded it. I cannot overstate how important her support was to the emergence of this project. Many thanks to Eugénie for her vital presence and for finally realizing that a PhD is not an internship. Additionally, thank you to Vincent, Maxime, Yassine, Matias, Reinhold and Florian for their support and constant availability. Some acknowledgement should legitimately go to Octavius, my cat. Octavius' intellectual influence on this work has been limited at best.

In the end, thank you to Florence Rudaz and Eric Rudaz for their continued support over these long years. They made this doctoral thesis possible. For better or worse, my inclination for research stems from the intellectual curiosity they always displayed.

TRANSCRIPTION CONVENTIONS

Transcription conventions for verbal interactions

Transcription of talk follows Jefferson's (2004) transcription conventions:

=	Latching of utterances
(.)	Short pause in speech (<200 ms)
(0.6)	Timed pause to tenths of a second
:	Lengthening of the previous sound
.	Stopping fall in tone
,	Continuing intonation
?	Rising intonation
°uh°	Softer sound than the surrounding talk
.h	Aspiration
h	Out breath
heh	Laughter
((text))	Described phenomena
.tk	Lips parting or a lip smack

Transcription Conventions for Embodied Conduct

Embodied actions were transcribed using Mondada's (2016) multimodal transcription conventions¹:

**	Gestures and descriptions of embodied actions are delimited between:
++	two identical symbols (one symbol per participant)
ΔΔ	and are synchronized with corresponding stretches of talk.
*->	The action described continues across subsequent lines
-->*	until the same symbol is reached.
>>	The action described begins before excerpt's beginning.
-->>	The action described continues after the excerpt's end.
...	Action's preparation.
--	Action's apex is reached and maintained.
,,,	Action's retraction.
ric	Participant doing the embodied action is identified in small caps in the margin.

Symbols and abbreviations used in transcriptions refer to the following multimodal dimensions:

HUM	Turn at talk from a participant (HUM, EMM, BOB, etc.)
ROB	Turn at talk from the robot
hum	Multimodal action from a participant (hum, emm, bob, etc.)
rob	Multimodal action from the robot
img	Screenshot of a transcribed event
Δ	Human's head

¹ <https://www.lorenzamondada.net/multimodal-transcription>. Accessed January 29, 2024.

£ Human's gaze
* Human's arms
@ Human's whole body
^ Movement in space
\$ Robot's arm
+ Robot's gaze
% Robot's belly screen (what is displayed on its tablet)
Position of a screenshot in the turn at talk

1. INTRODUCTION

1.1. Social robots or talking bundles of wires and plastic?

1.1.1. Wishful representations and real-life interactions

It is 2018. An animated video, commissioned by the marketing department of a robotics company, depicts the intended interaction between a robot receptionist (Pepper) and a human (Carla) in the hall of an office building. As Carla approaches, the robot locks eyes with her, inquiring how it can assist. With a warm smile, Carla introduces herself and mentions she has an appointment with Mr. Smith, emphasizing these last words with a wag of her finger. The robot greets her in return and confirms that Mr. Smith will be notified of her arrival. Moments later, it updates Carla that her visitor badge is prepared and that Mr. Smith awaits her. Still smiling, Carla flashes an “ok” sign to the robot and departs.

It is 2021. At the entrance of a tech-oriented company, a real-life interaction occurs between an employee and the receptionist robot through which it is mandatory to check in every morning. The human kicks the robot, grabs its head, shouts his full name, preemptively utters “yes” as the robot starts to formulate a question, then leaves.

Of the two encounters outlined above, the first, originating from a promotional video, constitutes a wishful representation of what should be a social interaction with a receptionist robot². The second showcases an expert user attempting to accelerate a check-in sequence as much as possible by exploiting the robot’s hardware and programming – a practical mastery honed over and over, during multiple interactions with the same robot. Indeed, this (real-life) employee’s behavior is not gratuitous violence. It is an extreme optimization of a task that he completed daily for one year with the (intended to be) social robot receptionist in the entrance of his workplace. Namely, by kicking the robot, the employee triggered a “looking around” animation during which the robot raised its head. Grabbing the robot’s head (and tilting it upward even more) then ensured that the robot rapidly detected a face at close range, which, in turn, launched its “check-in” script. In the end, the robot attempted to request its interlocutor to confirm his arrival: the employee responded before this line was fully pronounced, or, rather, he spoke as soon as the robot could register an answer to its (not yet formulated) question. By experimenting with different response timings over the course of his morning check-ins, he had built a sense of the earliest moment at which the robot’s speaking turn could be responded to.

The contrast between these two scenes (one imagined, one observed) illustrates a fact that roboticists are regularly confronted with: *the intrinsic properties of a robot (its shape, its ability to detect human bodies and faces, to respond to human utterances, etc.) do not guarantee its treatment as a social agent.* A robot deemed “social” because it exhibits specific

² The promotional cartoon from which stems this description is available on the Youtube channel of the social robotics company Aldebaran: <https://www.youtube.com/watch?v=yy8h2hgmPsM>. Accessed February 28, 2024.

technical features (labeled as such by its designers or engineers) does not necessarily emerge as social in situ, i.e., in the way it is treated by humans in co-present interactions. As hundreds of human-computer and human-robot studies have noted, an entity that one person “greet” and “speaks to” like a human may be, instead, “operated” or “maneuvered” by another as a stubborn computer whose vocal and embodied interface merely slows down whatever task needs to be accomplished. This rather mundane observation constitutes the starting point of the following work. Through the moment-to-moment analysis of a large corpus of human-robot interactions in different settings, we attempt to identify endogenous features (i.e., relevant to the participants themselves as they are involved in an ongoing encounter) which may partially *explain* the manner in which a robot is observably responded to during an interaction. That is, we aim to both *specify* what it is to treat a robot as a social agent and, in the same movement, to shed light on local properties of human-robot interactions that are consequential on the emergence of robots as such “social” entities – rather than mere moving and talking heaps of wires and plastic. That is, before pondering *why the real-life employee did not treat the robot like in the promotional video described above*, we ask *what it even means, in terms of empirically observable practices, to treat the robot “like in the promotional video”?*

1.1.2. Human-robot interactions as black boxes

While discussing a segment of talk between a doctor and a patient, Heritage (2001) notes that

“[i]n the past, social scientists have had little to say about how interaction works, treating it as an invisible or inscrutable ‘black box’. [...] If you had been presented with this segment in 1960, you would have found few systematic resources with which to analyse what is going on in this segment, and none which could offer any significant clues as to the details of the actions the participants are engaged in. In general, the social science of the period was highly abstract and unconcerned with the specifics of everyday conduct. In fact, it was believed that individual episodes like the doctor-patient exchange above are fundamentally disorderly and that attempts at their systematic analysis would only be a waste of time.”

One might contend that the general field of human-robot interactions (thereafter HRI) has, until recently, similarly treated as black boxes the very “interactions” it posits. Although related fields (e.g., computer-supported cooperative work) have taken a more granular approach to empirical data, often heavily influenced by ethnomethodology (Randall et al., 2021), HRI research generally stays at a safe distance from the inner workings of encounters between humans and robots – although it has increasingly turned its attention to natural or semi-natural settings (Sabanovic & Chang, 2016). Whether or not they are considered “fundamentally disorderly” (Heritage, 2001), single occurrences of human-robot encounters are, often, de facto treated as too intricate for a corpus analysis to identify patterns or yield generalizable results.

Yet, this reluctance to look at the details of the organization of interactions between robots and humans leads to an underspecification (Beach, 1994) of interactional phenomena taking place in these settings. In spite of the potential otherness of “social robots” as interactants, it is extremely common (and, arguably, unavoidable, for lack of a more specific vocabulary) to characterize them or their interactions with lay and unexamined concepts. Either in everyday discussions, sales speeches, scientific talks, or in robotic companies’

meetings aimed at designing new robots, there is often no choice but to briefly gloss human-robot encounters. For instance, stating that human participants make “contact”, get “briefly familiar” (Kaipainen et al., 2018), or “lose interest” (Leite et al., 2013) with a robot inevitably indexes specific interactional accomplishments. When a robot is said to elicit “more interactions (i.e., passersby touching the screen)” (Brenngman et al., 2021) with clients in a store, compared to a tablet service kiosk, this gesture of “touching the screen” may take place as a component of different sequences or actions – greeting the robot, producing an in-the-wild Turing test (Ivarsson & Lindwall, 2023), etc. – here aggregated under the label of “interaction”. In the end, when a lay observer intuitively recognizes and labels human-robot interactions (e.g., as a “conversation”, as a particular case of a “wow effect” (Fuentes-Moraleda et al., 2020), as “getting acquainted” (Lighthart et al., 2022), etc.) some features of these situations must make them recognizable as a certain type of activity – even if these distinctive features are “seen but unnoticed” (Garfinkel, 1964).

We certainly do not aim to specify all these terms, nor do we suggest their authors should have used alternative formulations for all practical purposes. In fact, given the lack of micro-analytic data dealing with human-robot interactions, and the limited vocabulary available to describe the here-and-now of these interactions, it is unlikely the accounts mentioned above could have been phrased differently without extensive additional research work. It goes without saying that scientific communication would not be possible if each account, description, or label had to be specified in regard to all its imaginable dimensions (micro and macro); if each social setting had to be detailed in all its possible levels of description. Rather, what we point to is the relative absence of studies detailing the interactional phenomena commonly glossed by lay formulations, like those reported above. This state of affairs, coupled with the tendency “for ordinary language descriptions to gloss or idealize the specifics of what they depict” (Heritage, 2013a), entails that, should we wish to discuss them, the “practicalities” (Heath & Lehn, 2004) of human-robot interactions are currently obscure. Few studies attend to the micro-order phenomena indexed by the terms we intuitively use to speak about human-robot encounters.

In particular, as we will argue throughout this work, one of the most underspecified terms in human-robot interaction is that of “social” robots. Based on McDermott (1976), Mitchell (2021) speaks of “wishful mnemonics” to refer to “terms associated with human intelligence that are used to describe the behavior and evaluation of AI programs” (Mitchell, 2021): for instance, stating that an AI “understands” although its method of processing information is distinct from the human process of “understanding” – or, at least, although what “understanding” involves was not specified. Under many facets, labeling robots as “social” or describing their activity as “interacting socially”, “conversing” (Button & Sharrock, 1995), “collaborating”, etc. are wishful mnemonics for social interaction – rather than for human cognition³. These terms conflate what the robot is intended to co-accomplish with its human users and what occurs (or does not occur) in concrete situated interactions. Their conventional use conveys a strongly internalist definition of “sociality” as based on robots’ intrinsic properties (Alač, 2016; Licoppe & Rollet, 2020): e.g., having two arms, being able to detect humans, to produce gestures, to answer questions, to build a world model, or any imaginable criteria. By positioning the core of “what sociality is” away from the details of situated encounters between robots and humans, this prevailing internalist definition may partially explain the scarcity of data on the inner workings of human-robot interactions. Such ex-ante

³ Arguably, the same could be said about describing some robots’ conduct as “greeting”, “requesting”, etc.

characterizations of a robot as “social” run the risk of naturalizing the observable result from micro-sociological processes.

In short, there is an immediate interest in studying how the previously mentioned glosses (getting “acquainted” with a robot, “opening an interaction” with a robot, having a “social interaction” with a robot, etc.) “[...] are organized in a fashion meriting descriptions offered (i.e., that the kinds of phenomena purported to exist warrant one kind of characterization in lieu of, perhaps in unison with another), that claims put forth regarding the social world are actually there and not simply a figment of researchers’ unexamined intuitions” (Beach, 1994). Indeed, such an endeavor is not purely linguistic, a mere quietist attempt at clarifying conceptual mix-ups, nor is it limited to translating well-known phenomena within a different level of description (List, 2019). By attending to the locally organized practices taking place in human-robot encounters, it becomes possible to provide another type of explanation as to “what went wrong” or “what went right” in an interaction: that is, *to produce explanations based on the features made relevant by the participants themselves within their endogenously organized interaction*, rather than on questionnaires or interviews. Plainly, participants’ experience of technological artifacts is mediated by their situated encounters⁴. Identifying these practicalities makes apparent local features of the interaction that matter to humans as they are immersed within the urgency of these ongoing encounters – precisely those features which can be leveraged to make the robot emerge as “social”. Although many exogenous variables may indirectly impact the local unfolding of an interaction (age, gender, attitude towards technology self-reported on psychometric scales⁵ before or after the experiment, etc.), these variables, and the potential “hidden social orders” (Livingston, 2008) they may document, “must enter the stream of discourse via particular mechanisms of production” (Beach, 1994)⁶. In essence, by paying close attention to the intricacies of human-robot interactions, one can prevent “premature theorizing” (Albert & de Ruiter, 2018) and hope to directly confront the shortcomings evident in current “social” robots’ situated encounters.

1.2. Topic of study, data, and plan

1.2.1. Topic of study

This work will examine *the members’ methods in and through which a robot emerges as a social agent in different settings*. Because of the enormous number of conducts that fall under the scope of such an endeavor, we will attempt to specify *some* recurring methods, practices,

⁴ Studying visitor’s behaviors in museums, Heath et al. (2005) remark that “conduct and social interaction at the exhibit-face remains unexplored territory and yet it provides the foundation, the very basis, to people’s experience of and the public understanding of science in science centers and museums” (Heath et al., 2005). The situation is similar with “social” robots.

⁵ E.g., Edison & Geissler (2003)

⁶ This stance is a knowingly assumed variant of what Sawyer (2015) calls “interpretivism”: “By ‘interpretivism’, I refer to a broad group of theories that have in common the belief that social reality can be studied only in terms of individuals’ interpretation of it. [...] Interpretivists believe that macrostructural forces never constrain individuals; rather, they are mediated by those individuals’ interpretation of them.” (Sawyer, 2015).

processes, etc. that participants display when organizing their activities with a robot – while maintaining this robot as a full-fledged and competent social participant.

To do so, we will study encounters occurring in different settings (natural and experimental), at different phases of an ongoing interaction, with robots displaying different conducts. Crucially, our goal is not to produce an abstract model of some hypothetically transsituational steps through which a robot emerges as an agent, independently from its design, its programming, and from the background expectations of human participants in different settings. Rather, this work merely strives to add to the list of documented interactional processes occurring during human-robot encounters (indexed to specific settings, sequential contexts, spatial configurations, etc.) through which a robot can be said to be, momentarily and locally, treated as a social agent. In doing that, we attempt to expand the list of local features – including the robots' design or conduct – documented as typically relevant in human-robot interactions. In other words, among the potentially infinite number of properties of a setting that can be perceived and attended to, we attempt to understand what is *used* by human participants (and, to a certain extent, by the robot) to progress the activity at hand. Yet, these features will be defined as “relevant” only in regard to a local and temporal context – in contrast with a necessary, ubiquitous and atemporal, relevance. In sum, should this work add to the existing body of literature on human-robot interactions, its contribution must be viewed as part of an incremental and cumulative approach to the local organization of social activities.

1.2.2. Corpora

Even the most widespread interactional phenomena identified in this work are acutely historical: they are connected to the quickly evolving state of the technology and with the general public's current familiarity with robots. As robots get more common, different manners of interfacing with them might appear as well: alternative ways to initiate a conversation, to repair a mishearing, to joke with or about them, etc. As is the case for all micro-analytic works attending to the use of new technologies, producing results that remain relevant at the instant of their publication is a race against time (Mlynář & Arminen, 2023). This is especially true for this research, which was undertaken amidst the appearance of publicly available chatbots relying on large language models like ChatGPT. Still, as argued in a recent trend of Ethnomethodological and Conversation Analytic studies (see section 8.2 at the very end of this study), the multiplication of environments and types of (non-)human interactants, rather than being a hindrance, may allow analysts to discern, by contrast, some fundamental and relatively stable properties of sociality (Mondémé, 2022). In this perspective, the marked instability of “ordinary” human-robot practices over time and across settings is advantageous: it creates a shifting fresco against which some of the core methods that humans use to construct social facts can stand out.

Hence, a core idea of this research is that varying the settings should lead to a more complete panorama of the recurrent methods, practices, categorizations, etc. that human participants display in their interactions with robots. Additionally, multiplying the settings provides more opportunities for the robot's potential “otherness” to breach or disrupt humans' ordinary background expectations (Garfinkel, 1967 – see section 8.2.1 for an extended discussion on this topic). Consequently, the analyses presented in this work are based on the following four corpora.

1. A corpus of experimental data recorded at the INSEAD-Sorbonne Université Behavioural Lab INSEAD in February 2020 (see chapter 5).
 - Composed of 80 videos of dyadic interactions with an autonomous robot.
 - The robot was programmed to act as a “travel agent” attempting to plan participants’ ideal holidays.
2. A corpus of naturally occurring data, recorded in the hall of the science-oriented museum “Cité des Sciences” in July 2022 (see chapter 6).
 - Composed of 100 videos of encounters between an autonomous robot and single visitors or groups of visitors.
 - The robot’s conduct was controlled by a rule-based “conversational” chatbot designed to respond to humans on a wide variety of topics.
3. A corpus of semi-experimental data recorded in a laboratory at the science-oriented museum “Cité des Sciences” in July 2022 (see chapter 7).
 - Composed of 108 videos of encounters between an autonomous robot and single visitors or groups of visitors.
 - The robot’s conduct was controlled by a rule-based “conversational” chatbot designed to respond to humans on a wide variety of topics.
 - This corpus includes data regarding participants’ gaze fixations during their interaction with the Pepper robot, recorded with an eye-tracking device.
4. A corpus of semi-experimental data, recorded in the entrance of an office building in July 2021 (see section 3.3.3).
 - Composed of 50 videos of encounters between an autonomous robot and employees familiar with this robot, as well as of 10 videos of encounters between this robot and visitors who had never spoken to a Pepper robot before.
 - The robot was programmed as a “receptionist” robot, with which visitors had to check in and check out of the building.

Each of these natural or controlled experiments uses the Pepper robot, built by the robotics company Aldebaran (formerly named Softbank Robotics) – see section 1.3. The particular programming for the robot as well as the exact setup for each experiment are detailed in the sections dedicated to their analysis (chapters 5, 6, and 7).

1.2.3. Plan

1.2.3.1 Preliminary methodological and theoretical considerations

This work is structured into chapters and sections as follows:

Chapter 1 (this chapter) introduces the general topic of the research, its structure, the corpora on which our results are based, and the humanoid robot featured throughout our experiments.

Chapter 2, Section 2.1 presents our analytical approach (Ethnomethodology and Conversation Analysis, thereafter EMCA) and the relevance of its cognitively agnostic stance for the study of human-robot interactions.

Chapter 2, Section 2.2 narrows down on some “conceptual loosening” that arise from using the set of analytical tools associated with EMCA (built from and for the study of “human-human interactions”) to shed light on so-called “human-robot interactions”. In particular, we investigate the limited relevance of the “next-turn proof procedure” in human-robot interactions: i.e., we examine to which degree the reactions of a robot can be said to confirm or contradict the analyst’s interpretation as to what human participants are “doing” in a recorded interaction. We argue that the applicability of the notion of membership knowledge to a robot is key to this inquiry. Finally, we assess the remaining grounds on which can be produced argumentatively sound interpretations of human-robot interactions.

Chapter 2, Section 2.3 briefly reviews the existing literature on rule-based AIs (the so-called Good Old-Fashioned AIs). It focuses on the well-documented limits of this approach to create artificial agents that function like everyday “expert” human conversationalists (who are argued not to follow rules as they pre-reflexively “cope” – Dreyfus, 2001 – with the urgency of social interactions). We then attempt to clarify the room for maneuver of “applied CA” (Antaki, 2011a; Ten Have, 2007) when it comes to the design of robots that rely on rule-based chatbots. We argue that the methods or norms that an EMCA approach can salvage from the observation of the practical activity of human conversationalists can, at most, turn rule-based robots into better “novices”.

Chapter 3 provides a definition for the central concept of “social agency” as a locally emergent property. We start by reviewing the long-standing ethnomethodological tradition that attends to the locally relevant features of a given setting as produced and maintained by co-present participants’ mutual conducts. This analytical stance is, arguably, the core “study policy” (Garfinkel, 1967) of ethnomethodology: it investigates each property of a situation “as if it were something that emerged from the activities of parties to that situation” (Dennis, 2003). Then, in line with this ethnomethodological perspective, we attempt to establish the interactionist definition of “social agency” to be used throughout this work – in opposition to an internalist definition of “social agency” as the unmediated consequence of a robot’s technical abilities (being able to speak, to move its arms, to detect and to respond to humans, etc.). To do so, we review the existing literature in Ethnomethodology and Conversation Analysis to identify the core interactional phenomena that compose what is commonly glossed by “social agency”. That is, we attempt to specify at least some of the typical processes that are indexed by this label in micro-analytic studies of human-robot encounters. Our goal is to formulate an “empirically viable” definition of social agency for our analytical purposes: one that a researcher can mobilize to distinguish between “social” and “non-social” treatments of a robot based on recorded, observational, data. We identify three defining criteria for the concept of “social agent”, as it is commonly used in the ethnomethodological and conversation analytic literature. To be a “social agent”, a robot must be an accountable participant in both senses of the term, i.e., in regard to intelligibility and responsibility (first two criteria) and its conduct must be treated as contingent to the local situation (third criteria). In other words, we state that, throughout this work, a “social agent” will be identical with an accountable entity granted with specific interactional rights and obligations, whose conduct is observably attended to as contingent on the ongoing situation.

1.2.3.2 Technological and industrial context

Chapter 4 describes the industrial context in which our analytical approach took place. Because an EMCA perspective is much less established in private technology companies than, say, user experience (UX) methods based on experimental psychology, we relate disjunctions that resulted from the application of this approach in a company rather unfamiliar with EMCA. We explore several instances of those ordinary frictions: temporality of EMCA research *versus* time constraints of a technology company, incommensurability between engineers or designers' theoretical background and an EMCA perspective, etc. In the end, after identifying diverging underlying definitions of “conversation” – between designers, engineers, and EMCA analysts – we provide some examples of the manner in which these representations or values are embedded in technical systems used for designing or programming conversations on the Pepper robot.

1.2.3.3 Empirical studies

Chapters 5, 6, and 7 present the three empirical studies underpinning this research.

Chapter 5 examines the very first moments of co-presence, during which a robot appears to a participant for the first time. It argues that, although these initial moments are often “off-the-record” in the data collected from human-robot experiments (video recordings, motion tracking, methodology sections, etc.), they do not constitute an interactional vacuum. This phase (which can be analogically described as a “pre-beginning” phase) is where interactional work from participants can take place so that the production of a first speaking turn – like greeting the robot – becomes relevant and expected. We base our analysis on an experiment that replicated the interaction opening delays sometimes observed in laboratory or “in-the-wild” human-robot interaction studies – where robots can require time before springing to life once they are in co-presence with a human. In the end, we identify several properties of the robot's behavior oriented to by participants as creating the adequate conditions to produce a first greeting – through which the robot was momentarily treated as an agent. Our findings highlight the importance of the state in which the robot originally appears to participants: as an immobile object or, instead, as an entity already involved in preexisting activity. That is, participants' orientations to the very first behaviors manifested by the robot during this “pre-beginning” phase produced a priori unpredictable sequential trajectories, which configured the timing and the manner in which the robot emerged as a social agent. We suggest that these first instants of co-presence are not peripheral issues with respect to human-robot experiments but should be thought about and designed as an integral part of those.

Chapter 6 is based on the analysis of a corpus of naturally occurring data collected at the Cité des Sciences et de l'Industrie, one of the biggest science museums in Europe. These interactions took place between the visitors of this museum and a Pepper robot, programmed to “converse” on a wide variety of topics and placed in a hall of this museum. Like in the previous section, we investigate the methods, processes, practices, etc. through which a robot labeled as “social” emerges, even momentarily, as a social agent. Yet, this section specifically focuses on one recurrent practice through which humans “worked” to maintain the robot as an agent: the description by bystanders, in a way that was made accessible to the main speaker, of the social action that the robot was taken to be accomplishing. We name this practice “playing the robot's advocate”. In doing so, bystanders maintained the robot's (sometimes incongruous) behavior as relevant to the activity at hand and preserved the robot itself as a

competent participant. Relying on these data, we argue that ex-ante definitions of a robot as “social” (i.e., before any interaction occurred) run the risk of naturalizing as self-evident the observable result from micro-sociological processes: namely, the interactional work of co-present humans through which the robot’s conduct is reconfigured as contextually relevant.

Chapter 7 investigates the impact of a nominal feature of the Pepper robot on its emergence as a social agent: namely, displaying what it “hears” from its interlocutors’ speaking turns – on a tablet attached to its torso. In other words, the robot constantly provides a public transcript of its own reconstruction of what is being said by its human co-participants: the “automatic speech recognition transcript”. This feature, although it is commonly found on smartphone vocal assistants, is rare among “social” robots. We attempt to demonstrate that it can be heavily consequential on the interpretation of the robot’s conduct as contextually relevant and, in particular, on the way miscommunications are detected and dealt with. That is, this public transcription reconfigures the informational ecology of the interaction: it opens a window on what is going on “inside the robot”, before this robot produces any verbal or embodied response (speech, sound effects, gestures, LEDs, etc.) to the previous turn of an interlocutor. Based on semi-experimental data, we describe several typical ways in which this “automatic speech recognition transcript” was used as a resource by participants in situations of miscommunication, and, more specifically, when the robot “misheard” the previous turn of a human. We finish by linking these qualitative findings with an interesting pattern visible in the data obtained from our eye-tracking device: as the interaction unfolded, the attention of humans (materialized by their gaze) slowly focused more and more on the robot’s tablet, while its head and gestures were gazed at less and less. We attempt to exemplify that an ethnomethodological and conversation analytic approach is fit to clarify the local interactional phenomena aggregated behind this progressive focus on the tablet, that the eye-tracking analysis reveals.

1.2.3.4 Conclusion and implications of the previous results

Chapter 8, Section 8.1 summarizes the articulation of the main empirical findings presented throughout this work:

1. Evidently, a robot pre-labeled as “social” does not necessarily emerge as a social agent *in situ*. However, even the very first instants of an encounter with a robot can be consequential on its momentary emergence as an accountable conversational partner, or, instead, on its persistent treatment as an object merely producing pre-recorded turns.
2. Yet, even when a robot may be glossed as being “treated like a social agent” in public interactions with groups of humans, a closer look reveals the recurring work stemming from bystanders to re-configure the robot’s conduct as relevant to the task at hand. That is, not only are some robots’ abilities to produce relevant contributions scaffolded *before* the interaction (as a significant body of literature has demonstrated), but the robot’s conduct can also be framed *a posteriori* by the audience to be meaningful for the person directly interacting with this robot.

3. Finally, an especially configuring parameter of the treatment of the Pepper robot as a competent participant⁷ is the “automatic speech recognition transcript”: that is, the public display by this robot, on a screen placed on its torso, of the words it “hears” from its interlocutor’s speaking turn. Indeed, among other pragmatic consequences, having access to what the robot (mis-)heard regularly led participants to *the a posteriori re-evaluation* of the situational relevance of the robot’s actions – even when these actions were so far treated as adequate responses. In different terms, this re-evaluation overrode these participants’ interpretation-so-far of the robot’s conduct as activity-relevant: it displayed a change in participants’ footing regarding the status of the robot’s conduct *as a response to what they really requested, asked, remarked, etc.* Consequently, we argue that, in an informational configuration where participants have direct access to a robot’s reconstruction of the external world, the “weave of interactional moves” (Linell & Lindström, 2016) produced between humans and robots is less likely to offer the (superficial) coherence which is the bedrock of an experience of the ongoing interaction as intersubjectively shared.

This chapter concludes by arguing that these results contribute (along with other EMCA-oriented works) to respecifying what a “social” robot is.

Chapter 8, Section 8.2 opens a discussion about a pervasive issue in EMCA-inspired examinations of interactions with non-humans: the specificity of the methods, processes, practices, etc. identified in these settings. Namely, are the practices deployed in such settings *sui generis* or, instead, identical to those observed in human-human interactions? For example, what ordinary human-human practices are “breached” or disrupted (if any) in first encounters between robots and humans? We argue that answers provided to this set of questions have deep implications regarding the relevance of EMCA’s tools (built from the study of human-human interactions) to examine activities taking place in settings involving robots and humans.

1.3. Technological and organizational context for this work

1.3.1. The Pepper robot

All the experiments presented in this work were carried out using Pepper, a 1.20-meter-tall humanoid robot produced by Aldebaran (see Figure 1.1). This section specifically describes the features of Pepper that are crucial for its ability to perceive or engage in multimodal social interactions.

⁷ Interactional competence is one of the criteria required for a robot to emerge as a social agent, as defined in section 3.3.2.

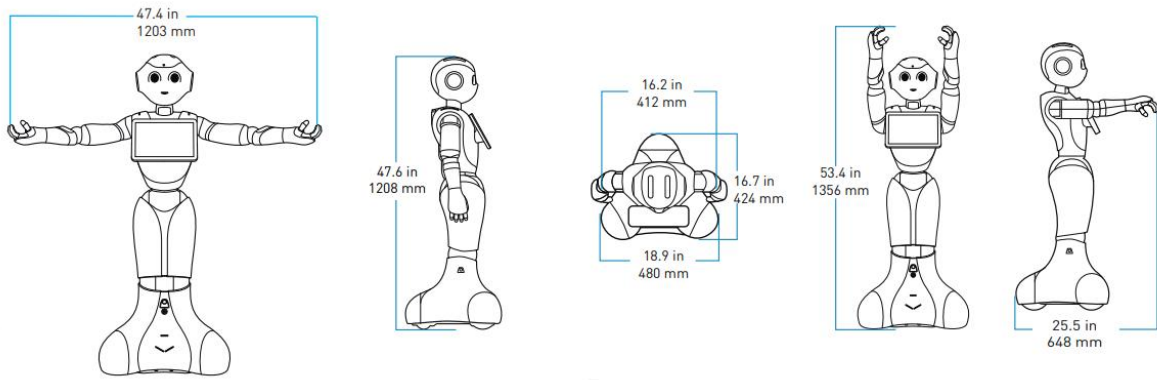


Figure 1.1. Pepper's dimensions⁸

1.3.2. Pepper sensors (what can be leveraged to perceive human actions)

Pepper can gather information from its environment via several cameras, lasers, sonars, microphones, tactile sensors, and a touch-sensitive screen on its torso (see Figure 1.2). It is through these instruments that Pepper registers specific features of its immediate surroundings. In a different vocabulary, what may analogically be described as Pepper's perceptual environment is shaped by the stream of data gathered by these sensors and processed through its algorithms.

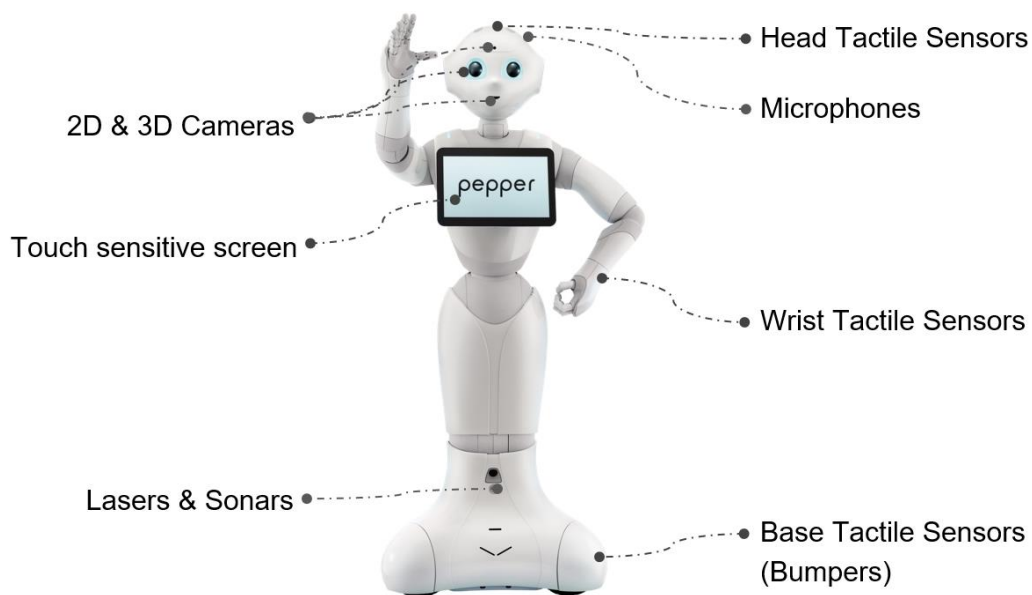


Figure 1.2. Simplified layout diagram of Pepper's sensors

From the information collected by these sensors, Pepper provides various resources (raw data or high-level variables – Ben-Youssef et al., 2017) to be used by interaction designers or

⁸ Retrieved February 22, 2024 from <https://support.aldebaran.com/support/solutions/articles/80000958735-pepper-technical-specifications>. Copyright (2022) Aldebaran & United Robotics Group.

developers. Overall, Ben-Youssef et al. (2017) categorize three levels of data streams for Pepper:

“Low level (Measured signals): Videos [...], Audio [...], Sonars, Laser

Intermediate level (Tracked variables): Face, Head motion [...] Eye gaze (direction, opening degree), User distance, User position, Engagement Zone, Robot behaviour [...]

High level (Output of other modules): Smile degree, Facial expressions (neutral, happy, surprised, angry or sad), Looking at robot, User dialog input (ASR) [Automatic Speech Recognition] [...]” (Ben-Youssef et al., 2017, emphasis ours)

For example, consider an engineer who wants Pepper to greet humans showing clear interest in interacting with it, as opposed to those merely passing by. To build a system that differentiates between “interested” and “uninterested” humans, this engineer may use changes in humans' positions relative to the robot, their gaze direction, and what these humans might be saying (as interpreted by the Automatic Speech Recognition), among other behavioral cues.

1.3.3. Pepper multimodal abilities (what can be leveraged to produce social actions)

To communicate, Pepper can be programmed to use gestures (e.g., head-tilt, arm or wrist movements – see Figure 1.3), postures (e.g., leaning forward), positioning (e.g., approaching a human), voice, sound, as well as the content displayed on the screen on its torso.

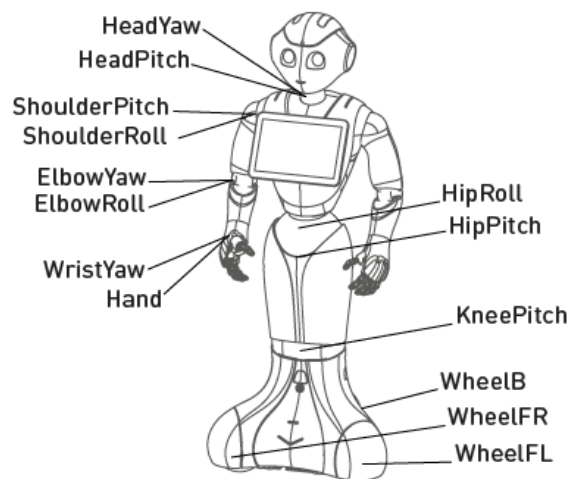


Figure 1.3. Pepper's degrees of freedom⁹

⁹ Retrieved February 22, 2024 from <https://support.aldebaran.com/support/solutions/articles/80000958735-pepper-technical-specifications>. Copyright (2022) Aldebaran & United Robotics Group.

1.3.4. Programming Pepper

1.3.4.1 *An autonomous robot*

The particular programming of Pepper for each experiment is detailed in their respective chapters. Overall, in each of those experiments, Pepper was configured as an “autonomous robot” in the weak sense (Beer et al., 2014) of the definition: it could carry out, to some degree, “its own processes and operations without external control” (Beer et al., 2014). In more concrete terms, the robot did not require intervention from an experimenter to respond¹⁰ to human inputs *during the experiments themselves*: no Wizard of Oz procedure was involved, and the robot was configured to react to some features of its interlocutors’ conduct – i.e., its observable behavior was not independent of inputs coming from physically co-present humans, unlike a fully scripted automaton or a fully teleoperated robot (Yanco & Drury, 2004).

1.3.4.2 *A rule-based robot*

In the recent versions of Pepper we used (featuring the operating system NAOqi 2.9), the multimodal abilities mentioned above could be programmed using the Pepper SDK plug-in¹¹ on Android Studio. See section 4.6 for an extended description of the programming and scripting tools available for Pepper 2.9. Hence, the Pepper robots mentioned throughout this work were systematically “rule-based”: in each experiment, their conduct was governed by a chatbot that relied on more or less complex dialog trees. The robot’s behavior exclusively depended on rules that connected specific inputs (e.g., Pepper hears “hello” and detects a human shape at close range) to specific outputs (Pepper waves and says “hello”). That is, significantly for the analysis produced afterward, the Pepper robots featured in this research did not rely on conversational agents based on large language models (e.g., ChatGPT) or other types of generative AIs. The data collection phase for this thesis spans from 2020 to 2022, whereas the initial attempts to implement ChatGPT on Pepper started in 2023.

1.3.5. Organizational context: EMCA in a technology company

Critically, this work was partially produced from within the company that creates the robot our participants interacted with: it results from an industrial partnership with the technology company Aldebaran (formerly named Softbank Robotics), which designs and manufactures the Pepper robot. I began working on this thesis in 2020, after having held the position of UX Researcher at Aldebaran for a year. Thus, from 2020 to 2024, I spent most of my time in Aldebaran’s office, working on my own human-robot experiments and occasionally helping with the organization and analysis of internal user tests and workshops – although this “UX testing” activity occurred less and less often as the thesis progressed. Formally or informally,

¹⁰ However, as discussed in chapter 6 of this work, the question remains as to how much intervention from human co-participants is required for the robot to emerge and persist as a competent agent. In this sense, we will argue that the robot could only “function” as a social agent thanks to the constant monitoring and interactional work (Chevallier, 2023; Due & Lüchow, 2022) of co-present humans.

¹¹ Documentation for the Pepper SDK plug-in can be found at <https://qisdsk.softbankrobotics.com/sdk/doc/pepper-sdk/index.html>. Accessed February 22, 2024.

I was consistently supported by the engineers or designers of the company whenever I needed to configure the Pepper in a specific manner and to script a particular behavior on it.

Consequently, chapter 4 documents the use of an Ethnomethodological and Conversational Analytic approach amid collaborations with designers, user experience researchers, and engineers in Aldebaran's office to "make the robot work" or to improve its design: whether as part of my early days as a UX researcher, or as part of my own experiments on human-robot interactions. Because the analytical stance of Ethnomethodology and Conversation Analysis is rather uncommon in technology companies, this partnership offered various examples of ineffability or incommensurability with alternative approaches to the study of human-robot interactions (as well as examples of friction with the design and programming tools pre-adapted to these approaches). I argue that these frictions and ineffabilities were especially meaningful in that they mutually revealed underlying representations as to what is a "conversation" or a "social action" – see sections 4.5 and 4.6.

2. METHOD

2.1. Ethnomethodological Conversation Analysis and Artificial Agents

2.1.1. Ethnomethodology and Conversation Analysis

The analytical perspective underpinning this research is Ethnomethodological Conversation Analysis (Mondada et al., 2020) – thereafter EMCA. This micro-sociological approach studies the temporal unfolding of events in an interaction in order to understand, on a moment-to-moment basis, how participants’ “actions—turns-at-talk, embodied actions or complex multimodal moves— are produced and recognised as performing meaningful social actions” (Tuncer et al., 2022a). In this endeavor, the fundamental stance of EMCA is that the observable orderliness of everyday interactions results from “specifiable interactional methods [...] that are used by members as solutions to specifiable organizational problems in social interaction” (Ten Have, 1990). Hence, because of the “remarkably minimalist character” (Button & Sharrock, 2016) of this approach, a definition of EMCA could reasonably end here: EMCA researchers merely aim to report “the methods persons use in doing social life” (Sacks, 1984a), that is, “to formulate the means used by the members in their situated interactions” (Ten Have, 1990). Even so, this fundamental stance can be broken down into several analytical practices:

First, EMCA adopts an emic point of view rather than an etic point of view (Mondada, 2018)¹². It offers methodological tools for understanding what is treated as publicly relevant by participants in a given situation – among the potentially inexhaustible number of features of this situation describable from an external perspective – without imposing the researcher’s perspective on the data.

Second, EMCA investigates social order as a product of the local organization of participants, i.e., as continuously maintained (or modified) through their local accomplishments (Garfinkel, 1967). A corollary of this treatment of social facts as accomplished is the potential variability of identities, roles, or any socially ascribed property, from one moment to the next. That is, “contextual aspects such as setting, goals, and biographical details (age, gender, ethnicity, social class, etc.) – though often visible – are not assumed a priori but are to be discovered in participants’ visible conduct in interaction” (Nguyen & Nguyen, 2022)¹³. This analytical perspective stands in contrast with “ex-ante models of human beings” (Meyer, 2019), i.e., with intrinsic definitions of members’ identities,

¹² “The relevance of resources is locally achieved and established by the participants themselves in and for their situated action, exploiting and orienting to them as publicly available, meaningful, and providing the accountability of their actions. This constitutes the fundamental emic dimension of multimodal details, consistent with the emic view on language, action, and social interaction characteristic of EMCA: *The relevance of details is always indexical; it cannot be decided a priori and once and for all.*” (Mondada, 2018; emphasis ours)

¹³ As we discuss later (see section 3.1.2), there exist methodological divergences between EMCA and the closely related field of Member Categorisation Analysis (thereafter MCA). However, to the best of our knowledge, these divergences do not extend to the core goals and practices of EMCA and MCA.

roles (vom Lehn, 2019), or qualities (Meyer, 2019) as identical with (or necessarily stemming from) specific material properties. As we will develop in chapter 3, a crucial consequence of this “study policy” (Garfinkel, 1967) for human-robot interactions is that an entity can only be meaningfully described as “social” – or even as a “robot” – by the researcher in regard to a local setting where it is accomplished as such by members. In a stricter sense, members’ actions do not possess “an abstract, decontexted meaning which they have independent of their occurrence” (Peyrot, 1982a). Rather, if they can be said to be “intrinsically meaningful”, it is “because they unavoidably participate in an organization of activity” (Peyrot, 1982a)¹⁴.

Third, EMCA draws on the fact that members’ methods are available to study for the external observer (e.g., for the researcher working on recordings; Deppermann and Haugh, 2022) because members themselves rely on the public accessibility of these methods to coordinate with each other in situ – or, alternatively, in order to “arrive at intersubjectively available common ground” (K. Fischer, 2021). Members construct their conduct in a way that overtly displays their interpretation of the local situation (Garfinkel, 1967; vom Lehn, 2019) and, conversely, they respond to what is made visible and hearable from their co-participants’ orientation. It is, in fact, “endemic to the organization of conversation” (Schegloff, 2007) that interlocutors’ turns-at-talk display their understanding of the preceding turn (e.g., as a greeting, a question, etc.; Sacks et al., 1974) and will be examined as displaying such understanding by co-present members. Taking advantage of this state of affairs, EMCA commonly relies on the minute analysis of video or audio recordings (Heath et al., 2022) – and of their detailed transcriptions – in order to identify the fine-tuned temporal unfolding (Mondada, 2019) of participants’ actions with a much higher degree of precision than mere observation would allow (Mondada, 2019). The in-depth analysis of large collections of data can subsequently reveal recurring patterns in the procedures and methods through which social order is accomplished (Have, 2007; Pitsch & Koch, 2010).

Fourth, EMCA is uninterested in the participants themselves. As pointed out by Button and Sharrock (2016), following Schegloff (2010), EMCA’s analytical perspective focuses on the (normative) “organisation to the action and interaction that stands apart from any particular actor”: it emphasizes the “doing” rather than the “doer” (Button & Sharrock, 2016). For example, if some feature of a setting is said to be “relevant” to participants, it is only relevant in that it is “used” (or observably “attended to” by participants) as part of specific practices. Participants’ mental categorization of this setting, their perception of what “matters” in this setting, their desires, etc., are not the focus (at least, not by themselves, i.e., unless those mental states are being made publicly accessible): only participants’ practices are. It is exclusively in regard to their praxeological relevance that the EMCA researcher is interested in specific properties of a setting. In sum, EMCA will consider “as phenomena only those practices of members which are used by them to produce, accomplish, sustain, reproduce, recognize, and give account of, to and for themselves, social order” (Psathas, 1980).

¹⁴ In Peyrot’s (1982) formulation, Garfinkel and Sacks (1970) hold that there is a “primordial embeddedness of action in an a priori meaningful organization of affairs” (Peyrot, 1982). Peyrot (1982) observes this idea to be analogous to Heidegger’s (1962) suggestion “that the primordial being-in-the-world be taken as the unavoidable starting place for a philosophical inquiry, that to do otherwise led away from the essential phenomenon” (Peyrot, 1982).

2.1.2. EMCA for Human-Robot Interactions

In recent years, “human-robot interactions” have been increasingly examined from the analytical perspective described above: EMCA's common practices, interests, and concepts have been, so to speak, “transposed” to situations involving an entity pre-categorized by the researcher as a “robot”. In this endeavor, EMCA has been suited to identify what, in a robot's multimodal behavior (its talk, gestures, flashing LEDs, motor noise, sound signals – Pelikan, 2021) is treated by involved participants as *social actions*: that is, as actions which make relevant “a set of potential next actions” (Tuncer et al., 2022a). In a restricted understanding, robots' conduct becomes “social” when it is responded to (and in a certain way) by co-present participants in the following turns-at-talk. In a more encompassing understanding, EMCA's micro-analytical level of description can be leveraged to explore what emerges as pragmatically consequential when humans' and robots' actions intertwine and respond to each other (often in ways unforeseen by designers; Pelikan et al., 2020), i.e., to retrospectively unpack how humans and robots co-constructed the interactions (Pitsch et al., 2013) which ended up being captured on camera. Therefore, researchers drawing on EMCA will focus on behaviors produced by a robot that are made publicly recognizable and accountable (as an action of a certain type, e.g., an offer, a question, etc.) by co-present humans, as these humans are immersed in a situated activity, where they face specific practical problems (Heritage, 2001) – e.g., whether to respond to a robot producing a “waving gesture” at the start of an interaction and, if so, how to respond to it. In this sense, the methodological tools of EMCA can be mobilized by researchers faced with the issue of analyzing robots' and humans' micro-adjustments over time as something else than an inscrutable “black box” (Heritage, 2001)¹⁵.

However, transposing EMCA's set of tools, interests, and practices to human-robot interactions comes at the cost of some “conceptual loosening”. We discuss these methodological issues in section 2.2 (on the next-turn-proof procedure), in section 3.1.4 (on the difficulty of transposing a strictly emergentist approach to situations pre-labeled as “human-robot interactions”), and in section 8.2 (on using “human-human” frameworks to describe potentially *sui generis* “human-robot” practices).

2.1.3. Studying the “sociality” of robots besides mental representations

A result of this approach is to study the “sociality of robots” independently from mental representations. Our analytical focus will be exclusively limited to what is oriented to by participants as an “action” from the robot, i.e., the way in which some of its “behaviors” are responded to or, alternatively, what actions from the robot are accountably displayed as absent by humans. In particular, an EMCA methodology does not aim at establishing how the robot is otherwise mentally represented (Hand & Catlaw, 2019) by the participants (as a social agent, as an object, as a human, as an animal, as a new ontological category...) or if they *engage in pretense towards* it (Severson & Carlson, 2010). Unless it is made observably accountable, it makes no difference for the status occupied by the robot in the interaction that participants are “behaving as” or “behaving as if” this robot is a “subject with internal states and perceptual experiences” (Severson & Carlson, 2010). Similarly, this analytical perspective

¹⁵ See section 1.1.2 for a discussion of the tendency displayed by the field of Human-Robot Interaction to treat as a “black box” the very interactions it posits.

is independent of the set of questions related to whether participants' actions involve a mental representation of the robot as they interact with it, or if these behaviors are produced as part of a non-representational “mindless coping” with the situation (H. L. Dreyfus, 2002). In particular, we do not intend to suggest that participants discretize and categorize the stream of conduct of the robot into a preexisting list of action types (Enfield & Sidnell, 2017): the robot can be positioned interactionally as an agent without any of its behaviors being mentally constituted as “actions” by participants as they react to them. In other words, EMCA takes an agnostic stance regarding cognition (Jarske et al., 2020; Maynard, 2006) as it is interested in participants’ (humans or robots) “observable and hearable conduct [...] at the interactional surface” (Pitsch & Koch, 2010).

2.2. Next-turn-proof procedure in human-robot-interaction as a “conceptual loosening”

2.2.1. Next-turn-proof procedure and membership knowledge in human-human interactions

A potent “practical tool” (Button & Sharrock, 2016) of conversation analysis is the next-turn proof procedure. Its relevance – both for participants in situ and for the professional analyst examining a strip of interaction after the fact – was seminaly described by Sacks et al. (1974). They state:

“[s]ince it is the parties' understandings of prior turns' talk that is relevant to their construction of next turns, it is THEIR understandings that are wanted for analysis. The display of those understandings in the talk of subsequent turns affords both a resource for the analysis of prior turns and a proof procedure for professional analyses of prior turns” (Sacks et al., 1974)

Because participants’ reactions are “resources intrinsic to the data themselves” (Sacks et al., 1974), this procedure safeguards the analysts from imposing their own (etic and a posteriori) characterization of what was publicly relevant for participants as they were involved in the immediacy of a situated interaction. As such, even though conversation analysts study situations from which they are temporally and spatially distant, they can rely on the next-turn proof procedure to provide evidence “that it is that action which co-participants in the interaction took to be what was getting done, as revealed in/by the response they make to it” (Schegloff, 2007)¹⁶. This opportunity (to rely on strong evidence of what an observable conduct constitutes for involved participants) is one of the powerful means through which conversation

¹⁶ From the point of view of the involved participants, another participant’s response does not just give evidence that a previous action was of a certain type, it *transforms the situation*, so that, for all practical purposes, it is this line of action to which the participants orient as being what is going on. Yet, what we attempt to demonstrate throughout this section is that the analyst's analytical stance is not fully encapsulated by the previous sentence. That is, the ordinary practices of EMCA researchers faced with recordings of human-robot encounters *are not entirely insulated from exogenous questions as to whether the participants being studied share a form of membership knowledge beyond the immediately observable interaction – i.e., whether their conducts enact and document (and possibly negotiate, contest, re-establish, etc.) a common and taken-for-granted background knowledge.*

analysts aim to make “as little appeal as possible to intuitive judgments – they may, willy-nilly, guide research, but they are not explanations and they certainly do not circumscribe the data; the emphasis is on what can actually be found to occur, not on what one would guess would be odd (or acceptable) if it were to do so” (Levinson, 1983).

Crucially, by leveraging the next-turn proof procedure, participants' displayed understanding of other participants' conduct can be argued to make this conduct publicly available as (un)ordinary of the group that these participants are (expert) members of. The manner in which interactants respond to specific features of their co-interactants' conduct as constituting an action (e.g., asking for the salt, joking, closing the interaction, etc.) is evidence that this action is, indeed, “what is going on” (Beach & Sigman, 1995) from the perspective of involved members. Such occurrences can be treated by the researcher as *emic* windows on, e.g., what response is normatively expected in a specific setting, what is treated (or negotiated) as publicly relevant, how a specific behavior is, at this moment, categorized and named by participants, etc. That, in their next-turn, participants can make publicly accountable the immediately prior turn as atypical or, even more, as an “outright violation [...] of relevance rules” (Robinson, 2016) is of precious use to the conversational analyst. Indeed, “CA is primarily concerned with relevance rules that are intersubjectively understood by a large swath of a culture/society” (Robinson, 2016) and “only secondarily concerned with idiosyncratic relevance rules, or those shared uniquely by a single dyad, such as two relational partners” (Robinson, 2016)¹⁷. In this understanding, practices or methods typical of “life as usual” (Garfinkel, 1967) display and enact members' “membership knowledge” (Have, 2005). As Meyer (2019) formulates it, “[...] the 'documentary' character of social practices is 'incarnate' in precisely these practices. When I do or say something, I simultaneously express that 'I am confident that you understand me when I use practices that have proven successful in my interaction with you and others in situations in the past, that I judge similar to the one present'”.

2.2.2. The limited applicability of the next turn proof procedure in human-robot interactions

A set of questions arises when applying the previous considerations to interactions where a potential “participant” is, from an etic point of view, a robot. Can the robot be said to display “membership knowledge” or, at least, to form actions that “produce intelligibility” (Licoppe & Rollet, 2020; my translation) about “what is going on” in a given situation? Does a robot's conduct rely on and make public “a wide range of background assumptions, contextual knowledge, and other elements of common sense” (Clayman, 2015)? Does the answer to the previous questions vary depending on the technology being used: e.g., basic rule-based chatbots versus chatbots relying on large language models? In other words, in each specific experiment being analyzed by researchers, using more or less advanced conversational systems, can the robot's responses document the “so-called competencies involved in being a member of a collective” (Ten Have, 2016)? Indeed, if “members of society must have some shared methods that they use to mutually construct the meaningful orderliness of social

¹⁷ Of course, the mutual documentation of the action being achieved by a participant is collaboratively constructed over the course of the interaction: even in exceptional cases where explicit descriptions are produced about what a participant has been doing (Enfield & Sidnell, 2017), the nature of one's previous action can be contested or negotiated. Still, the responses being produced over the course of a conversation between human participants are crucial for the analyst to grasp what kind of actions these participants accomplished.

situations” (A. Rawls, 2003), what happens when conversation analysts can confidently assert – for example, because they programmed the robot – that the robot in question does not rely on these methods and common-sense resources? Or that, as Jarske et al. (2020) summarize, “robots are not invested in the interaction in the same manner as humans are and they do not comprehend the world of normative practices and institutional realities”? As technology evolves, these questions might be more and more tricky to answer. However, in the present day, most HRI researchers conduct experiments with robots that can hardly be said to possess the competencies stated above – unless a Wizard of Oz procedure is used, in which case the robot becomes, to some extent at least, the medium of the human who controls it.

Yet, if the robot’s actions do not produce intelligibility about what actions are being achieved by human participants, this is consequential on the argumentative solidity of conversation analysis’ findings. In Licoppe & Rollet’s (2020) terms, there is a form of “conceptual loosening” (translation ours) in applying a conversation analytic approach to interactions where an artificial agent is treated as a participant. It could be suggested that EMCA for “human-robot interactions” is, in practice, a different discipline than EMCA for “human-human interactions”. If we postulate that the robot’s reactions are not relevant information, by themselves, about what actions a co-present human has just produced, some interactional phenomena become difficult to characterize with the same degree of rigor – compared to human interactions that do not involve robots.

For example, let’s consider a fragment of a naturally occurring interaction between a Pepper robot and a group of humans, recorded in July 2022 in a French museum:

Fragment 1.1.

1. *(1.5)

hum >> *leans towards ROB

2. HUM **est-ce que tu sais danser#**
do you know how to dance

img

#img.1

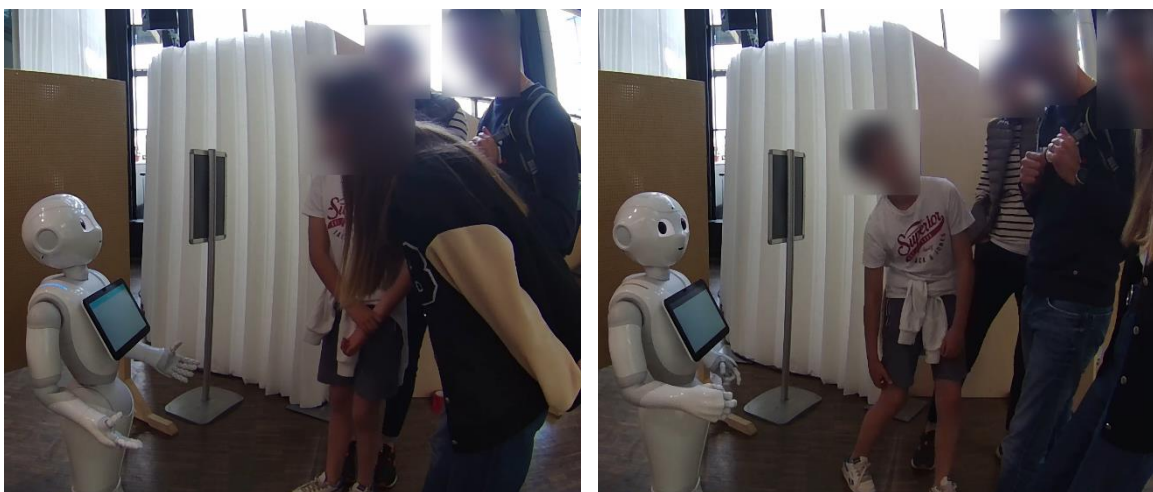


Figure 2.1. Images 1 & 2 – HUM asks ROB if it knows how to dance (img. 1), then laughs after ROB starts dancing (img. 2).

3. (0.4) * (1.9)
hum -->*

4. ROB okay! (0.7) on danse
okay! let's dance

5. (2.1) % (1.6)
rob %dances>>

6. HUM ((laughs)) #
img #img.2

In this fragment, the robot's (ROB) embodied response (L.5) can be argued to treat the human's (HUM) utterance (L.2, "do you know you to dance") as a request to dance and not as a mere factual question regarding the extent of the robot's abilities. Indeed, the robot does not produce a confirmation token to the human question (e.g., "yes I can dance") but, instead, utters an agreement token ("okay!") and, a few seconds later, effectively starts to dance. Making such distinctions between different candidate actions is a frequent concern for conversation analytic works: when studying single cases, it is a typical conversation analytic endeavor to explore, e.g., if a question has been understood "as something other than a straightforward request for information" (Hoey & Kendrick, 2017b). Repeating this micro-analytic work on numerous fragments allows the analyst to produce coherent collections of conversational practices (of formulating a complaint, a request, etc. – Sidnell, 2012).

Consequently, in the specific human-robot interaction displayed in the fragment above, can it be argued that the robot's response (L.4 & 5) effectively *constitutes evidence* that the human participants produced a request (to dance) and not a straightforward question? What does the action demonstrably "ascribed"¹⁸ (Levinson, 2012) by the robot to a participant's turn tell us about what the human "did" – i.e., what action was formed¹⁹ (Schegloff, 2007) by the human? In other words, can the robot's observable response be used as part of a demonstration that HUM's turn (L.2) was, indeed, an indirect speech act (Searle, 1975) of a certain type? Of course, to characterize HUM's turn (L.2), *we can partially rely on HUM's own – or on any of the co-present human participants' – reaction to the robot's conduct*. Human participants did not produce a third turn repair after the robot's response to HUM, nor did they produce an account of the robot's behavior (e.g., as incongruous or inadequate), etc. By agreeing to dance (L.4), the robot produced a second pair part of the type and form made conditionally relevant by HUM's first pair part, which, in turn, informs the kind of action (a request to dance) that this first pair part may have been achieving among multiple "possible actions" (Schegloff, 1996a). At this point, because humans observably treated the robot's response as such, a request to dance was what was going on for all practical purposes. However, crucially, *does the robot's response itself provide any argumentative strength to this analysis*, as part of a next-turn proof procedure? Is it a resource that the analyst can use to

¹⁸ Assuming that the concept of "action ascription" remains meaningful in the case of a robot.

¹⁹ In other terms, what multimodal resources were "designed to be, and to be recognized by recipients as, particular actions—actions like requesting, inviting, granting, complaining, agreeing, telling, noticing, rejecting, and so on—in a class of unknown size?" (Schegloff, 2007).

clarify the action produced by HUM? This question would persist if, in the previous fragment, the robot had responded to the human's turn as projecting a complaint (Sidnell, 2012), an insult, etc. As long as we make the preliminary (and technical) assumption that a robot is not a competent member, the possible responses from this robot do not constitute usable evidence of “what is going on” in the recorded interaction. In the case of the robot featured in fragment 1.1, its behavior stems from a limited number of invariable rules predefined by its developers. We argue that it does not produce direct intelligibility about the “there and now” of the situated interaction being studied.

Schegloff (1980) observes that “[w]hich use is being made of an action projection on any given occasion is something worked through by the parties in the ensuing talk”. For example, “when one says ‘What are you doin’ one might be initiating an invitation sequence, or simply asking the other for assurance that one is not eating supper or doing some other activity that would compromise a phone chat. Or one might suspect that the other was on call-waiting” (Hopper, 2005). Yet, a core issue during the analysis of a dyadic human-robot interaction is that there is, in the current state of technology²⁰, only one doubtlessly competent “party” (the human) in position to work out what was projected by an action. In light of this indeterminacy of the robot's competence, we suspect that human-robot interaction analysts are, much more often than their human-human interaction colleagues, forced to acknowledge “sequentially ambiguous” (Hopper, 2005) conducts, without having access to an “ensuing talk” between two competent parties. The situations to which is regularly confronted the HRI analyst are analogous to human-human interactions that display “actions to which there is not a clear response” (Deppermann & Haugh, 2022a) or “uptake” (Stivers & Rossano, 2010b). Naturally, in human-robot interactions, the robot sometimes *does* react to the human's conduct (in the causal sense: something about the human's conduct triggers a response from the robot). Yet, unless this robot is assumed to be a competent conversationalist, this reaction does not clarify “how are we to ascertain what action the speaker designed” (Stivers & Rossano, 2010a). Consequently, compared to common analytical practices when discussing fragments of human-human conversations, HRI researchers have one less resource at their disposal to make (rigorous) sense of humans' conduct in human-robot interactions. In the current state of robotic technologies, heavily relying on rule-based chatbots, robots' conducts can hardly be said to display an “orientation to the local situation” which sheds light on human participants' own orientation to this situation. Discussing human interactions that do not involve robots, Heritage (2018) remarks that the “next-turn proof procedure is not infallible and cannot be dispositive in all cases”. More radically, in “human-robot interactions”, if it is considered by the analyst that a robot being interacted with does not rely on a form of “membership knowledge” (simulated or not), it must be considered that this robot's reactions to humans' conduct are, by themselves, never dispositive.

2.2.3. (The absence of) membership knowledge as a working hypothesis

The previous argument rests on the following assumption: *whether or not a participant's conduct produces intelligibility about a situation cannot be decided by looking exclusively at a single recording – which entails, in turn, that whether or not next-turn proof procedure is*

²⁰ As these lines are being written, conversational agents based on *multimodal* large language models (or any multimodal generative AIs) have not been seamlessly integrated into robots. Were this to happen, this may entirely turn around the question of what is *evidenced* by the robot's responses to a human.

relevant cannot be decided by looking at this single recording alone. In other terms, we posit that researchers display a specific orientation to membership knowledge when they analyze strips of interaction. During their examination of data, whether a participant “has membership knowledge” partially functions as a working hypothesis produced by the researcher based on what is known about this participant's experience as a member of a group, or after repeated observations of the way in which this participant is treated by others – i.e., if this participant's conduct is not occasionally and fortuitously adequate (Pelikan et al., 2020; Tuncer et al., 2023) but, instead, is responded to, over the long run, as situationally relevant by co-present members. The question of whether a participant “has membership knowledge” holds, then, a mixed status in the concrete practice of the analyst: it is a preliminary assumption about a conduct that is, simultaneously, confronted to the local responses of other members to this participant's conduct. That is, *membership knowledge is, in part, conceived by the analyst as a presumed property of a participant (established beforehand through ethnographic or technical investigations), on which the researcher relies to understand to which degree this participant's reaction to a situation manifests ordinary methods and common-sense resources of the specific group being studied.* The analyst, like the involved participant, must produce a preliminary “foundational interpretation” (Lindemann, 2005; Muhle, 2015) as to whether or not the robot (or any entity) can be treated as a “You” – that is, as a social actor (Muhle, 2015) – before interpreting this robot's actions. Although a participant's status as a “member” is constantly maintained and reconfigured through its local accomplishments (Garfinkel, 1967), the analyst, in practice, unavoidably relies on their preliminary knowledge about participants (and their settings) to make sense of recorded data (Turner, 2017). In the end, this critical concern (for professional analysts and possibly for participants in situ) could be summarized as follows: Are embodied or virtual agents “transformers” in Schegloff's (2010) sense? That this, if “people talking-in-interaction are the ‘transformers’ *who take the general and formal organizations of practice [...] and realize them in and through the particulars of the moment*” (Schegloff, 2010b – emphasis mine), can it be said that artificial agents, too, “construct the talk via the generic organizations of practice as embodied in that moment's detail” (Schegloff, 2010b). We argue that, if technology continues to evolve, this question will become increasingly consequential both for humans interacting with artificial agents in situ and for the professional analysts studying them.

Compared to human-human interactions, one key difference strengthens the need to produce a preliminary assessment of the robot's competence as a member, i.e., to treat “having membership knowledge”, at first, as a pre-existing set of properties that inform the researchers during their examination of the data – rather than exclusively as a social accomplishment “intrinsic to the data themselves” (Sacks et al., 1974). Namely, the robot's internal mechanics are generally accessible to the researchers. Take, for instance, a robot that relies on a very basic conversational tree, responsive to a very limited set of inputs (e.g., a few keywords), and which, despite having a humanoid shape and eyes, does not “perceive” human gestures. Consider that the hypothetical researchers experimenting with this robot in different social settings are professional conversation analysts. These researchers happen to know that their robot is unable to see or hear most of the highly varied and fine-tuned multimodal social actions produced by the lay participants it speaks with: they have an exact understanding of both the robot's “sensory world” (what data its sensors provide and what the

robot can process from this stream of data²¹) and of the programming rules it follows. Significantly, if these researchers were to blindly follow a next-turn proof procedure, their analytical stance would occasionally become strangely dissonant with their technical knowledge about the robot's inner workings. That is, when the robot happens to be treated as a competent member by its co-participants – i.e., when its responses are treated as relevant to the task at hand²² and that they are not made accountable as breaching normative expectations –, its responses to humans will be considered (by the researchers strictly following a next-turn proof procedure) as relevant next-turn proofs of these human's actions, even though, because they have programmed the robot, these researchers are aware that, in some cases, the robot's conducts cannot be responsive (in the causal sense) to its interlocutors' immediately preceding conduct. Indeed, they know some of the responses they observe from the robot to be fortuitously relevant. This situation typically occurs, e.g., when a robot is treated as having responded to humans waving their hand (e.g., Rudaz et al., 2023; Tuncer et al., 2023) although it does not, in fact, possess the ability to recognize gestures²³. Such conducts from a robot do not respond to (and therefore, by themselves, cannot possibly inform about) the embodied conducts produced by co-present humans: are these humans greeting the robot, waving goodbye, stretching, testing the robots' abilities (Pelikan et al., 2022; Tuncer et al., 2023), etc.? Of course, that human co-participants treat these fortuitous behaviors from the robot as relevant second-pair parts is extremely significant to the analyst to clarify what is expected from the robot in a given setting and sequential context (Tuncer et al., 2023). Yet, as next-turns, the robot's behaviors do not provide “proof” of what is being accomplished in the immediately previous turn by its human interlocutor, nor of the ordinariness of those prior human conducts.

In this acceptance, “having membership knowledge” is not treated by the analyst as merely a practical accomplishment, unlike what it is to be a “social” robot, a “funny” person, a lawyer (Garfinkel, 1967), etc.²⁴ It is not an emergent phenomenon entirely independent from the robot's programming and physical features: it is, rather, a pre-existing set of properties of the robot, which function as prerequisites to produce ordinary methods in situated interactions. This set of questions regarding the local and co-constructed participation status of a robot (or a human, or an animal; Mondémé, 2016) is arguably the core of an ethnomethodological endeavor; yet, in human-robot interactions, it can be informed by a preliminary investigation aimed at grasping the robot's perspective (Pelikan, 2023). We argue that, if a robot has not been programmed with or trained on “‘common-sense’ resources, procedures, and practices through which members of a society interpret their everyday life” (Williams, 2001), its conduct cannot be considered as a producing intelligibility, with any degree of confidence, about the typical methods which human members use to coordinate themselves in daily interactions. Hence, being agnostic about the robot's inner mechanics (its programming, its sensors, possibly on which data its chatbot was trained, etc.) lowers the analyst's ability to identify whether or not the robot's conduct displays, by itself, typical methods of a “bona-fide collectivity

²¹ For example, they know that the robot can identify a human shape yet that it cannot single out the arm gestures produced by this human.

²² See chapter 6 for examples of human bystanders maintaining the robot's conduct are relevant to the task at hand in spite of interactional trouble encountered by the main speaker

²³ See also fragments 5.3 and 5.4 in this manuscript.

²⁴ In many ways, postulating that a robot is a competent “member” (i.e., with membership knowledge) functions similarly to analyzing a non-human entity (e.g., a dog) as a “participant” in a strip of interaction: analytical categories inevitably “construct the objects [they] want to bring to light” (Mondémé, 2016a).

member” (Garfinkel, 1967; Have, 2005) which are responsive to other participants' conduct (Pelikan, 2023) and to local circumstances in general.

2.2.4. Human-robot interactions and loose hermeneutics

Turner (2017) remarks that the analysis of a conversational phenomenon follows two steps. In the first one, the “sociologist inevitably trades on his members' knowledge in recognizing the activities that participants to interaction are engaged in; for example, it is by virtue of my status as a competent member that I can recurrently locate in my transcripts instances of 'the same' activity” (Turner, 2017). Then, in a second step, the sociologist “must then pose as problematic how utterances come off as recognizable unit activities” (Turner, 2017). As Have (2005) summarizes, “[w]hat Turner suggests is that ethnomethodological research is done in two phases. In the first the researcher uses his own membership knowledge to understand his materials, while in the second he analyses this understanding from a procedural perspective”. This second phase heavily relies on the next-turn-proof strategy, thanks to which “the analyst can inspect subsequent utterances to see whether these display any specific understandings of previous utterances, either by other participants, or by the original speaker himself or herself” (Have, 2005).

As we attempted to illustrate throughout this section, when applying the process described by Turner (2017) to the analysis of dyadic human-robot interactions, the second phase becomes more delicate to achieve: we hold that, with current “social” robots, only human participants' conduct is in position to “display any specific understandings of previous utterances” (Have, 2005). In other words, when it comes to deciding what actions human participants are producing, the researcher is most often left to decide “by virtue of [their] status as a competent member” (Turner, 2017) if the observable conduct of a participant can be considered as composing a particular social action. Of course, this issue is not entirely absent from the study of human-human interactions. Common practices in conversation analysis have been suspected of producing too many “intuitive characterizations of the actions embedded in turns” (Levinson, 2012), where “identification is largely based on an appeal to our knowledge as societal ‘members’ or conversational practitioners” (Levinson, 2012). Yet, when a researcher is interested in the action produced by a human interacting with a robot, interpretations of this human's potential actions are especially vulnerable to this criticism of “loose hermeneutics” (Levinson, 2012). In the case of dyadic human-robot fragments whose features are similar to those of fragment 1.1 provided above, analysts' conclusions can be (more readily) accused of being unreliable etic reconstructions of the action produced by a human participant on a recording, as these analysts cannot rely on the reactions of the robot to confirm or contradict their interpretations²⁵.

²⁵ This asymmetry – between robot and human – in producing intelligibility about the situation may be one of the reasons behind the recurring use of the verb “orient” (e.g., “the participant orients to the robot's silence”) in conversation analytic discussions of human participants' conduct in human-robot settings – probably more recurring than it is in analyses of human-human interactions (Billig, 2013). In other words, it is generally possible to identify features of the situation to which human participants orient when producing their own conduct, but much more complicated to identify what these human participants “are doing” (without a next-turn proof about how another competent participant understands these human participants' actions). The orientation of a human towards a feature of the robot's conduct as relevant requires less empirical data to be demonstrated, compared

2.2.5. The next-turn proof procedure applied to inter or intraspecies interactions

Interestingly, methodological issues with the next-turn proof procedure are also apparent in interspecies interactions (Mondémé, 2022) where, sometimes, “pets might not confirm the humans’ interpretations of their conduct in any way” (Harjunpää, 2022)²⁶. However, in contrast to human-robot interactions, these issues appear less recurring or hindering for EMCA works focused on interspecies interactions. Mondémé’s (2016) considerations on this topic are directly relevant to this comparison. She remarks that “[o]ne is a member, in the practical sense, of a community where one acts in an appropriate manner” (Mondémé, 2016; translation ours) and that this definition of a member (based on its “competencies in a given community”; Mondémé, 2016; translations ours) can be rigorously applied to animals. For example, following Laurier et al. (2005), Mondémé (2016) notes that “urban dogs deploy practical know-how to behave appropriately in parks” (translation ours). Crucially, pets, *as long-standing members of a household*, can produce interpretations of their owners’ actions (Mondémé, 2022) that rely on a background of shared experiences. Even more, she notes that animals can account for the absence of normatively expected moves from their human counterparts: “when the expected receipt is not produced by the human participant, this absence is made noticeable by the animal through diverse procedures such as repairs or expansions of the sequence” (Mondémé, 2022). In this perspective, animals are full-fledged participants enacting a next-turn proof procedure. They display and are attended to as displaying “know-how”, developed through prior interactions with humans. Their practices can be said to be intrinsically “documentary” (Garfinkel, 1967) as formulated in Meyer’s (2019) quotation in the first part of this section: “When I do or say something, I simultaneously express that ‘I am confident that you understand me when I use practices that have proven successful in my interaction with you and others in situations in the past, that I judge similar to the one present’” (Meyer, 2019).

Evidently, animals have also been examined as competent members of their own group: the next-turn proof procedure was used in studies of *intraspecies* interactions. Baboons have been analyzed as producing “greeting” sequences, during which is manifested the normative expectation that a first “greeting” must be followed by an adequate second “turn” for the recipient (Mondada & Meguerditchian, 2022). Similarly, the extension of a bonobo’s wrist towards his mother (Rossano, 2013) can be understood as accomplishing an invitation or a request, which is made evident when this “first pair part” is, indeed, oriented to by the mother as a request in her subsequent response (Mondada & Meguerditchian, 2022). The conduct of each participant – what action a participant observably ascribes to the other – is, then, a resource for the analyst. As members, these animals’ actions produce intelligibility about the “(normative) expectations connected to the organization of the sequence” (Mondada & Meguerditchian, 2022) and, more generally, about “what is going on”: they make apparent those features which are currently relevant to local participants’ ongoing activity, among all the candidate features conceivable by the human observer.

to the characterization of the exact action that this human is doing when responding to this specific conduct from the robot.

²⁶ Harjunpää (2022) notes that these methodological issues with the next-turn proof procedure are “not only a problem for the analyst but also for the human participants. How do they interpret the signs produced by non-human companions and orient to the interspecies social actions as intelligible and accountable?”.

2.2.6. Remaining grounds to produce demonstrable interpretations of human-robot interactions

Considering the difficulties exposed previously, and without directly using the robot's behavior as part of a next-turn proof procedure, how can an analyst make sense of some especially intricate conducts displayed by human participants in human-robot settings? Two possibilities stand out:

- A) To absolutely avoid producing interpretations of human participants' actions, unless these actions are described, accounted for, or, more generally, observably oriented to as actions of a certain type by these human participants themselves in the following talk or by a co-present human participant. This solution implies limiting descriptions of human actions to cases where the main speaker or another human (third-party, co-participant, bystander, etc.) can produce conducts providing evidence ("next-turn proofs" or not) as to which action has been achieved by the current speaker. However, descriptions of other participants' actions are rare (Enfield & Sidnell, 2017) and cannot be found in dyadic interactions between a single human participant and a robot – unless it is produced by the robot, which leads us back to the issue stated throughout this section. Verbal characterizations of one's actions in dyadic interactions with a machine are equally scarce: this is one of the arguments behind Suchman's (1987) focus on collaborative tasks (rather than solo participants) in her work on the use of copy machines²⁷. Even more, participants' description or explicit categorization of their own actions "does not guarantee that this is the action they will be understood to have done" (Sidnell & Enfield, 2017).
- B) To rely on:
 - a. The analyst's own competence and mastery "of the natural language in play" (Macbeth, 2018) as a member
 - b. The construction and temporality of participants' verbal and embodied turns
 - c. Demonstrable similarities between observed participants' conducts and practices documented in conversation analytic studies

If the first item in this list (a.) helps the analyst in "recognizing the activities that participants to interaction are engaged in" (Turner, 2017), the last two (b., c.) are fit to provide evidence that a given conduct composed "a possible X" (Schegloff, 1996c) for the interactants in situ²⁸. This approach to human-robot interactions is deprived of

²⁷ According to Theureau (2004), Suchman approached "describability as a property of the practical activities of social actors in the process of being accomplished" (translation ours). Because they were instructed to collaborate, participants observably indexed or described their mutual conduct or their interpretation of the activity at hand: "In the interest of the collaboration, each makes available to the other what she believes to be going on: what the task is, how it is to be accomplished, what has already been done and what remains, rationales for this way of proceeding over that, and so forth. Through the ways in which each collaborator works to provide her sense of what is going on to the other, she provides that sense to the researcher as well." (Suchman, 1987)

²⁸ Schegloff (1996a) notes: "[...] to describe some utterance, for example, as 'a possible invitation' [...] or 'a possible complaint' [...] is to claim that there is a describable practice of talk-in-interaction which is usable to do recognizable invitations or complaints (a claim which can be documented by exemplars of exchanges in which such utterances were so recognized by their recipients), and that the utterance now being described can be understood to have been produced by such a practice, and is thus analyzable as an invitation or as a complaint. This claim is made, and can be defended, independent of whether the actual recipient on this occasion has

some resources typically available to “human-human” conversation analysts to recognize and provide evidence of what is being accomplished by a participant. Yet, because it relies on “members’ competence for the understanding of members’ competence” (Macbeth, 2018), it draws instead on “the very premise of ethnomethodology’s program” (Macbeth, 2018). This is the method we used in each of the studies presented in this work, as long as some evidence was available (e.g., in the turn construction) regarding what action a participant had produced. This analytical stance embraces the “soft underbelly” (Levinson, 2012) inherent to an EMCA approach to human-robot interactions. Indeed, even in human-human interactions, the absence of a total reliance on a next-turn proof procedure is regularly required to produce intelligibility about some properties of the talk being done. For example, Schegloff (1996a) denounces the view “according to which a recipient’s understanding of some utterance is definitive of its import and the utterance itself has no ‘objective’ import”. This view, he argues, subverts “the possibility of analytically specifiable ‘misunderstanding’; for if a recipient’s understanding is definitive, what leverage is there for claiming it to be a misunderstanding?” (Schegloff, 1996a).²⁹

This last remark from Schegloff applies to the characterization of humans’ conduct in human-robot interactions. If the robot’s observable treatment of each turn produced by humans was to be taken as definitive proof that “this is what is being done by this human” (unless the human treats this robot’s response as troublesome or explicitly verbalizes it as a misunderstanding), the claims produced by analysts, in addition to being regularly incongruous, would be oblivious to what the construction of humans’ turns can tell us – about what action is being achieved through them³⁰. For example, in our collection of recordings, there are many occurrences of robots treating as “questions” turns that were constructed (in terms of prosody and grammatical structure) as “assertions” by human participants. Regularly, these robots stand uncorrected in the following talk. In these occurrences, the robot’s behavior is not made accountable as a breach of normative expectations by participants. When

treated it as an invitation or not, and independent of whether the speaker can be shown to have produced it for recognition as such on this occasion. *Such an analytic stance is required to provide resources for accounts of ‘failures’ to recognize an utterance as an invitation or complaint, for in order to claim that a recipient failed to recognize it as such or respond to it as such, one must be able to show that it was recognizable as such [...]* (emphasis ours).

²⁹ A congruent idea can be found in Button & Sharrock’s (2016) characterization of the next-turn proof as a “practical tool”. They argue that “[t]he ‘proof’ procedure is not a means for initially establishing the identity of a turn in conversation; it only supplies further confirmation of a determination that the analyst has already made as to what a given utterance might be – if, for example, the analyst figures some utterance is a question and, apparently, fellow conversationalists respond by engaging in answering, this shows that the identification does not manifest the analyst’s idiosyncratic reasoning but, seemingly, the same kind of reasoning on the part of the conversationalists whose talk is being analysed.” (Button & Sharrock, 2016)

³⁰ As a consequence, we argue that occurrences of humans orienting to the robot as having misunderstood their action (third position repair) – and, to a lower extent, occurrences of humans producing commentaries, assessments, descriptions or, more generally, any form of account about the action that the robot just produced or about their own immediately prior action – are especially enlightening in the study of human-robot encounters. Although “participants can always post hoc manipulate the status of their prior actions in line with their current goals and aims” (Stevanovic, 2023), these fragments provide the analyst with practices that display additional information as to “what is going on” in a strip of interaction – thereby reducing (or delaying) the need to take a strong methodological stance on the robot’s competence as a member of society. These recordings allow the analyst to develop analytical insights while still relying on a weak form of methodological safeguard (human participants’ accounts of their or the robot’s actions).

analyzing these fragments, if the analyst's own competence is to be put entirely aside, nothing can be said about the discrepancy between the robot's response and the action ordinarily produced by the human's preceding conduct.

Finally, what EMCA loses of the intelligibility of a situation when a participant is a robot, it can partially compensate thanks to the explainability of the robot's conduct. Indeed, there exists an (additional) set of objections against the claim of ethnomethodology (and, on different grounds, conversation analysis³¹) to merely capture the practical logic of the members involved in an activity, i.e., objections against ethnomethodology's claim to be a reproductive rather than a productive "understanding of human practice" (K.-M. Kim, 1999). EMCA "applied" to HRI is less vulnerable to this risk of producing a "hermeneutic reconstruction of the practical logic of the members" (K.-M. Kim, 1999) regarding one participant: the robot, whose internal operations are normally accessible to some degree to the analyst – depending on the technology being used. In our case, each experiment described throughout this work was conducted using a Pepper robot that featured a rule-based chatbot running locally. Given this robot's programming, it was generally possible to provide a technical answer (or a set of probable causes) for the observed conduct of the robot, as well as to collect logs about what the robot perceived from its external environment, how it discretized specific features of this environment (if it perceived an object or a human, if it categorized this human as "available to interact" or as "unavailable", etc.) and with which degree of confidence. "Why that now?", a core question of conversation analysts, and a critical concern for conversationalists *in situ* (Schegloff & Sacks, 1973a) was therefore possible to answer with additional certainty. In fact, as long as a robot does not function as a black box, a conversation analyst producing an *a posteriori* account of this robot's conduct can generally rely on more tools to answer this question of "why that now?" (from the robot's perspective), compared to human participants who were directly involved in the local interaction with this robot. Because the analyst can gather data about the robot's sensors and look at its code, both the robot's *umwelt*³² (i.e., its perceptual environment; Mondémé, 2016; Uexküll, 1982; Von Uexküll, 1988) and its internal logic are accessible. That is, with the type of robot we used, it was most often possible for the analyst to establish, not only what the robot was treated to be doing by its co-participants on

³¹ See, e.g., Lynch (2019) for a discussion of some objections to Conversation Analysis' claims. Although it is not typically portrayed as such, one could also count Enfield & Sidnell's (2017) work as criticism of the ability of common conversation analytic practices to reveal the practical reasoning or the "underlying ontology of action" (Enfield & Sidnell, 2017) of participants *in situ*. They remark that discretizing and "binning" a conduct produced by an interlocutor (as a specific action) is, more often than not, a theoretical problem for the analyst rather than a practical concern for the members immersed in a situated interaction. That a participant responded to an interlocutor's conduct "as" a specific action (e.g., as what may appear to the analyst as an "invitation") does not imply that the interlocutor's conduct was, in fact, categorized by this participant. Indeed, "responding to an action is not equivalent to describing it, nor does it entail that a description of it has been made at any level" (Enfield & Sidnell, 2017).

³² The question of whether a robot can have an *umwelt* remains complex and unresolved: see Emmeche (2001) and Nöth (2001) for a summary of this discussion. We focus here solely on the minimalist and restrictive definition of an *umwelt*. In this restrictive definition, "[a]ll it takes to constitute an *Umwelt* in the sense of a phenomenal experienced species-specific (or 'devicespecific') world is a certain circular information-based relation between sensor devices and motor devices as described by the notion of a functional circle." (Emmeche, 2001). That is, an engineer using Pepper has access to what information the robot gathers and processes through its sensors and algorithm and how, in turn, this information leads to the activation of specific devices on the robot (motors, speakers, screen, LEDs) through the application of rules (in the programming sense of "rules").

the video recordings, but “why” the robot produced a specific conduct and “what” it perceived and discretized from the world.

2.3. “Applied CA” and Good Old-Fashioned AI: Turning rule-based robots into expert conversationalists, a (long-documented) hopeless endeavor?

Loid: I created a conversational flow chart based on what kind of person she is. First, you're going to ask her question one. For example, about her favorite food. Yes. From there, I wrote down her most probable answers, A, B, C, and D. Then I added the best response you can give to each choice, and then I added her possible responses – again, as A, B, C, and D – along with your next response. Just memorize this chart, go with the flow, and choose your answers appropriately, and you're sure to –

Franky: How am I supposed to do that, you dumbass?! That's tens of thousands of possibilities! What the hell is “appropriate,” anyway?

[...]

Loid: Then you're just going to have to practice having conversations.

—*Spy Family*, Season 1, Episode 16

“Sing while you're playing” (Sudnow, 1993)

2.3.1. “Actualizing rules” in situ

Two major issues have long been documented in programming exclusively rule-based agents to “converse” with human participants – and, more generally, for these rule-based agents to imitate human-level expertise in any domain. A first difficulty pertains to what Coulon (1993) terms the “actualization” of rules in situ. That is, the challenge of specifying the members' methods or background knowledge through which context-free rules or instructions can be applied locally. By definition, such rules or instructions cannot fully specify their conditions of application (Wittgenstein, 1953)³³ and always rely on an “etcetera” clause (Garfinkel, 1996). Consequently, for “applied CA” (Antaki, 2011a; Ten Have, 2007)³⁴ dealing with artificial

³³ Congruently, Kim (1999) glosses ethnomethodology's main activity as follows: “Taking seriously Ludwig Wittgenstein's dictum that ‘a rule does not contain its own applications’, ethnomethodology attempts to delineate the process of how actors ‘manage’ the indeterminacy and contingencies necessitated by the applications of such implicit rules to a wide variety of concrete interactional settings. And, in this sense, social organization of action is conceived as an ‘ongoing accomplishment’ of social actors.” (K.-M. Kim, 1999)

³⁴ In the terminology used by Antaki (2011b) we refer to “applied CA” in its “interventionist” sense: “where CA can be applied to a practical problem as it plays out in interaction, with the intention of bringing about some sort of change” (Antaki, 2011b). It is also in this sense that these words are used by Antaki (2011b) and in the

agents, the core of the problem does not lie in identifying the (conversational) rules that humans orient to. Rather, its whole challenge is to *bridge the gap* between rules, norms or instructions and their local application (Coulon, 1993). Indeed, simple rule-based programs cannot reproduce the “subdued phenomenology” (Lynch, 2000) behind “some of the most ubiquitous technical features” (Lynch, 2000) of the results from conversation analytic approaches. Conversation analysis does not offer a fully formalized system – exportable in any local context: to be meaningful, its results still require a “phenomenological orientation” (Lynch, 2000). Lynch (2000) takes the example of an “object” such as a turn-transition place:

“Although this 'object' [a turn-transition place] is described as a feature of a context-free rules set [...], *it is impossible to identify a particular 'turn-transition relevance place' by consulting context- or content-free specifications.* [...] Syntactic and narrative structures may be relevant for locating a possible ending of an utterance on the floor, but the relevance of any such analysis depends on locally situated, normative judgments about the sort of *action* that is (or could be) taking place.” (Lynch, 2000; emphasis mine).

This is significant for the EMCA-inspired roboticist, as the implementation of EMCA findings cannot “import” their “subdued phenomenology” on a rule-based robot. Deriving “rules”, interaction flows, guidelines, etc. from EMCA observations inescapably peels off the situatedness of these findings. Pitsch (2016) synthesizes the previous issue as follows: any formalization is a “transformation that changes the phenomenon itself” (Pitsch, 2016). That is, “rule-based approaches to discourse modeling stand in direct conceptual contrast to the openness and unpredictability of social interaction, and it is unclear on what grounds a technical system can select an appropriate and relevant subsequent action” (Pitsch, 2016)³⁵. As Pitsch (2016) notes, this issue is a particular case of what Suchman (1987) coined as the tension between “plans” and “situated action”. From the point of view (or the code) of rule-based artificial agents, there is no distinction between “situated action” and “plans”³⁶. If “rules leave loop-holes open, and the practice has to speak for itself” (Wittgenstein, 1969), then, there is no “practice” that can fill these loopholes for a rule-based robot.

These brief remarks take place within a significant body of literature, produced over the last 60 years (Coeckelbergh, 2019; H. L. Dreyfus, 1992; Mitchell, 2021), that approaches the issue for AI of relying on the “background knowledge” (H. L. Dreyfus, 1992) on which human conducts are built. The specifics of this debate (and whether technologies like generative artificial intelligence fall under this criticism) are beyond the scope of this section as we focus on the case, arguably much less intricate, of robots relying on rule-based chatbots. Yet, its central argument remains crucial for the EMCA analyst attempting to derive interaction designs from their empirical findings: a form of background knowledge is indispensable for members of a society to “apply” rules, norms, and instructions in situ.

various contributions featured in Antaki (2011a). To the best of our knowledge, this is the currently accepted definition of “applied CA”.

³⁵ Besides rule-based conversational agents, this issue can also be encountered when using machine learning to identify specific interactional phenomena like “closings” and “pre-closings” (Rollet & Clavel, 2020; Schegloff & Sacks, 1973b). In Rollet & Clavel's (2020) formulation, there is a “tension between the principle of describing the uniqueness of cases – defining the analytical mentality of Conversation Analysis – and the requirement of generalization for the training of automatic models.” (Rollet & Clavel, 2020)

³⁶ The tension between plans and situated actions is itself connected to the “computo-representationalist paradigm” (Relieu et al., 2004), initially identified in the setting of a technology company by Suchman (1987).

2.3.2. Following rules and orienting to rules

A second difficulty for the EMCA-inspired designer working on rule-based robots is that an artificial conversational agent encoded with rules (i.e., the programming sense of “applying a rule”) does not “follow” normative rules as human conversationalists do (the ethnomethodological sense of “applying a rule”). Indeed, a fundamental position held by ethnomethodologists and conversation analysts is that rules or norms are resources (Button, 1990; Garfinkel, 1967; J. J. Turowetz & Maynard, 2010)³⁷. That is, participants do not blindly follow social norms or written rules (if those are locally relevant), nor do they act entirely independently of them. They display practical mastery of the norms or rules and can *orient to them* (Meyer, 2019) as part of their ongoing activity – as they could orient to any other locally relevant feature of the situation.

This orientation to rules and norms is observable in the way in which participants generate their conduct. For example, they may produce a speaking turn whose construction observably indexes which norms are being breached by its production: when apologizing while interrupting the current speaker’s turn (Weatherall & Edmonds, 2018), when refusing an invitation (Bilmes, 2014), etc. In this understanding, norms and rules are only important as long as they are used or invoked by participants in situ (Emirbayer & Maynard, 2011). Doing so, ethnomethodology avoids sociological explanations exclusively based on a “hidden order” (Livingston, 2017). Although members of a society orient to norms, they are not “cultural dopes” that mechanically apply those³⁸.

Yet, this ethnomethodological observation gets turned on its head when robots using rule-based chatbots are being programmed³⁹. Rule-based chatbots do not “orient” to rules in the sense where humans do. From an internalist point of view (i.e., no matter how robots are treated by co-present participants in situ), rules and norms are not “resources” (Garfinkel, 1967; J. J. Turowetz & Maynard, 2010) nor a “guide for action” (Dupret, 2014) for the rule-based agent. The robot has “rules” in a mechanical sense, it does not mobilize them as a *resource* for all practical purposes⁴⁰. Button (1990) specifies this conceptual difference between “rules” as applied by rule-based artificial agents and “rules” used by humans in situated interaction. He distinguishes “rules that people can be shown to orient to, and rules that are said to be an interior mental machinery. On the latter understanding, rules stand behind action; on the former, rules are embedded within action” (Button, 1990). Hence, *what EMCA uncovers (normative rules which are oriented to by humans) is not what is, then, implemented into simple rule-based robots (“if;then” rules)*. In this understanding, transferring

³⁷ “[T]he rule set which is posited in [Sacks et al. (1974)] does not lie behind the actions of constructing a turn, allocating a turn, and co-ordinating speaker transfer and thereby causing things to happen. We do not get a turn because of the rules. Rather, the way in which a turn is taken displays an orientation to the rule. A rule is followed as part of accomplishing the action. The sense of rule here, is then, part of the logical grammar of the action. The rule does not precede the action. Rather the rule is discoverable in the action” (Button, 1990).

³⁸ “Rules, to repeat, are features of actions rather than explanations for them.” (J. J. Turowetz & Maynard, 2010).

³⁹ The issue may present itself differently with conversational chatbots relying on other types of technologies, e.g., those based on large language models. See section 2.2 for a discussion of those types of robots as potentially “competent” actors and the implications of such a statement for EMCA studies of human-robot interactions.

⁴⁰ To push the analogy to its extreme, rule-based artificial agents can be said to be the most extreme imaginable form of “cultural dopes” (Garfinkel, 1967). Additionally, still from an internalist point of view, these artificial agents can hardly be said to “experience” rules and norms as a “normative pressure” (Heritage, 2013b).

practical habits into formal rules (which are followed in the programming sense of “following a rule”, i.e., not as an orientation) inevitably “loses the phenomenon” (Garfinkel, 1996).

2.3.3. What can we salvage from practical “coping” when doing rule-based design?

In light of the previous considerations, what exactly are we doing when we attempt to derive interaction designs for artificial agents from EMCA findings? What is the underlying premise of “applied CA” (Antaki, 2011a; Ten Have, 2007) when its results will ultimately serve to create rule-based behaviors on a robot? A brief phenomenological detour may help to underline this premise. We base our argument on a large section of the phenomenological literature which holds that our everyday mode of encountering the world occurs in the form of a “mindless coping” (H. L. Dreyfus, 1990, 2001; Heidegger, 1996) and/or that expert practitioners do not rely on symbolic representations of their activity as they are absorbed in it (Cappuccio, 2023; Coeckelbergh, 2019; H. L. Dreyfus, 2014; S. E. Dreyfus & Dreyfus, 1980; Hoffding, 2014).

For the professional analyst, relevant features of a strip of interaction are encountered as distant objects, on a video recording that can be paused and replayed at will. That is, the relevant features of these strips of interactions are not encountered as a set of local expectancies in reaction to which an expert conversationalist is “drawn” (H. L. Dreyfus, 2001) to respond as if they occurred in real time⁴¹. Analysts are generally not tempted to utter “hello” when they observe a participant produce a greeting on a video recording. Crucially, in these conditions, removed from the situated activities that participants were accomplishing on the recording, the analyst can *describe* an action (e.g., by explicitly categorizing and labeling it as an action of a certain type) rather than merely *respond* to it (Enfield & Sidnell, 2017). That is, the analyst can attempt to verbalize “silent knowledge” (Hirschauer, 2007), i.e., to articulate “phenomena that are not yet available in linguistic form in the first place” (Meier Zu Verl & Meyer, 2024). In a different vocabulary, relevant features of strips of interaction are intentionally turned into a “present-at-hand” (Heidegger, 1996) mode of being (at least, as much as they can be) for the sake of uncovering and verbalizing them.

Provided that the description above is correct, EMCA-inspired designs derive formal rules or action types (Enfield & Sidnell, 2017) from conducts that emerged within an absorbed mode of “coping” with the world (H. L. Dreyfus, 2013; Hoffding, 2014) or from a preconceptual “subprehension” (Enfield & Sidnell, 2021) of the appropriate response to co-participants' conducts. This endeavor assumes that those occupying the role of the “academic” (Schegloff, 1988) or “professional” (Schegloff, 1996b) analysts – those who examine social action at a distance, detached from the urgency (K.-M. Kim, 1999) of the expectancies that participants were experiencing in situ; those who step back and change their activity “to that of inquiring into the methods used rather than using the methods” (Psathas, 1980) – can nevertheless produce descriptions or models of practical activities that can be, then, turned into viable designs. In this understanding, “applied CA” constitutes an unavoidable case of “calculative rationality” (H. L. Dreyfus, 1987).

⁴¹ For example, the analyst will not “respond” to a video recording of a greeting sequence by producing multimodal behaviors (gestures, speech, gaze fixations), themselves experienced as a meaningful gestalt rather than as discrete actions.

2.3.4. Expert conversationalists and rules

Because of its endeavor to go from practical coping to rule-based design, “applied CA” dealing with rule-based agents stands partially at odds with phenomenological accounts of experts’ practice (such as that of ordinary expert conversationalists) as non-rule governed. Notably, building on a Heideggerian perspective, Dreyfus & Dreyfus (1980) propose a hierarchy of experience levels based on the observation that the core of experts’ activity, in any given domain, cannot be encapsulated by rules (S. E. Dreyfus, 2004, 2014), while, on the contrary, a defining property of novices is their straightforward application of context-free rules. Very briefly summarized, Dreyfus & Dreyfus (1980) oppose the beginner chess player who follows a “few simple rules” to the chess expert to whom “an appropriate move or tactical idea presents itself” (S. E. Dreyfus & Dreyfus, 1980) when they observe a particular arrangement of pieces on the chessboard – or to the expert speaker for whom “situations simply elicit [...] appropriate linguistic responses” (S. E. Dreyfus & Dreyfus, 1980)⁴². Bourdieu (1977) distinctly formulates this relationship between beginner status and the reliance on an ex-ante representation of a “repertoire of rules”:

“So long as he remains unaware of the limits inherent in his point of view on the object, the anthropologist is condemned to adopt unwittingly for his own use the representation of action which is forced on agents or groups when they lack practical mastery of a highly valued competence and have to provide themselves with an explicit and at least semi-formalized substitute for it in the form of a repertoire of rules, or of what sociologists consider, at best, as a “role”, i.e. a predetermined set of discourses and actions appropriate to a particular “stage-part”. It is significant that “culture” is sometimes described as a *map*; *it is the analogy which occurs to an outsider who has to find his way around in a foreign landscape and who compensates for his lack of practical mastery, the prerogative of the native, by the use of a model of all possible routes*. The gulf between this potential, abstract space, devoid of landmarks or any privileged centre - like genealogies, in which the ego is as unreal as the starting-point in a Cartesian space - and *the practical space of journeys actually made, or rather of journeys actually being made, can be seen from the difficulty we have in recognizing familiar routes on a map or town-plan until we are able to bring together the axes of the field of potentialities and the “system of axes linked unalterably to our bodies, and carried about with us wherever we go”, as Poincaré puts it, which structures practical space into right and left, up and down, in front and behind.*” (Bourdieu, 1977, p. 2; emphasis mine)

Although we are mainly interested in rules and norms that conversationalists observably orient to, these phenomenological considerations extend to any form of expertise. Among other examples, *ethical* expertise has been analyzed as a situated activity that does not rely on the application of a general set of rules to particular cases (H. L. Dreyfus, 2005;

⁴² As mentioned before, the “rule following” of the robot is not the “rule following” of the human. However, if we limit ourselves to S. E. Dreyfus & Dreyfus’ (1980) definition, a robot relying on a simple context-free interaction flows is, analogically, the very essence of a novice that follows a limited number of context-free rules and orients to a set of “non-situational” (S. E. Dreyfus & Dreyfus, 1980) features: its scripted behavior is determined by systematic rules produced ex-ante (e.g., to say “oh sorry” when it starts speaking at the same time as its interlocutor), based on the designers’ representations about what social situations this robot may face (Rollet & Clavel, 2020).

Stichter, 2016). Examining the Aristotelian concept of phronesis as practical wisdom (Reeve, 1992) – produced by expert social actors (Pickup, 2016) as a reaction without deliberation (H. L. Dreyfus, 2005) to the particulars of a situation – Polak & Krzanowski (2020) note that Aristotelian ethics would not be directly transposable on a rule-based robot. They state:

- “1. Phronetic ascend cannot be encapsulated in a set of rules, because it deals with the specificity of particular cases (i.e., it is not procedural).
2. Phronetic ascend is case-specific (i.e., the focus of the phronetic decision is a particular case).
3. The decision for the specific case is not achieved through logical analysis (which originally meant Aristotelian logic) or reasoning through principles (like in science) but rather through intuitively grasping the outcome or having 'foresight of consequences.’” (Polak & Krzanowski, 2020)“

2.3.5. Some examples of EMCA-inspired interaction flows or conversational guidelines

To ground in concrete practices this disjunction between phenomenological accounts of expert practice and EMCA-inspired designs for a rule-based robot, I will list some ideas of rule-based designs which may be classified as resulting from “applied CA” (Antaki, 2011a; Ten Have, 2007). These candidate designs were considered, discussed, and sometimes explored with my colleagues, formally or informally, over the course of my PhD. They were either based on our own (more or less refined) examinations of human-human interactions, or directly inspired from existing EMCA papers. Sometimes, the detailed technical implementation of these ideas had been thought through, often, not at all. For the sake of conciseness, I will merely skim through them.

1. When the robot detects someone speaking in overlap with its current turn, it must *speed up its rate of speech*. The intent is to simulate a “rushthrough” (Walker, 2003) that will be interpreted as a form of turn-holding.
2. When the robot detects someone speaking in overlap with its current turn, it must *repeat the last word it uttered*. As in the case above, this “self-repetition” (H. Kim, 2002) intends to convey that the robot “attempts” to hold its turn.
3. When the robot detects that it started speaking at the same time as its interlocutor, it must *stop speaking and say “oh sorry, go on”*. The intent is to manifest that the robot “recognizes” and accounts for an ongoing “simultaneous start” (Sacks et al., 1974; Schegloff, 2000).
4. When the robot is able to respond to a human’s turn before this human finishes their speaking turn, the robot must occasionally *raise its finger and produce pre-recorded vocal fillers or bricolage turn beginnings* (Gardner, 2007). The intent is that the robot “signals” it is “trying” to take the turn.
5. After the robot is approached by a human, and assuming that it is equipped with the adequate detection algorithm, *the robot congratulates the human on their hat, if they*

are wearing any. The intention is to roughly simulate a basic form of “registering” as described by Pillet-Shore (2008)⁴³

A few supplementary design ideas could not be captured by a small set of rules. Provided that we ultimately had to work with rather simple rule-based chatbots and dialog trees (rather than more configurable types of conversational technologies), these additional ideas were, instead, discussed as guidelines when designing such dialog trees for the robot: they were candidate principles or rules of thumb for the “Conversational UX designers” (Moore & Arar, 2018) that scripted the verbal and gestural conducts of the robot in response to a given input. For example:

6. In cases where the robot must request assistance from a human co-participant, its turn construction should reflect its entitlement to ask such a service (as evaluated by the person writing the script) and the obstacles or contingencies associated with accomplishing what it requests (also evaluated by the script writer). For example, based on Drew (2017), the robot should rather use imperatives (“pass me the charger”) for “high entitlement/low contingency” requests, while it should use conditional forms (“I wonder if...”) for “high contingency/low entitlement” requests (Drew, 2017). For requests that rest at the middle of this spectrum, the use of modals (could you...) is preferred. Naturally, this is merely a guideline for script writers – who, on a practical level, may already “know how” to produce an appropriate request in their everyday interactions.
7. Although they are not detailed here, additional design ideas or guidelines were found in Moore & Arar (2018), who provide various examples of typical conversational practices to leverage. Significantly, in the view of these authors, conversational designers “must be able to articulate the mechanics of human conversation so they can design it, instead of simply knowing it tacitly like everyone does” (Moore & Arar, 2018). For instance, “a conversation expert may describe the function of the word 'oh' to mark speakers' realizations [(Heritage, 1985)] or how the phrase, 'to the what?,' in response to 'I'm going to the workshop,' elegantly elicits a repeat of a single word 'workshop' [(Schegloff et al., 1977)]. Conversational UX designers use such observable patterns of the machinery of human conversation in building conversational machines.” (Moore & Arar, 2018).

Crucially, two properties of the examples of rules or guidelines provided above stand out:

- A. Their triggering conditions rely on very limited criteria: e.g., “speaking” and “hearing someone speak” will trigger an acceleration in the robot's rate of speech. This holds true for examples 1 to 5.

⁴³ Pillet-Shore (2008) observes: “During co-present openings, speakers can deliver utterances through which they interactionally register some feature of the setting that they are ‘just now’ starting to share with their interlocutors – features that are available for mutual perception and experience. Through these registrations, speakers display themselves to be attending to the selected features, inviting or directing co-participants to also attend to them, thereby proffering these features for sequence expansion” (D. M. Pillet-Shore, 2008). Note that a comparable intent lies behind the system created by Glas et al. (2017). Their end-goal was a robot that produces “personalized greetings” when it “observes a novel feature, such as a new hairstyle, or a consistent behavior, such as visiting every afternoon” (Glas et al., 2017).

And/Or

- B. They require human expertise (or, possibly, highly evolved multimodal AI systems) to *recognize* the context in which specific conducts could be relevantly produced. This is the case for example 6. Like the “transition relevance place” in Lynch's (2000) interpretation, identifying the adequate situation in which to produce these typical conversational practices requires “locally situated normative judgements about the sort of *action* that is (or could be) taking place” (Lynch, 2000).

2.3.6. “Applied CA”: Making rule-based robots into better novices?

Overall, as has been argued for over 50 years about rule-based AIs (H. L. Dreyfus, 1992; Mitchell, 2021), exclusively rule-based robots cannot function similarly to human experts. Attempts to specify the practical logic of involved participants can only get us so far. Unless another technology is used (e.g., possibly, multimodal large language models or any form of generative AI), the EMCA-inspired designers cannot turn the robot into an expert but can merely hope to make it *less* of a (metaphorical) novice. The reconstruction of the practical activity of conversationalists may create “convenient heuristics” (Enfield & Sidnell, 2017) but it cannot⁴⁴ provide the robots with the means to function as a conversational expert or, in another vocabulary, as an ordinary member of society (Garfinkel, 2002). In this situation, speaking about “EMCA-informed designs” is to refer to the humble and limited activity of increasing the number of context-free rules that will be applied by a robot (rather than expertly oriented to).

In light of these challenges, a relatively common stance (Ten Have, 2007) – which was adopted throughout this research – is that these constraints must be embraced. If the robot cannot function as an expert conversationalist, then, we might as well provide it with as many rules as possible – based on what it can grasp from its environment. A realistic goal for an EMCA-inspired design approach is therefore to narrow the gap (Coulon, 1993) between the robot's *plan* and the *emergent local practices* of its human interactants. This endeavor is precisely what Ten Have (2007) suggests for “applied CA” beyond artificial agents (e.g., attempts to “improve” doctors' practices during medical interviews):

“I have stressed that a CA study (or an ethnomethodological study generally) is committed to elucidate the local logic, the emic rationality, of situated practices. One way in which one could conceive of 'applied CA' in such fields would be to try to influence the plans for such practices, in the sense of *making them more respectful of the local rationalities one has discovered.*” (Ten Have, 2007, emphasis mine).

In this understanding, although its findings are applied to robots which (from an internalist point of view) function differently from human experts, EMCA-oriented designs are more aligned with these experts' ordinary methods – compared to non-EMCA-oriented designs. As ex-ante attempts to imitate what is originally encountered as a local and often preconceptual

⁴⁴ Following Leiter (1996), Heidegger or Bourdieu both denied the possibility of fully formulating practical activities in language: “at the foundation of this field of intelligibility are mindless coping skills that resist explicit articulation in propositional form”. In this view, the issues connected to EMCA-inspired designs can be mitigated but not eliminated.

orientation to “norms as resources”, these designs are, at least, produced while considering *more* of the common-sense resources and procedures that conversationalists ordinarily use and orient to. Returning to the analogy that Bourdieu draws with the paradigmatic problem of the map and the territory: *applied CA strives to draw better maps*. To recognize that the map cannot be the territory does not erase the obvious usefulness of maps. An expert may always operate better without following rules, or without consulting a map during the most hectic phases of their activity; yet, this does not invalidate the endeavor to continuously sketch more accurate maps (for the level of description required in specific settings) – or, in our case, to adequately capture the rules that expert conversationalists orient to during social activities. If exclusively rule-based robots are doomed to never outperform novices as conversationalists, they may at least be turned into better novices. Thus, in light of these largely inescapable difficulties for rule-based agents, a more practical question can legitimately come to the forefront: how can EMCA findings be leveraged to design a rule-based robot whose conduct is treated as relevant by co-present humans – regardless of its internal functioning.

3. AN EMERGENTIST DEFINITION OF SOCIAL AGENCY

Those to whom the principle of reciprocity of perspectives is extended and who are thereby included in the proto-moral universe of discourse are constituted as individual personae. In the same way, the-status of “being a person” is withheld from those who are excluded from the proto-moral universe of discourse. (Bergmann, 1998)

3.1. The “social” of “social robots” as an emergent property

3.1.1. Social facts as accomplishments

A long-standing ethnomethodological tradition has attended to the locally relevant features of a given setting as produced and maintained by co-present participants' mutual conducts. This approach investigates each property of a situation as emergent – or, at least, treats it analytically “*as if it were* something that emerged from the activities of parties to that situation and that has no “existence” independently of those activities” (Dennis, 2003). The previous stance, arguably the core “study policy” (Garfinkel, 1967) of ethnomethodology, was first introduced by Garfinkel (1967) in the following, now canonical, passage: “the objective reality of social facts as an ongoing accomplishment of the concerted activities of daily life, with the ordinary, artful ways of that accomplishment being by members known, used, and taken for granted, is, for members doing sociology, a fundamental phenomenon” (Garfinkel, 1967). Significantly, Garfinkel's (1967) formulation is both a definitional statement and an empirical program: from an ethnomethodological perspective, methods, activities, practices, etc. *are* what is encountered by members as taken-for-granted social facts (Psathas, 1980) and, consequently, they are what ethnomethodologists strive to report. In other words, from this analytic standpoint, “where others might see 'things,' 'givens,' or 'facts of life,' the ethnomethodologist sees (or attempts to see) process: the process through which the perceivedly stable features of socially organized environments are continually created and sustained” (Pollner, 1974). This stance is not intended as a philosophical assertion regarding the transcendental nature of reality but as an analytical perspective or a specific “level of description” (List, 2019) of social life: it is what Psathas (1980) names “the ethnomethodological attitude of reduction”, through which “the ethnomethodologist suspends belief in society as an objective reality, except as it appears and is 'accomplished' in and through the ordinary everyday activities of members themselves” (Psathas, 1989). To sum up, in what is still, to this day, a “classic” (Dennis, 2003) expansion on Garfinkel's (1967) original observation on the ongoing accomplishment of social facts, Zimmerman & Pollner (2013) remark that:

“From the analyst's point of view, the presented texture of the scene, including its appearance as an objective, recalcitrant order of affairs, is conceived as the

accomplishment of members' methods for displaying and detecting the setting's features. For the member the corpus of setting features presents itself as a product, as objective and independent scene features. For the analyst the corpus is the family of practices employed by members to assemble, recognise, and realise the corpus-as-a-product." (Zimmerman & Pollner, 2013)⁴⁵.

3.1.2. Identities and roles as accomplishments

A corollary of the accomplished nature of social facts is the potential variability of identities, roles or any socially ascribed property from one moment to the next. Because of the "ephemeral here-and-nowness of the situational practical accomplishment entails that participants have to deal with each social situation—because of their under-determinacy—for the first time anew" (Meyer, 2019), identities are not intrinsically stable over time. As Antaki et al. (1996) formulate it, rather than an identity "uniform throughout" the interaction, "the play of identities was more like the disposition of meat and fat in a salami sausage: differently patterned wherever you cut into it" (Antaki et al., 1996).

This view is directly opposed to "ex-ante models of human beings" (Meyer, 2019), i.e., to intrinsic definitions of members' identities, roles (vom Lehn, 2019), or qualities (Meyer, 2019) as identical with (or necessarily stemming from) specific material properties. From a strict ethnomethodological standpoint, a barking dog-shaped entity does not emerge as a "dog" in a social vacuum. In the case of a pet-owner relationship (e.g., Laurier et al., 2005; Roberts, 2004; Tannen, 2004; Tedeschi, 2016), the presence of a "dog" (rather than a "wild animal", a "carbon-based biological body", etc.) is embedded in a concrete activity in which the "dog" and the "owner" mutually configure the other as a participant with different qualities, rights, obligations, etc. In vom Lehn's (2019) phrasing, "it therefore is not the material and visual environment in which a person acts and the uniform he wears that make him a security guard but he becomes a security guard by virtue of his actions and by virtue of the ways in which others orient to his actions [...] role is not a property of an actor but a practical achievement" (vom Lehn, 2019). Among other matters, this ethnomethodological endeavor to uncover how properties of a social setting are locally achieved was notably applied to gender as an emergent feature by West & Fenstermaker (1993) and West & Zimmerman (1987). Attending to gender as a moment-by-moment accomplishment hence "transformed an ascribed status into an achieved status, moving masculinity and femininity from natural, essential properties of individuals to interactional, that is to say, *social properties of a system of relationships*" (West & Zimmerman, 2009, emphasis ours).

A different body of work relying on an "ethnomethodologically informed method" (Stokoe & Attenborough, 2014) has dealt in detail with the question of the local emergence of identity and with the situated production of social properties. This corpus of empirical research strongly connects but does not perfectly overlap with the previously mentioned studies. In this literature, usually grouped under the term "membership categorization analysis" (MCA), Maynard & Heritage (2023) identify two "strands". A first strand focuses on "the ways that

⁴⁵ Relevantly to our extrinsic and relational definition of the "sociality" of a robot (rather than an intrinsic and context-free property), Zimmerman & Pollner (2013) underscore "the occasioned character of the corpus in contrast to a corpus of members' knowledge, skill, and belief standing prior to and independent of any actual occasion in which such knowledge, skill, and belief are displayed or recognized" (Zimmerman & Pollner, 2013).

persons make who they are (or what they are doing) visible and recognizable to others ... as that kind of person” (Wieder & Pratt, 1990). The second strand builds on a similar inductive approach and on a corpus of empirical findings in common with the previous studies; yet, it focuses on the use of “descriptive categories” (Maynard & Heritage, 2023) by situated participants: it works to exhibit that “[t]he cultural meanings carried by categories are never innately given, but rather moulded and shaped around the social and cultural action(s) being attempted in that particular sequence of interaction” (Stokoe & Attenborough, 2014). This analytical focus is not strict, however, as Stokoe & Attenborough (2014) argue that membership categorization analysis aims to produce a “systematic ethnomethodological analysis of phenomena like 'culture' and 'identity' that is grounded in both categorial and sequential concerns” (Stokoe & Attenborough, 2014). That is, it attends to “members methodical practices in describing the world, and displaying their understanding of the world and of the commonsense routine workings of society” (Fitzgerald et al., 2009).

The endeavor of this “second strand”, although relevant, will not be at the center of our focus in this work: we will be interested in labelings or categorizations (Antaki et al., 1996) of the “robot” (as social, as a robot, as a friend, as a he or a she) as part of the overall orientation of participants towards this “robot” as, e.g., a participant with specific rights and obligations and accountability, more or less knowledgeable or more or less legitimate to express its knowledge on a specific topic, etc. Indeed, we argue, the treatment of the robot “as” something (e.g., a judge, a security guard; vom Lehn, 2019) can be produced in the total absence of description, categorization or labeling of this entity by participants. In this sense, we align with Schegloff's (1997) position (although rejected by Stokoe & Smithson, 2001) that “the explicit mention of a category term [...] is by no means necessary to establish the relevant orientation by the participants” (Schegloff, 1997).

3.1.3. Consequences for local settings which involve “robots”

The emergentist view described in the previous sections constitutes, to the best of our knowledge, both the conceptual ground and the empirical program for most EMCA-inspired works in HRI. Jones (2017), observes a “paradigm shift” in the field of social robotics in which “the robot becomes social by virtue of how people regard it” (Jones, 2017). More than a “social turn” (Sabanovic & Chang, 2016), i.e., more than a *shift in attention* towards the concrete ways in which robots are experienced and used in different social settings, it is a shift in the very definition of sociality mobilized by researchers as “an emergent property of the interaction itself” (Jones, 2017). As a matter of fact, in one of the first ethnomethodologically inspired research attending to robots, Pitsch & Koch (2010) attempt to describe how “a user's perception and categorization of a robot emerges step by step during the interaction with the system” (Pitsch & Koch, 2010).

Congruently, in another foundational analysis of agency from an interactional perspective, Alač (2016) strives to describe “how the non-creature-like, machinic, or thing-like features of a robot are not only evoked but – as part of semiotic moves and practical doings – contingently constituted in a multiparty encounter”. This approach “move[s] away from the idea that the robot's sociality has to be understood as an intrinsic and categorial property of the robot's *inside* [...], the robot is a technology that can be enacted, in one breath, as an agent and a thing” (Alač, 2016). For this reason, in their survey of the prevailing definitions of social agency in human-robot interaction research, Jackson & Williams (2021) categorize Alač's

(2016) approach as mobilizing the most “extrinsic” definition of social agency: i.e., as attending to the robot’s sociality “as enacted and emergent from how a robot is experienced and articulated in interactions” (Jackson & Williams, 2021).

Other pioneering investigations of human-robot interactions through an EMCA lens explicitly formulate an analytic definition of some properties of a robot as relational or emergentist. Pelikan et al. (2022) study human-robot settings from a dialogical perspective for which “agency never exists in a vacuum but always emerges in interaction with others” (Pelikan et al., 2022). Paraphrasing Heritage’s (2013) remark that “institutions are talked into being”, they observe that robotic agency is “embodied into being” (Pelikan et al., 2022). Similarly, Krummheuer (2015) explicitly studies technical agency as “a situated and emergent product of social practices” (Krummheuer, 2015a). Finally, Licoppe & Rollet (2020) provide an interactional and relational account of the emergence of the “sociality” for a robot: they define “social” not “as a set of isolatable features, but rather as a result” (Licoppe & Rollet, 2020; our translation). That is, they state, the “social character of an entity is not a property in itself, but the result of a process, of transactions and interpretative activities” (Licoppe & Rollet, 2020; our translation).

A notable implication of this analytic perspective in its most radical form is the following: there is no preexisting social robot that simply “awaits” to be treated as social by adequate users. From an ethnomethodological standpoint, it is meaningless to speak of an intrinsically social agent that was merely not accepted yet by the general public – e.g., because of resistance to technology as measured on a psychometric scale (Edison & Geissler, 2003; Heerink et al., 2010; Krägeloh et al., 2019). No robot, or feature of a situation, can be isolated and declared *ex-ante* as social “in potential” (e.g., because of its humanoid shape) while, in the meantime, other pieces of technology (e.g., a stapler) are defined as ontologically “non-social”. In other words, when a human refuses to activate a robot, an ethnomethodologically oriented researcher cannot speak of a social agent denied by a recalcitrant person. In this case, there is merely no “social agent” at all. This point, made by several of the works presented in this section, may distinguish interactionist or emergentist approaches from some underlying assumptions behind lay and commercial uses of notions like “technology acceptance” or “acceptability”.

3.1.4. “Human”-“Robot” “Interactions”. A toned-down approach to emergence

Our analytic approach follows the same endeavor and analytical perspective as the ethnomethodological and conversation analytic studies focused on human-human interactions mentioned earlier (sections 3.1.1 and 3.1.2). However, in light of this summary, it should be noted that we cannot pretend to the same degree of ethnomethodological “radicalness”. The following analyses will focus on the (non-)emergence of a social agent in situations that feature what is *a priori* intended to be a “social robot” by designers and engineers. That is, we aim to uncover if and how a preexisting assemblage of “metal, plastic, sensors, and other material” (H. Clark & Fischer, 2022) – discretized and categorized in advance by an engineer or by the analyst – starts to be treated by locally present members, even momentarily, as a full-fledged social agent. This approach is incongruent with a “pure” ethnomethodological standpoint, from which a property of a situation becomes a property only when it is treated as such by local members. There is a potentially infinite number of features of a setting that can be perceived and oriented to a relevant by local participants. Our “mixed” endeavor unavoidably imposes

on local situations the presence of a property (i.e., a robot) to, then, study how participants react to this predefined and discretized element of the situation. That is, although we do not define how the “robot” should be perceived, treated, described, or if it should even be seen by participants, we predefine this “robot” as a feature of the situation and focus our analysis on the manner in which participants respond (or not) to this pre-established property. In regard to typical fields of investigation for ethnomethodological studies (i.e., in the so-called “human-human” interactions), an equivalent approach would be to intentionally place a human preliminarily defined as possessing specific social attributes – e.g., a Japanese person (Nishizaka, 1995), a “recently-qualified-but-cynical-medical-student” (Antaki et al., 1996), someone pre-categorized as belonging to a specific gender or ethnic group – in a specific social setting (a museum, a classroom, etc.) and to check if and how this person is made “visible and recognizable to others ... as that kind of person” (Wieder & Pratt, 1990). Such a method does not perfectly overlap with a purely inductive endeavor where the situated relevance, in a particular setting, of a person (as the bearer of specific, locally accomplished, properties) is what the analysis reports – and not the original question around which the empirical setting is organized. In this sense, even human-robot studies that take part in a fully “natural” setting (i.e., outside a lab) are closer to a hypothesis-testing experiment than to a purely inductive study.

In sum, as Meyer (2019) notes, “the notion of ‘actor’ is ascribed by members and specifically created by social practices in social situations [...] The focus of ethnomethodology is by no means on actors and their subjective knowledge stocks or shared knowledge, but rather on the practices that bring them into being” (Meyer, 2019). Predefining the “actors” (be they robots or humans) of a setting is dissonant with an analytic definition of the relevant features of a situation as locally emergent. Yet, it is a necessary evil for most ethnomethodologically oriented studies which hope to shed some light on the local organization of situations involving both “humans” and (etically defined) “robots”.

3.2. The polysemy of “social agency”

3.2.1. Participant, member, conversational partner, social actor or social agent?

Mondémé (2016) notes that, when studying interactions between humans and other species, speaking of “participants” carries a less heavy theoretical load than referring to “members” (Mondémé, 2016b). Describing a robot’s status with the notion of “social agent” may impose an even more labyrinthine, heterogeneous, and sometimes tacit, theoretical background onto the situation. As Jackson & Williams (2021) demonstrate in their review of the use of this concept of “social agency” applied to human-robot interactions, there exists a wide spectrum of internalist, externalist, or mixed definitions of this emergent and/or intrinsic property. Depending on the scientific field, “what makes a robot social” (Henschel et al., 2021) can rely on radically different (and potentially incommensurable) theoretical foundations. Yet, among the eclectic vocabulary used to describe the interactional or ontological status of a robot (of which only a small part is listed in the title of this subsection), we believe that the concept of “social agent” remains, for better or worse, the most suitable. Despite interdisciplinary inconsistencies, “social agency” is, we argue, a satisfactory label, devoid of technical jargon,

for the coherent set of practices that EMCA analysts, in fact, study. That is, as an intradisciplinary term, it efficiently encompasses the processes, methods, practices, etc. on which EMCA researchers focus their attention when they scrutinize recordings of human-robot encounters – i.e., the phenomena that do matter to EMCA researchers, independently from how they are labeled.

Consequently, and given the existence of several recent side-by-side comparisons of what constitutes robots' "sociality" and/or "agency" across fields (Henschel et al., 2021; Jackson & Williams, 2021; Sarrica et al., 2019), it is critical for our study to ensure it aligns with what lies behind the concept of "social agency" (and, more generally, "social interactions", "social encounters", "social robots", etc.) mobilized in the EMCA literature and, if possible, to specify these terms even more. In short, we need to summarize what we are searching for in our data – and to do so without worsening the polysemy of "social agency". As we will demonstrate, our definition of social agency does not differ from most implicit or explicit definitions of "agency" or "sociality", heard in a very loose sense, which can be found in the EMCA-oriented literature – in particular, it appears entirely congruent with Pelikan et al.'s (2022) in-depth interactionist definition of social agency. This comes as no surprise; EMCA researchers are labeled as EMCA researchers precisely because of their interest in the same underlying phenomena. As we will attempt to show, the concrete interactional accomplishments glossed by the categories of "social agent", "social interaction", "conversational agent", etc. are, in EMCA-inspired works, overwhelmingly identical.

3.2.2. A minimal and empirically viable definition

Based on the previous analytic perspective (see sections 3.1.1 to 3.1.4 on local features of interaction as emergent), what the researcher etically defines as "the robot's sociality" emerges from local practices from co-present participants – or, rather, it *is* these practices. A "social robot" is a gloss for empirically observable conducts produced by members of a setting towards this entity: i.e., the specific activities, methods, or processes through which the robot is treated as accountable (Pelikan et al., 2022), as having interactional rights and obligations (Rollet et al., 2017), etc. Significantly, members may not label the result of their practices (those which are identical with a robot being "social" in the researcher's vocabulary) as a "social robot": this label is an etic, exogenous, label.

Still, a minimalist transsituational definition of "social agency" is required, at least as a guideline, for our study. For the sake of using categories that can be understood by the reader (rather than *sui generis* labels meaningful exclusively for co-present members of situated human-robot interactions), we need an analytic definition for what local processes the "sociality" and the "agency" of the robot corresponds to. This implies specifying at least some of the concrete situated phenomena that are glossed by these terms both in human-robot research and by lay participants *in situ*. Crucially, this definition has to be analytically viable rather than theoretically exhaustive: we need to formulate *empirical criteria* to define (some of) what Lindemann (2005) calls the "border of the social world"⁴⁶. That is, we need a definition that an analyst can mobilize to distinguish between "social" and "non-social" treatments of a

⁴⁶ This concept was notably applied to robotics by Muhle (2015).

robot, based on recorded, observational, data. Even if this entails leaving many observed cases uncategorized.

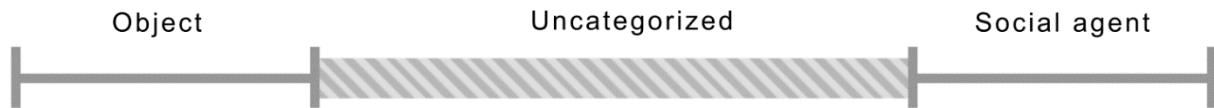


Figure 3.1. Range of applicability of an interactionist definition of social agency to observational data

As a matter of fact, a definition of “social agency” is unlikely to allow the analyst to, then, systematically classify each human-robot interaction recording or fragment. It is common for EMCA researchers to conclude that a fragment does not allow to demonstrate the presence nor the absence of a given phenomenon. The attempt to identify interactional phenomena from which an entity emerges as a “social agent” should be no different. It is possible to provide sound proofs that a robot is treated (non-)socially in canonical or marked cases in regard to predefined criteria. Yet, conversely, it is likely that a pool of cases will sit between both extremes of the spectrum – neither clear treatments of a robot as an object nor as an agent – and will be challenging or impossible to rigorously categorize (see Figure 3.1).

3.3. Social agency as accountability and contingency

To start with, ethnomethodological and conversation analytic uses of the term “social agent” predominantly point to a form of accountability. In this understanding, a “social agent” is an accountable participant in both senses of the term, i.e., in regard to “intelligibility” and “responsibility” (Robinson, 2016). That is, a “social agent” is, in part, an entity locally treated as accountable for *recognizing and producing recognizable* actions (the intelligibility dimension of accountability) as well as for producing *normatively relevant and sequentially implicative* (Schegloff & Sacks, 1973a) actions (the moral element of accountability – Robinson, 2016).

3.3.1. (A) Social agents as morally accountable (accountability as normativity)

3.3.1.1 Social agency as producing actions within a moral order

EMCA-oriented literature in HRI offers several overlapping variations on an understanding of “sociality” as being embedded in a moral order. Jarske et al. (2020) have most explicitly articulated this definition of robots as involved in a “social interaction” when their conducts are treated as normatively relevant actions. Relying on Garfinkel (1963)’s notion of “trust”, they indicate that “when persons trust, they produce actions that are understandable and discoverable as actions in-a-normative-order” (Jarske et al., 2020). For the robot to be

interacted with “socially”, some properties of its conduct must be treated by human co-participants as normative practices: e.g., the robot must be understood as “greeting someone with a wave” (thus making relevant a return greeting) rather than merely “raising an arm”. In other words, rather than merely involving entities “using speech” (Jarske et al., 2020), social interactions must involve participants who “talk”, i.e., situations “where a person produces actions that are discoverable within a normative order and assumes other participants to be able to perceive them as actions within that order. [...] On the other hand, 'using speech' refers to situations lacking this reciprocal attitude” (Jarske et al., 2020).⁴⁷ However, crucially, Jarske et al.'s (2020) definition does not completely abandon internalist preconditions for (genuine) social interactions, namely that robots should, too, “be able to interpret the situation here and now, understand the morality of interaction and the implications of the lack of responses” and not just be treated as *if they were* by humans. This definition of social interactions re-distributes the interpretative work on both humans and robots: for the “sociality” not to be a deceit, all participants must be (and not just assumed to be) able to grasp locally relevant interactional expectancies. In Jarske et al.'s (2020) perspective, because the robots we study in this work do not possess the aforementioned technical abilities, all encounters in which these robots will be involved can only display, at best, “the appearances of social interactions” (Jarske et al., 2020)⁴⁸.

Although they stay agnostic in regard to apparent versus genuine social interactions, (Rollet et al., 2017) provide a congruent perspective on moral order as key to the robot's emergent status. They do not speak of “social robots” but, rather, focus on describing how a robot emerges as a “participant in its own right” and provide criteria to specify this category empirically. From a detailed analysis of human-robot interactions, they remark that the robot may fall under an “analogous commitment as in human-human interactions” (Rollet et al., 2017). That is, the robot is a “participant in its own right” when it is treated as possessing *similar rights and obligations*, as part of a participation framework. In these situations, rather than neutral features of the setting, the robot's contributions are treated as embedded within the (constraining) “structures providing for the organization of the endogenous activity systems”. Note that – although Goffman's perspective is not systematically congruent with a conversation analytic framework (Schegloff, 1988) – human behaviors that display “face-work” (Goffman, 1967) towards the robot indirectly fit into this category: face-work documents the ascription of interactional rights and obligations to the robot, which make relevant specific conducts, e.g., closing an interaction with farewells (Goffman, 1967) rather than leaving without warning (Licoppe & Rollet, 2020)⁴⁹. In the same line of reasoning, because it attends to the robot as “ritually delicate object” (Rollet et al., 2017), a greeting “models the appropriate

⁴⁷Although, as Jarske et al. (2020) mention, robots cannot currently orient to rules (“they do not comprehend the world of normative practices” – Jarske et al., 2020), we will focus on the manner in which they are attended to as orienting to rules by co-present participants – without denying that the gap between what robots can “do” and what they are taken to be doing by human participants can be a form of “ethical 'crimes'” (Jarske et al., 2020).

⁴⁸ “We can interact with socially interactive robots in a similar appearance to how we interact with humans, but the interaction does not constitute what human interaction constitutes.” (Jarske et al., 2020)

⁴⁹ Licoppe & Rollet (2020) integrate facework, or at least visible forms of tact, as one of the local manifestations of the “sociality” of the robot. However, they are intentionally agnostic regarding the locally emergent (endogenous) or etically imposed (exogenous) nature of the category of “social robot”. Their endeavor is not to specify the methods employed by local participants (or by the researchers if it is exogenous) to produce this “social” category.

and expected way of acting and interacting that constitutes the addressee as a particular kind of entity” (Alač, 2016).

This outline of robots as participants or social agents partially relies on the idea that conversational sequence organization involves a “proto-morality” (Bergmann, 1998; Robles, 2015). Given this premise, the very involvement of the robot in a sequence organization supposes that it is attributed specific rights and responsibilities in a “micro-level moral order” (Stivers et al., 2011) or, rather, it is in part the demonstrable attributions of specific rights and responsibilities to the robot that the analyst indexes when describing the robot as involved in a sequence organization. Yet, the moral order of interaction is not limited to sequence organization. To the best of our understanding, the aforementioned studies provide examples of accountability where interlocutors (including the robot) “are morally responsible at all times (i.e., omnirelevantly) for recognizing, understanding, and adhering to relevance rules, and that breaching relevance rules is ‘accountable’” (Robinson, 2016). Hence, accountability is far from relating only to conditional relevance (i.e., a “crystallized type of sequential implicativeness” – Küttner, 2020)⁵⁰. It does not limit to, e.g., responding “hello” to a “hello”. Instead, accountability can be extended to any relevance rule (Robinson, 2016): e.g., not making a request without legitimate reasons to do so, not engaging in criticism when it is not “appropriate” (Robinson, 2016), telling good rather than bad news (Maynard, 2003).

In sum, since, “accountability status can be made relevant in interaction, such as through accounts, account solicitations, and other actions that index accountability (e.g., accusations)” (Robinson, 2016), we will focus on these observable practices as momentary treatments of the robot as accountable. Namely, if these practices of accountability are “windows through which relevance rules can be examined” (Robinson, 2016), they are, also, the very empirical phenomena that we point to when we speak about a robot being ascribed rights and obligations as part of a moral order. It is through the identification of these practices that the emergent and momentary interactional status of the robot can be demonstrated. They do not document an elusive “social agency” floating above the robot; they are what “social agency” glosses.

3.3.1.2 Social agency as rights and obligations for both human and non-human participants

In the previous definition, the robot is social in that it is treated as having interactional rights and obligations. However, because it has “rights” along with its “obligations”, the robot is not merely immersed in a normative order that binds its own conduct and only its own. This normative order supposes that some actions *also become normatively relevant for its human co-interactants*. In this line of thought, Pelikan et al. (2022) explicitly posit “that a robot

⁵⁰ “It has been proposed, however, that conditional relevance is but a special, crystallized type of sequential implicativeness [...]—a term coined by Schegloff and Sacks to refer to the sequentially organized implications of an utterance for subsequent conduct more broadly, simply meaning “that an utterance projects for the sequentially following turn(s) the relevance of a determinate range of occurrences (be they utterance types, activities, speaker selections, etc.)” [Schegloff & Sacks, 1973]. After a declining response to an invitation, for example, it is somewhat expectable that the inviting party will produce a next turn that deals with the declination in one way or another (e.g., accepting it, pursuing the invitation by proffering alternatives [...]). To be sure, such sequential relevancies and expectations are slightly more contingent and certainly not as strong (or as normatively binding) as those that hold between FPPs and SPPs in adjacency pairs (so departures from them may be slightly less accountable), but they clearly matter for the organization of sequences of actions, especially beyond the two parts of an adjacency pair [...]” (Küttner, 2020)

emerges as an 'autonomous' agent when it initiates and performs an action that is treated as relevant and accountable by human participants, as evidenced in their next actions". For example, when addressing a question to a human, a robot is treated as a social agent *in that it produces a normative pressure to produce an answer of a certain type and form* (Kendrick et al., 2020), demonstrable by its co-participants' answers, or their justifications for not answering, etc.⁵¹ After uttering a greeting token at the beginning of an encounter, the robot can be said to be treated socially when human interactants' conduct displays or observably indexes a "moral obligation to greet back" (Jarske et al., 2020). In this eventuality, the robot's conduct is retrospectively enacted "as a meaningful sequence initiation that requires a fitting next action, which [the co-participant] also produces" (Pelikan et al., 2020). Conversely, the robot is not treated as a social agent when "it is being poked in the eyes and talked about" (Pelikan et al., 2022) or when it is dealt with "as an object that can be scrutinized and explored rather than as a participant relevant to interact with" (Pelikan et al., 2022). Similarly, in this definition, positioning a robot as non-accountable for its own actions constitutes a non-social treatment of the robot: e.g., when side-sequences emerge between humans (Krummheuer, 2015b) as a resource to solve ongoing interactional trouble between the robot and a human interactant... yet from which the robot is excluded in spite of its physical co-presence (Krummheuer, 2015b).

3.3.2. (B) Social agents as competent participants (accountability as intelligibility)

Along the moral dimension of accountability, the "intelligibility side of accountability" (Stevanovic, 2023) is also central to how "sociality" or "social agents" are defined in EMCA-oriented works. This dimension of accountability as "a tool of sense-making" (Stevanovic, 2023) supposes that "accounts not only (and perhaps simultaneously) function to "save face" [...], but also to "save intelligibility" (Robinson, 2016). That is, treating the robot as a "social agent" requires to hold it accountable for producing "recognizable and understandable" (Garfinkel, 1967) actions for the activity at hand. This constituent of the robot's sociality largely overlaps with the notion of interactional competence which is, precisely, the "speakers' ability to make social actions recognizable to one another while taking into account individual identities and social role relationships" (Dai & Davey, 2023). In regard to this dimension of accountability, speaking of a robot as a "social agent" is a gloss for the different processes through which it is locally responded to as interactionally competent: the robot is a "social agent" if it is treated as a participant capable and bound to make its actions recognizable. Its actions are expected to make publicly intelligible "what is going on", and the robot can be explicitly required (as an account-able participant) to explain, describe, clarify, etc. what it is

⁵¹ Heritage & Clayman (2012) consider that adjacency pair organization "embodies a simple conversational norm that, upon the production of a first action (for example, a greeting, question, request, etc.), a recipient should respond with a corresponding second action" (Heritage & Clayman, 2012). As such, this norm "is unquestionably treated as an incorrigible feature of social life" (Heritage & Clayman, 2012).

“doing”. That is, a social agent's conduct is oriented to as displaying actions *designed*⁶² to be intelligible⁵³ (Stevanovic, 2023).

However, in practice, a strict line is hardly possible to draw between intelligibility and responsibility. To the best of our understanding, a synthetic definition of (robotic) “autonomous agency” as “the recognized ability to participate (i.e., contribute relevant actions) in an evolving sequence” (Pelikan et al., 2022) puts forward both these dimensions as criteria. Intelligibility and responsibility compose the interaction order as “a social and moral order, in that it forms the basis for mutual intelligibility and self-presentation” (J. Turowetz et al., 2016). In other words, there is a *normative order* weighing on the production of *intelligible conducts*. As Turowetz et al. (2016) observe, “members use of practices and methods have an inherently moral dimension in the sense that we take it for granted, and assume others take for granted, that practices and methods indeed will be conjoined with our talk to render what we say meaningful” (J. Turowetz et al., 2016).

3.3.3. (C) Social agents as producing contingent conducts

3.3.3.1 An edge case of non-social interactions

To make our empirical criteria as exhaustive as possible, let us look at some properties particular to an extreme case of treatment of a “robot” as an object, i.e., a case that would intuitively be categorized as a “non-social” interaction by most observers. Besides the enactment of the robot as non-accountable, we can attempt to specify if something else is characteristic of this “non-social” encounter: are there additional properties of the interaction – oriented to as relevant by the human participant – which are lacking or, at least, which are underspecified in the previous definition of “sociality”?

The following fragment is similar in many points to the one with which this work was introduced. It is an occurrence of the daily interactions that took place between an employee and a “receptionist” Pepper robot from 2020 to 2022 – during which time this employee was required to check in and check out of the building using this robot. Hence, rather than a first encounter, this constitutes an expert and ritualized interaction between a familiar human (or, in another vocabulary, a “power user”) and a robot that has systematically followed the same pre-scripted steps for years. In this peculiar occasion, however, to reveal the participant’s practical habits even more, the robot’s response latencies had been increased (i.e., it took more time to react), unbeknownst to him. Additionally, the robot’s processing time after the employee’s name had been pronounced was now conveyed gesturally, through a gaze aversion animation (L.5). This animation was intended to be a turn-holding cue which displayed that the robot was “thinking” (i.e., searching for the employee’s name in its database). Although this is far from a radical breaching experiment, these sudden changes in the robot’s script may allow us to better grasp the participant’s expectancies when confronted with a robot whose behavior is not “as usual”.

⁵² For example, connecting with Fischer's (2021) distinction between a proximal and a distal scene, a robot's conduct is treated as social if it is attended to as “ostensively attempt[ing] to reach for [an] object on [a] shelf” (i.e., a conduct designed to be publicly recognizable) rather than merely lifting its arm (K. Fischer, 2021).

⁵³ The recognizability of actions also appears to be a prerequisite for Tuncer et al.'s (2023) definition of interactional competence as “the capacity to produce a timely first or responsive action” (Tuncer et al., 2023).

3.3.3.2 Fragment 3.1: A familiar user checks-out on a receptionist robot

```

1.          % (2.0) @ (0.5) @ (1.4) *# (0.4) *%
rob        >>%-----display "accueil"-----%displays check-in screen-->
kar
img

```



Figure 3.2. Image 1 – KAR slaps ROB

```

2. ROB      you can show me the guest QR code or* give me your name
kar
           *seizes ROB's head-->
3.          (0.2) * (0.9)
kar        -->*tilts ROB's head to align with his gaze-->
4. KAR      Karl Smith
5.          (3.3) + (1.8) +
rob        +lowers its head and tilts it to the left+
6. KAR      Karl Smith
7.          (1.3) * (0.4) #
kar        -->*seizes ROB's head with both hands-->
img        #img.2

```

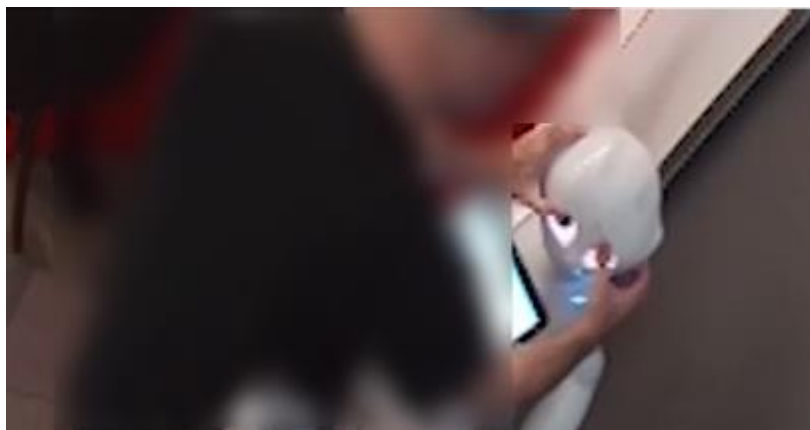


Figure 3.3. Image 2 – ROB gazes on the left, KAR seizes ROB's head with both hands

```

8.          *# (1.2) * (0.4)
kar        -->*tilts ROB's head to align with his gaze*
img        #img.3

```

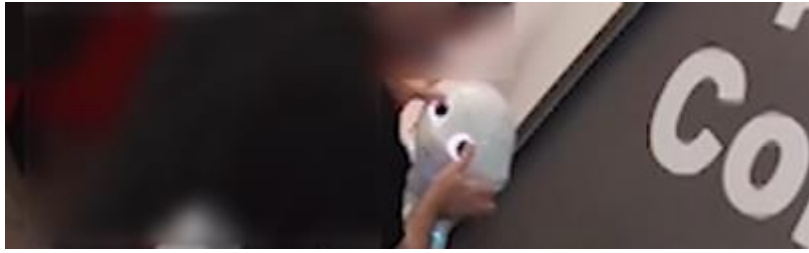


Figure 3.4. Image 3 – KAR turns ROB's head towards him

```

9. KAR      Karl Smith
10.         (1.8)
11.ROB     %is this you?
           rob      %displays KAR's name-->
12.KAR     =yes.
13.         (7.1)%(0.1)
           rob      %displays "do you wish to confirm your departure"
14.KAR     yes.
15.ROB     do you wish to confirm your [depa]rture? [ok.]
16.KAR                                     [yes.]      [yes.]
17.         (2.5)%(0.7)@(0.1)
           rob      %displays a green checkmark>>
           kar      -->@turns around>>
18.ROB     your check-out is registered

```

3.3.3.3 Description

After approaching the robot (not recorded), KAR (Karl) instantly kicks (L.1) the “bumper” part of ROB’s (the receptionist robot) base. Bumpers are plastic parts placed around ROB’s wheels that can detect impacts: when a bumper is kicked, ROB is usually programmed to search for a human (likely to be near ROB) by moving its head around, in a scripted animation. Forcing ROB to move its head would have, usually, allowed KARL to be instantly detected by ROB. This detection would have, in turn, instantly triggered ROB’s request to provide one’s name (L.2) and accelerated the check-out beyond its typical pace. Unbeknownst to KAR, this feature had been disabled on this particular robot. After waiting silently for 1.4 seconds, KAR slaps ROB on the right part of its face (L.1). Immediately after, ROB displays the check-out screen (L.1) and asks KAR to show a QR code or to provide his name (L.2). KAR seizes ROB’s head with his left hand, orients it even more towards his own face (L.3) and utters his name (L.4). From a technical point of view, seizing ROB’s head allows KARL to force ROB’s identification of a human standing in front of it, to ensure that ROB does not disengage from the interaction (because no human is detected anymore). After 3.3 seconds, ROB lowers its head and tilts it to the left (L.5) in an animation newly implemented as a “gaze aversion filler” – intended to indicate that the robot is internally searching for the human’s name in its database. This off-the-usual-script conduct is physically countered by KAR: while ROB’s head tilt is still ongoing, KAR repeats his name (L.6), waits 1.3 seconds (L.7), then seizes ROB’s head with both hands and forcefully turns it to face him (L.8). He then repeats his name (L.9) and stays silent while maintaining ROB’s head in place (L.10). After 1.8 seconds, ROB displays KAR’s name and asks him to confirm his identity (L.11), which he does (L.12). After a processing time of 7.1 seconds (L.13), ROB starts to ask KAR to confirm his departure (L.14). However, in overlap with the very first syllable of ROB’s question, and less than 0.1 seconds after ROB’s screen

displays “do you wish to confirm your departure” (L.13), KAR answers “yes” (L.14) and repeats it (L.16). Immediately after ROB displays a green checkmark (L.17), KAR turns around (L.17) and leaves.

3.3.3.4 Analysis: What does it mean to treat a conduct as non-contingent?

A striking feature of this encounter which has not been insisted upon in the previous definition is that the robot’s conduct is treated as non-contingent. That is, the human participant does not “act on the assumption that [the robot is] capable of choosing among alternative courses of action” (Bergmann, 1998). This orientation to a non-contingent conduct concerns both *the features of the interaction the robot will respond to* and *what the robot will respond*.

a) *What the robot will respond to is treated as predetermined: limited contingency*

Karl demonstrates a precise understanding of the extremely limited list of conducts on his side (namely, saying his name, saying “yes”, forcing the detection of its face, or touching the tablet) which can initiate a response from the robot. In other words, the robot’s reactions are not treated as a priori “contingent” to *anything Karl could do or say*, in the sense of contingency defined by Yamaoka & Hagita (2007)⁵⁴ as “a correspondence of one’s behaviour to another’s behaviour”. The robot is, instead, treated as able to respond to a *limited list of inputs* rather than to a *non-denumerable list of conducts*. Karl expertly knew which analogical “senses” of the robot to rely on (i.e., which sensor and programming rule to trigger) to obtain the expected result: hitting the robot’s “bumper” so that the robot moves its head around and detects him (L.1), physically orienting the head of the robot to force the identification of a human face (L.2 and L.7) so that the robot does not disengage from the interaction, saying “yes” as soon as the robot is able to register this input (L.14). Familiar users like Karl “packaged” their verbal and embodied turns only in regard to these sensors and programming rules: everything else in their verbal and gestural conduct (e.g., when discussing with someone at the same time) was not addressed, directly or indirectly, to the robot. These participants fully knew in advance, rather than *discovered*, what features of their conduct and of the setting the robot would treat as “relevant” (again, analogically) to the task at hand⁵⁵.

b) *What the robot will say or do is treated as predetermined: complete pre-scripting*

Another rather unusual conduct, by human-human standards, is displayed by Karl: he produces a turn which pre-emptively responds to the robot’s incoming question, before the robot even started uttering this question (L. 14)⁵⁶. Because this “answer” occurs less than 0.1 seconds after the robot displayed a prompt requesting to confirm the check-out (L.13), it is also highly unlikely that the participant responds to a renewed reading of this check-out screen. Rather, Karl’s response time indicates a mastery “by heart” of what the robot will say or display – and when – rather than the monitoring of possibly unexpected behaviors. If the

⁵⁴ This definition was incorporated by Pitsch et al. (2009) into an EMCA approach.

⁵⁵ This situation may not perfectly overlap with institutional talk (Drew & Heritage, 1992), or any kind of standardized talk (e.g., scripted telephone survey interviews – Houtkoop-Steenstra, 1995) where one would have to stay “on script”. In institutional talk, one can always go beyond the script and expect its interlocutors to respond to this out-of-script conduct. In the case of the receptionist robot, familiar participants knew in advance, and displayed knowing, everything that the robot could possibly respond to.

⁵⁶ This practice of responding to the robot before it spoke and before it displayed anything on its tablet was common among familiar users of this receptionist robot.

robot had attempted to ask Karl any other question at this step of the check-out, Karl would have responded “yes” before the question could be heard or read.

3.3.3.5 Contingency as a condition for accountability

Undeniably, Karl’s orientation towards the robot can also be analyzed through the lenses of accountability. For example, the robot is not made accountable for the out-of-the-ordinary conduct that constitutes its “gaze aversion filler” head tilt (L.5). Although this embodied behavior is not responded to as a relevant contribution, the robot is not publicly criticized, nor pressured to explain or justify why it acted in such a way, nor is it involved in a repair sequence. This seemingly off-script conduct from the robot is physically interrupted as a form of technical issue rather than treated as a recognizable or accountable action. Yet, what we want to emphasize is not the lack of accountability of the robot but, rather, Karl’s orientation to the complete absence of contingency of his interlocutor’s conduct. At no point does Karl treat ROB’s head tilt animation as potentially responsive to the local situation (relevantly or not), as indexing any property of the setting or of his own conduct. We suggest that, in this particular episode, the lack of accountability of the robot mechanically results from this treatment of the robot’s behavior as fully determined: the absence of “actions that index accountability” (Stevanovic, 2023) merely documents the more fundamental orientation displayed by Karl here. Contingency is, after all, a condition for accountability and, more generally, for any form of moral responsibility. As Bergmann (1998) notes, “ascribing to others the capability of choosing among different courses of action means that we take them as fellow human beings who could have acted differently. This possibility of choice is the presupposition for the attribution of responsibility. Once the others are equipped with this capability of choice, they may be held responsible for their doings” (Bergmann, 1998).

Conversely, and beyond this particular example, the packaging of some turns has been shown to observably index the contingent nature of an interlocutor’s potential response. For example, some “low entitlement demands” (Antaki & Kent, 2012) (e.g., “I’m wonderin’ if a doctor could call and see [name] please” – Antaki & Kent, 2012) have been argued to display acknowledgment of doctors’ contingencies and their impact on the granting of a request. Such analyses are useful resources in our endeavor to produce an empirically viable definition of social agency. Along with the example of Karl, they confirm that at least some practices can constitute argumentatively sound proof that the robot’s conduct is treated as contingent or non-contingent (to local or non-local parameters).

3.3.4. The last piece of our definition: contingency as a third criterion to define a “social agent”

3.3.4.1 Embeddedness in the here-and-now

This quick analysis provides us with the last piece of our interactionist definition of a “social agent”. The robot’s conduct must be treated as responsive to the local situation (at least potentially) rather than pre-scripted. It cannot be attended to as the mere actualization of the ex-ante expectations of its designers and programmers. That is, for the robot to be an accountable social agent, *what the robot will respond to* (1) and *what the robot will respond* (2) must not be treated as fully determined. Indubitably, these criteria are intricate to

demonstrate based on observable data – besides edge cases of long-term use of robots as in Karl's case – and might only rarely be practically usable as part of our delimitation of social interactions between robots and humans. Still, *when it is demonstrable that a human treats a robot's responses as entirely pre-determined (i.e., as if the robot cannot be impacted by local contingencies and is not recognizably embedded within the here-and-now of the interaction), then it is not interacted with as a social agent.*

In sum, to be accomplished as an agent, a robot must be treated as “capable of choosing among alternative courses of action” (Bergmann, 1998). Instead, in the previous fragment, Karl does not treat the robot as a social agent in that he exclusively orients to the “predefinite rules” (Porcheron et al., 2018) that the robot follows – in the programming sense of “following”. This is unlike human-human interactions where “such ‘rules’ are established as achievements in and through interaction by conversationalists, situationally and moment-by-moment” (Porcheron et al., 2018). This view of the behavior of a social agent as not identical with a set of rules is congruent with the view of “rules as resources” in ethnomethodology rather than as “abstract algorithms that predetermine behavior” (J. Turowetz et al., 2016). By contrast, when the robot is treated as strictly adhering to rules (the equivalent of rules as “interior mental machinery” (Button, 1990) rather than orienting to them (Button, 1990; Garfinkel, 1967)⁵⁷, it is not attended to like a social agent.

3.3.4.2 Contingency and social agents in human-robot interaction research

These observations on contingency are congruent with the definition of an agent as able to “autonomously choose what to do and then carry it out by themselves” which Pelikan et al. (2022) identified as a Western ideal implicit in the way we attend to human co-participants. In short, to be accountable, the robot must be “autonomous” (Pelikan et al., 2022) in the sense of able to choose between different conducts – i.e., “in control” (Pelikan et al., 2022) to some degree of what it does. An alternative yet complementary idea postulates that in order *for the human participants' actions themselves to be social*, the robot must be treated as autonomous: Muhle (2015), building on Luhmann's (2013) theory of double contingency, states that a condition for an action to be “social” is that “the behaviour of the other person cannot be expected to be a determinable fact” (Luhmann, 2013). In many ways, our approach is an interactionist attempt to specify what it is to attend to the robot's conduct as something else than “a determinable fact” (Luhmann, 2013) but, rather, as a “choice between different possibilities” (Luhmann, 2013).

More fundamentally, the robot does not emerge as a local actor when it is responded to by co-present participants as the mere executant (or, in Goffman's participation framework, as an extreme form of “animator” – Goffman, 1979) of a content programmed ex-ante by developers removed (temporarily and physically) from the interaction. In Suchman's (1987)'s terminology, it is, then, encountered as a pure “plan” incapable of situated action. As “pre-strategy professionals” (Hopper, 2005), computer programmers are not local actors⁵⁸: at best – when discrepancies between plans and situated action do not lead to a complete breakdown

⁵⁷ “There is a distinction then between rules that people can be shown to orient to, and rules that are said to be an interior mental machinery. On the latter understanding, rules stand behind action, on the former, rules are embedded within the action.” (Button, 1990)

⁵⁸ See section 2.2 for a development on the extent to which a rule-based agent can be said to be a “local” actor, embedded in the here-and-now of the interaction.

in the interaction –, a pre-scripted robot is an efficient reflection of programmers’ a priori anticipation (Rollet & Clavel, 2020) of what would occur during an interaction with this robot in a predefined context (an office entrance, a shopping mall, etc.)⁵⁹. Hence, in the definition we have built so far, social agents must emerge, at least in part, as “emergent-when” actors (Hopper, 2005): their conduct must be perceived as recognizably responsive to the contingencies of the ongoing interaction⁶⁰.

3.4. Result: A summarized definition of “social agent”

3.4.1. An interactionist formulation

In short, in the following work, a “social agent” will be identical with *an accountable entity* (Pelikan et al., 2022) *with specific interactional rights and obligations, whose conduct is attended to as contingent on the ongoing situation*. In their review of existing definitions of social agency, Jackson & Williams (2021) distinguish “what social agency is” from “human behaviors that indicate ascription thereof” (Jackson & Williams, 2021). We argue that, from an emergentist perspective such as the one depicted here, the observable embeddedness of an entity in a social and moral order (J. Turowetz et al., 2016) “is” social agency. These criteria notably concur with Pelikan et al.’s (2022) detailed analysis of the emergence of social agency in human-robot interaction, which draws on Enfield & Sidnell’s (2021) definition of agency. For the latter, “agency is defined as the combination of an actor’s flexibility and accountability” (Enfield & Sidnell, 2021)⁶¹. These dimensions are similarly required for a robot (as locally emergent properties, accomplished during the interaction, and not as intrinsic abilities or potentialities) if it is to be treated as a social agent. Hence, based on the previous components A), B) and C) of our definition of sociality, we argue that, for any fragment of “human-robot interaction” under study, the analyst interested in social agency should ask: Is the robot’s

⁵⁹ “[...] the design of robot behaviours is based on the designer’s ordinary interactional knowledge: he/she assumes that these hypothetic interactions should start with a welcome phase, and that saying hello will trigger a hello.” (Rollet & Clavel, 2020)

⁶⁰ Note that we do not refer here to the robot’s technical ability to produce “contingent” actions, i.e., to “deal with the dynamic and flexible nature of human interaction” (Pitsch et al., 2009). That is, we do not speak about the intrinsic “potential” of a robot (evaluated by another group, e.g., the robot’s designers, based on their own specific methods) to handle the now canonical issue of plans and situated actions (Suchman, 1987), or to adequately represent the here-and-now of a situation. In this section, we regularly resorted to cumbersome sentence structures such as “the robot’s conduct was oriented to, attended to, treated, etc. as responsive to the contingencies of the ongoing interaction”. This phrasing aims to depict the robot’s non-determinacy as a local accomplishment. No matter if the robot is, in a technical sense, fully scripted – what matters for its emergence as a social agent is that it responded to as in control of its actions (Pelikan et al., 2022) to some degree.

⁶¹ In this definition, agency supposes control, composition, and subprehesion:

“1. An agent controls a perceptible event to the extent that they can determine the timing and placement of the execution of a piece of behavior; and as a corollary, that they can attend to the timing and placement of another’s behavior.

2. An agent composes a meaningful action to the extent that they select and execute a specific behavior that should stand for something; and as a corollary, that they can recognize others’ actions as standing for things.

3. An agent subprehends to the extent that they are not surprised by what happens next as defined by the conventions of the activity, as calibrated to the conventionalized rights and duties of participants.” (Enfield & Sidnell, 2021)

conduct treated by co-present members as forming recognizable actions, responsive to the local situation, that “project (empirically) and require (normatively) that some 'next action' [...] should be done by a subsequent participant” (Heritage, 1998)?

3.4.2. The analytical status of the category “social agents”

What exactly, then, is this category of “social agent” as it is typically used in the EMCA-oriented literature? Is it an “emergent category” (Mondada, 2002, 2017; Pitsch & Koch, 2010) which properly encapsulates participants' own practical orientation towards the local situation, or is it a technical and transsituational nomenclature? Are we engaging in “constructionist” work (Button & Sharrock, 2016) by imposing identities on participants (here, the robot) or, even more, by pre-assigning the status of “participant” to the robot as a preliminary analytic category (Mondémé, 2016b)? In other words, is there something akin to a status of “social agent” (even labeled otherwise or, more likely, not labeled at all) that humans practically orient to during their interactions with a robot? Or is “social agent” merely a convenient term for the analysts to group together different conducts they deem relevant? Can it be that the category of “social agent” refers to interactional phenomena, activities, relationships, etc., meaningful for both the researcher studying recordings of interactions with a robot and, at the same time, for the humans on these recordings while they were immersed in the urgency of their ongoing interactions?

This issue disappears when we summarize what our definition of “social agent” glosses. When trying to determine if the robot emerges as a “social agent”, what we do study is whether the robot's conducts are woven into an interaction order (understood as “a complex web of standards, expectations, rules, and proscriptions” – Sidnell, 2010) in which they are recognizable and accountable (Rollet & Clavel, 2020). In sum, we are not interested in what the robot “does” or “is” (beyond the exogenous concern of the analyst for settings that involve what is a preliminary labeled as a robot) but in what local organization it is embedded. To the extent where “there is an organisation to the action and interaction that stands apart from any particular actor” (Button & Sharrock, 2016), we are not interested in the “doer” but in the “doing” (Button & Sharrock, 2016). In the definition presented above, social agency *is* the social and normative organization in which different members (etically labeled as “humans” or “robots”) are embedded: it *is* the publicly observable and describable practices, methods, processes, etc. accomplished and reflexively indexed by these members' conducts. In sum, dwelling on the distinction drawn by West & Zimmerman (1987) between “sex categorizations” and “the accomplishment of gender”, our main endeavor in this work is not to establish the processes through which a robot ends up being merely *labeled* as “social” or “humanlike” by human participants but, rather, how it *emerges* as a participant with specific rights and obligations whose conduct is accountable and contingent. That is, we are interested in the (potential) *embeddedness* of the robot in the local organization of the interaction produced and maintained by all co-present members.

4. AN EMCA APPROACH IN A TECHNOLOGY COMPANY: INEFFABILITY AND FRICTIONS

You are not obliged to assault people with discourses that are out of their road, when you see that their received notions must prevent your making an impression upon them: you ought rather to cast about and to manage things with all the dexterity in your power, so that, if you are not able to make them go well, they may be as little ill as possible; for, except all men were good, everything cannot be right, and that is a blessing that I do not at present hope to see.

—Thomas More, *Utopia*, 1516

4.1. Ethnomethodological conversation analysis and the corporate world

4.1.1. Ineffability and ethnomethodology: bug or feature?

Pollner (2012) suggests that Garfinkel “may be something of a Wittgensteinian Lion (Wittgenstein, 1953) whose form of life is so different –being connected with the form of life– that when he speaks we cannot understand him”. He goes on to state that “translation to understandable terms is to subvert Garfinkel; it is to translate him back to the very form of life –including its mode of representation (i.e., formal analysis)– which he attempts to avoid” (Pollner, 2012). Given such a long-lasting and intentional resistance to translation, the ineffability of EMCA’s findings or practices to non-practitioners does not come as a surprise. It is not a bug, but a feature of ethnomethodological activity. Whereas other professional disciplines studying social life (social psychology, behavioral economics, etc.) have been struggling for a long time with the issue of explaining their findings to unfamiliar audiences and have developed countless narrative strategies to overcome this gap, ethnomethodology has been, at best, moving at a much slower pace. In fact, the arduousness (or impossibility) of translating even the most toned-down and non-radical ethnomethodological analysis is sometimes considered encouraging proof that the analyst is not “losing the phenomenon”.

Beyond professionals of “formal analysis”, a relatively sparse body of literature addresses the regularly observed reluctance or lack of interest of lay members of society to thematize their own taken-for-granted methods. Going from *methods as a resource* to *methods as a topic* is not an ordinary activity. As Psathas (1980) remarks, “[t]o direct one’s attention to the resources being used in accomplishing an activity is to change one’s activity to that of inquiring into the methods used rather than using the methods. The fact remains that members, when engaged in the production or accomplishment of an activity in the world of everyday life, are interested in that production and not in understanding or explicating how they are doing it”. Going from “using the methods” to “inquiring about the methods” (Psathas,

1980) is not, either, an effortless activity: it requires going from a practical mastery of common sense procedures (Have, 2005) as “seen but unnoticed” (Garfinkel, 1964) to an a posteriori verbal description of this “know-how”. Indeed, “[t]he effective test of ‘knowing how’ for members in the world of everyday life is doing whatever it is, rather than being able to explicate how they do it” (Psathas, 1980). The difficulties of this task are likely to be found in any endeavor to transition from a practical, possibly non-representational, “skillful coping” with the world (Coeckelbergh, 2019; H. L. Dreyfus, 2001), to a representational, rule-based and theoretical understanding of a labeled activity (constituted as an “object” processed by a “subject” – Dreyfus & Dreyfus, 2005). In the end, no matter the cause, members are generally “uninterested” (Psathas, 1980) in thematizing their ordinary methods: as mentioned by Psathas (1980), a strong proof of this persistent lack of interest is that “ethnomethodological studies are relatively recent in the history of the social sciences”. That is, “if members had been interested in such analyses, they would have undertaken them long before this time” (Psathas, 1980).

Behind these considerations on the relative ineffability and taken-for-granted nature of ordinary methods might lurk an explanation, or an etiology, for the lack of mainstream interest in conversation analysis and ethnomethodology as disciplines – compared to the typical examples of popular research fields that are social psychology and behavioral economics. For better or worse, the concepts and vocabulary of these two fields – nudging (Thaler & Sunstein, 2009), etc. – have arguably become folk knowledge, or at least common concepts in the corporate world. By contrast, despite a substantial community of practitioners, distributed across various academic fields (sociology, linguistics, psychology), and despite a certain degree of “normalization” of its “radical” roots (Lynch, 2016), EMCA’s practical adoption outside of academia is rather limited – including, as will be our focus, in technology companies. In lay terms, one could argue that EMCA is not “where the money is”. Unlike many subfields of experimental psychology, marketing science, management science, design science, data science, etc., the affinities between the private sector and EMCA are rather thin – both in terms of concrete exchanges and in terms of shared concepts or theoretical backgrounds. Rather, precisely because EMCA first developed in opposition to formal analysis (Garfinkel, 1996; Pollner, 2012), it naturally positions itself at odds within a culture of engineers or designers, immersed daily in a positivist conception of science, in which quantitative and/or experimental research is the only rigorous method for producing knowledge about the world. As I will argue, the idea that there might be “order at all points” (Sacks, 1984b), even in a single fragment of talk, is especially difficult to enforce in this context. The aforementioned incommensurability (Reeves, 2022) of mental horizons or of “forms of life” may reach its apex when the audience is composed of experts in formal analysis: engineers, data scientists, user experience researchers with a background in quantitative psychology, etc.

The following subsections detail concrete disjunctions or frictions that occurred between the pre-existing methods, theoretical backgrounds, design or programming tools, etc. to which Aldebaran’s engineers or designers were accustomed and the practices or concepts (sequentiality, locality, etc.) connected to an EMCA approach. That is, I want to describe the main sticking points which I encountered while conducting an EMCA oriented research in a tech company – without dwelling on whether these difficulties are inescapable consequences of EMCA’s incommensurability with alternative “conceptualisations of sociality and social organisation” (Reeves, 2022), or merely contingent on the local specificities of the professional context in which I evolved. Still, I hope that the subsequent points of friction, between an EMCA approach and the ordinary activities of a private robotics company, can shed light on

the different representations or practical habits from which these frictions emerge.

4.1.2. Disjunctions with the practical constraints and theoretical backgrounds of roboticists

To begin with, the next few pages outline reservations that I encountered regarding the adoption of an EMCA perspective during my thesis. Some of these were produced by engineers, others by ergonomists or designers (“UX researchers” or “UX designers”), and some by the middle or upper management. The intent of this list is not to refute these arguments but to summarize the points of friction that they reveal. In fact, given the professional constraints of each of these actors (time constraints, accountability to hierarchical superiors, readability of the reports for non-acquainted readers), I consider several of these points to constitute valid criticism of my personal attempt to “carry out” an EMCA approach in the context of a private technology company. Specifically, I share some of the reservations, detailed below, on the systematic relevance of EMCA findings to inform the interactional design of rule-based robots. These arguments aggregate scattered remarks made by many different colleagues throughout my years at Aldebaran – they were rarely formulated in one go by a single individual. To ensure they are presented in their strongest form (or, so to say, “steelmaned”), these points are rephrased (rather than transcribed) and replaced among the existing scientific literature. Additionally, for the sake of confidentiality, the specific status of their author(s) (as employees of Aldebaran) is not disclosed, as well as the circumstances of their enunciation.

4.2. Do we need EMCA if we have common-sense? Practical mastery *versus* analytical understanding of conversation

The following objection could be roughly summarized as “do we need EMCA if we have common-sense?”. That is, *in the concrete situations where human-robot interactions are designed (when designers are at their desk pondering different designs in front of an interaction flow, when they are “brainstorming” during meetings, etc.) and for all practical purposes, do designers need to construct “conversation” as a distant, theoretical object, rather than keep a tacit, intuitive, relationship to it?* As Pelikan (2020) remarks, “when evaluating and designing for interaction with robots, we draw on our tacit knowledge about such interaction patterns. For instance, in our everyday lives we constantly face interactional challenges like how to open a conversation with a stranger, how to instruct someone in using the dishwasher, and how to behave in a traffic situation that is not covered by the rules” (Pelikan, 2020). Similarly, Rollet & Clavel (2020) note that “the design of robot behaviours is based on the designer’s ordinary interactional knowledge: he/she assumes that these hypothetical interactions should start with a welcome phase, and that saying hello will trigger a hello”. These observations constitute a solid basis to justify the use of an EMCA perspective to inform design practices and to support (or, at least, formulate) the designer’s tacit ordinary knowledge rather than “blindly accept our gut feelings as scientific concepts” (Pelikan, 2020). This is the position generally held by EMCA-inspired researchers or designers and effectively synthesized by Moore & Arar (2018):

“They must be able to articulate the mechanics of human conversation so they can design it, instead of simply knowing it tacitly like everyone does. For example, a conversation expert may describe the function of the word “oh” to mark speakers’ realizations (Heritage, 1985) or how the phrase, “to the what?,” in response to “I’m going to the workshop,” elegantly elicits a repeat of a single word “workshop” (Schegloff et al., 1977). Conversational UX designers use such observable patterns of the machinery of human conversation in building conversational machines.” (Moore & Arar, 2018)

Yet, the very same observations (about the inevitable reliance of interaction designers on their ordinary knowledge as members of society) can be used to argue that a tacit, practical, understanding of human-human conversation is “good enough” for designers – as everyday conversationalists – to design human-machine interactions. Much like the jazz pianist whose advice to become a proficient jazz improviser was merely to “sing while you’re playing” (Sudnow, 1993) rather than a granular description of his moment-to-moment practical decision-making, the conversational designer can be argued not to require precise theoretical principles to identify “what a conversational agent should do” at a given moment⁶². In a phenomenological sense, this latter argument questions the need to purposely turn everyday routines into distant objects of study, to, then, extract a set of granular guidelines or rules from this detached perspective. As expert members of the social world, we practically master taken-for-granted methods for organizing our social activities. We “cope” (Coeckelbergh, 2019; H. L. Dreyfus, 2014) with conversations daily without requiring a precise formulation of our interlocutors’ conduct – or, most of the time, without the need to discretize and “bin” these conduct in a “pre-existing inventory of discrete actions” (Enfield & Sidnell, 2017) to appropriately respond to the situation. This argument is far from limited to the skilled design of conversational agents and has been discussed – with widely different theoretical foundations – in regard to many activities: sports (Cappuccio, 2023), music education (Deslyper, 2013) or expert musicians (Hoffding, 2014), clinical problem-solving skills (Peña, 2010), and, more generally, any discipline in which a “deliberative rationality” (H. L. Dreyfus, 1987; H. L. Dreyfus & Dreyfus, 2005) tends to compete with a “calculative rationality”⁶³ (H. L. Dreyfus, 1987). As Fjelland (2020) notes:

“The bicycle rider keeps his balance by turning the handlebar of the bicycle. To avoid falling to the left, he moves the handlebar to the left, and to avoid falling to the right he turns the handlebar to the right. Thus he keeps his balance by moving

⁶² In many ways, this argument from Aldebaran’s employees reformulates and supports Psathas’ (1980) remarks about members’ limited incentives to “know” how they accomplish everyday actions:

“Members ‘know how’ to produce an event or social situation through their actions, but they do not ‘know’ how they do it; similarly, they ‘know how’ to recognize a social situation and identify it, but they do not ‘know’ how they do ‘recognizing.’ Their efforts to explain ‘how’ they do it are simply inadequate to the task. There is, for them, no practical reason for addressing such a question in the first place since if one ‘knows how’ to do something there is no need to be able to say what the ‘knowing’ consists of. *The effective test of ‘knowing how’ for members in the world of everyday life is doing whatever it is, rather than being able to explicate how they do it.* The fully socialized member of society accomplishes everyday actions in a manner sufficient for all practical purposes. The manner in which he accomplishes such tasks remain for him an already ‘understood’ taken-for-granted ‘fact’ despite its actual mystery.” (Psathas, 1980; emphasis mine).

⁶³ According to H. Dreyfus & Dreyfus, (1986), “calculative rationality” relies on the attempt to give as little room as possible to embodied know-how and to regulate practice by a system of rigid rules, while “deliberative rationality [...] does not seek to analyze the situation into context-free elements but seeks to test and improve whole intuitions”.

along a series of small curvatures. According to Polanyi a simple analysis shows that for a given angle of unbalance, the curvature of each winding is inversely proportional to the square of the speed of the bicycle. But *the bicycle rider does not know this, and it would not help him become a better bicycle rider* (Fjelland, 2020; emphasis mine).

Hence, the two stances presented at the beginning of this subsection offer alternative answers to the same underlying question: does a theoretical understanding of practical conversational processes genuinely help to produce better designs – rather than hindering the designer’s spontaneous ability to grasp what should be done in a given situation as an expert member of the social world? This question is not merely one of time constraints or cost-benefits. If EMCA unavoidably reconstructs (K.-M. Kim, 1999) a situation experienced *on a pre-reflexive level* by conversationalists, then what degree of granularity in the formulation of conversational practices (i.e., in “putting things into words”; Hirschauer, 2007) allows for efficient guidelines, and what degree paralyzes the design work? In Schegloff’s terms, in the absence of a “demonstrable analysis”, “we are left with ‘a sense of how the world works,’ but without its detailed explication” (Schegloff, 1991). The concern put forth by Aldebaran’s engineers and designers can thus be reformulated as such: isn’t this “intuition” or “sense of how the world works” (Schegloff, 1991) enough when it comes to design rule-based behaviors for a robot? From their perspective, the granularity of EMCA – as well as its endeavor to turn participants’ methods into the topic of the analysis – becomes a hindrance: by trying not to “lose the phenomenon” (Eisenmann & Lynch, 2021), it leads the researcher to miss quick and easy common-sensical improvements to the robot’s design. As Relieu et al. (2004) argued 20 years ago, “the goal of ethnomethodology is not so much to design situations as to contribute to the construction of an empirical body of knowledge on specific theoretical objects”⁶⁴ (my translation). In this understanding, heuristics or guidelines stemming from less intricate approaches – which leave more room for the designer’s own “common sense” or “intuition” – may constitute better compromises for short-term design problems.

4.3. Why use precision tools for the analysis if we must use a sledgehammer for the design?

Another regular concern at Aldebaran pertained to the tension between the granularity of the analysis allowed by an EMCA approach and what could be realistically designed and implemented into the robot. There was an uncanny gap between, on the one hand, the analysis of thoroughly transcribed interactions detailing the fine-tuned mechanisms through which humans coordinate each other, and, on the other hand, the aridity of what the robot could process from the external world. For example, provided that the robot’s cameras, algorithms, and processing power did not allow it to detect most gestures, was it relevant to spend hours studying a subtle embodied practice displayed by human participants (e.g., a practice produced through a head tilt)? What was the “added value” of granularly describing the conduct of a waiter speaking to a customer while serving a dish, to, then, design a serving

⁶⁴ Relieu et al. (2004) add that, among other factors, this “may explain the unsystematized nature of the relationship between the detailed analytical practices of situated action and the recommendations/prescriptions for design work” (my translation). Indeed, they remark that “in practice, the tangible contribution of these studies to design is often limited” (Relieu et al., 2004; my translation).

robot whose ability to hear voices was, at the time, non-existent? In sum, why use a microscope to untangle the moment-to-moment organization of naturally occurring interactions if, afterward, we must design the robot's behavior with a hammer?

A quick description of the perceptual abilities of the Pepper robot should clarify where the previous concerns were coming from. The following two main constraints weighed on engineers or interaction designers:

A. The robot's perceptual abilities were limited

The perceptual abilities of the robot were significantly restricted compared to the able-bodied humans on which our data were (initially) collected: in another vocabulary, the robot's "umwelt" was widely different from most human beings. Developers could configure Pepper through a framework named QiSDK (see section 4.6.1 for a description of the QiSDK): this software development kit did provide a wide range of information⁶⁵ (e.g., whether the robot identified a human shape at close range, whether a sensor had been touched, etc.) to anyone attempting to script specific conducts for Pepper. Yet, to prevent a severe increase in its response latencies (given the robot's hardware and processing power), a Pepper robot using an unmodified version of its operating system (NAOqi) and of the QiSDK did not run algorithms that identified human gestures from the information provided by its cameras.

As a consequence of the limited information that Pepper could gather and use from its environment, engineers and designers faced many constraints when attempting to design rule-based criteria to determine, for example, if humans were projecting to disengage from an interaction – or, conversely, to identify if passersby' conduct displayed that they intended to engage in a focused interaction with the robot. During my time at Aldebaran, this was a long-standing problem: many engineers and designers battled against the robot's technical constraints to establish consistent yet efficient engagement or disengagement criteria⁶⁶. After many debates, and given the processing power of the Pepper robot, it was deemed preferable not to go for a more efficient gesture recognition algorithm, as this would result in an even higher response time for the robot. In other words, more recent hardware (better cameras and computing power) was considered a prerequisite to significantly improve the gesture recognition abilities of the robot without impacting parallel processes (e.g., without worsening the robot's lag before responding to an utterance).

For a time, a relatively refined set of rules was tested internally on Pepper to detect "uninterested humans", "interested humans" and "humans seeking engagement"; yet, it could not rely on subtle embodied cues (e.g., someone leaning towards the robot). This decision tree is represented in Figure 4.1 below. For example, it defined a "human seeking engagement" as a static or approaching human that was standing in the "interaction zone" (i.e., at less than 1.5 meters from the robot) and whose face was still visible less than 2 seconds ago (i.e., if this human's face could be seen, they were not turned in the opposite direction). Although the "human engagement" system which relied on this set of rules was never released on a public version of the QiSDK, it should be noted that – in my experience

⁶⁵ https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/perception/tuto/people_characteristics_tutorial.html

⁶⁶ Recognizing engagement or disengagement is also a long-standing issue for academic researchers. See for reference Rollet & Clavel (2020).

and during several technical tests I had the opportunity to produce before my thesis officially started – it worked as expected. Its simplicity took advantage of the robot’s capabilities to detect humans’ faces, trajectories, and positions relative to the robot: although it was slow (compared to human-human standards) and did not facilitate smooth opening sequences, this system produced a surprisingly low amount of false positive or false negatives (i.e., engaging uninterested humans or refusing to engage interested humans).

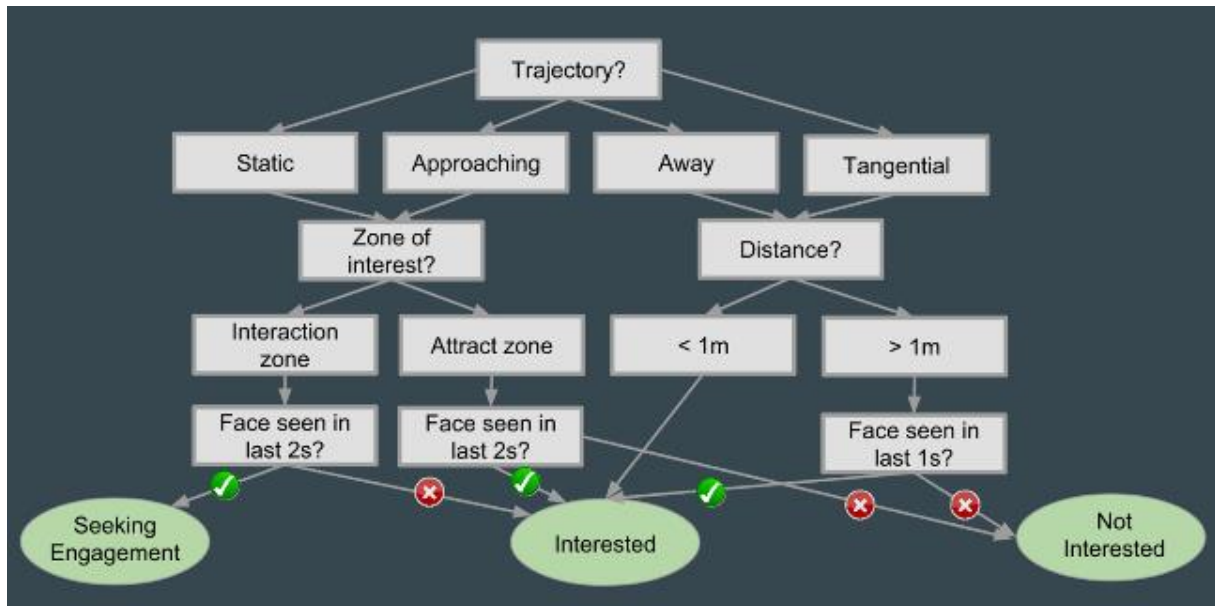


Figure 4.1. “Engagement Intention Flow Diagram”. This diagram summarizes the criteria for the identification of humans “Seeking Engagement”, “Interested” and “Not Interested” (Image courtesy of Miriam Bilac, used with permission).

Yet, with such perceptual limits, any decision tree to engage an interaction or to disengage from an interlocutor had to be based on very scarce criteria – far removed from the subtle patterns relevant to humans. For example, Mortensen & Hazel (2014) describe several embodied practices produced and oriented to by participants to establish whether a “greeting will effectively lead to a sustained interaction” (Mortensen & Hazel, 2014) or, on the contrary, to establish if this greeting does *not* project an imminent focused interaction. In the vocabulary of the engagement flow presented above (see Figure 4.1), co-participants rely on subtle cues to display or progressively negotiate if one is “seeking engagement” or “uninterested”: “participants rely on several bodily visual resources such as smiling, the ‘determined’ gait of the student’s walk, and sustained mutual gaze” (Mortensen & Hazel, 2014). All these cues (and many others) were not usable by a designer on Pepper as they were absent from the robot’s perceptual environment – including the sequential context of these cues. In other words, relying on S. E. Dreyfus & Dreyfus’ (1980) definition of expertise, the robot was analogically confined to a beginner’s status (see section 2.3 for a discussion of this analogy). First, inevitably, because a rule-based system was used (H. L. Dreyfus & Dreyfus, 1992, 2005); then, because, within this system, the robot followed a very limited number of rules.

B. The robot’s latencies were comparatively high for human-human interaction standards.

Compared to an average of a few hundred milliseconds between human conversationalists

(Kendrick, 2015; Skantze, 2021), Pepper required a significant delay before responding to a co-participant's speaking turn. Similarly, its reaction time to human presence (e.g., the time Pepper needed to greet a group of humans after being approached by them) sometimes exceeded a few seconds. In other words, the robot could not imitate the usual pace of able-bodied human-human interactions. As observed by Pitsch (2016), “perceptual delays” can be heavily consequential on an interaction. It also imposed severe limits on what the robot could be designed to respond to – and how. To take an example among many: even imagining that the robot was able to finely perceive humans’ hands, it would have been impossible to design a human-like responsive handshake feature. The mutual monitoring taking place during a handshake between two humans gives rise to moment-to-moment adjustments (Tuncer et al., 2023) in the milliseconds range, much faster than the robot's average reaction time.

I believe that the technical limits presented above were those that contrasted the most with the finely tuned human practices that I routinely attempted to specify. These technical constraints, well-known to engineers and designers who faced them daily, significantly impacted the granularity of the designs (EMCA-informed or not) that could be programmed.

4.4. “Losing the phenomenon”: Micro-analytic approaches and synthetic reports

4.4.1. EMCA and the pace of a technology company

Evidently, EMCA’s conceptual background, transcription conventions, and themes of interest (Relieu et al., 2004) are not optimally adapted to the formats of hard-hitting, easy-to-read, and synthetic reports. As Tavory (2022) phrases it, “ethnomethodology deliberately loses the battle for parsimony in its insistence on the detailed production of orderliness”. In Aldebaran, as is generally the case, field observations or user tests were summarized in user experience (UX) reports through graphs, bullet points, verbatim, and ethnographical examples. The research methods that led to these “user experience” findings could be roughly pictured by anyone with a limited background (even informal) in experimental psychology, ergonomics, or ethnography. In contrast, much more preliminary explanations were required as soon as was introduced a statement resulting from the EMCA-oriented analysis of a corpus of interactions. Indeed, such statements could not rely on a preexisting familiarity of the audience with the methods and “study policy” (Garfinkel, 1967) of ethnomethodologically inspired approaches (see chapter 2). Introductory explanations were mandatory to justify the very idea of looking at collections of single cases and the inductive nature of this endeavor: comparatively more time was required to clarify “how the analysis was conducted”, leaving less space for presenting the findings themselves⁶⁷.

There is nothing ground-breaking about these brief observations. EMCA’s commitment to avoid aggregating different interactional phenomena under common labels (Schegloff,

⁶⁷ This issue persisted when EMCA-based reports had to be relayed up the hierarchical chain. Stakeholders who intended to pass on information about these reports to their hierarchy were in the delicate position of having to repeat the previously mentioned phase of preliminary clarifications, yet without being familiar with EMCA themselves.

1993) renders it dissonant, by design, with the fast-paced design processes of some technology companies; that is, when research is not conducted within a laboratory that operates on its own timeline. Although my thesis work was insulated from this rapid rhythm, some of the smaller-scale studies I participated in were specifically designed for an audience operating under strict deadlines and unfamiliar with an EMCA approach. I would like to document how these constraints impacted, concretely, my attempts to produce reports relying on an EMCA stance: that is, how my original endeavors were adjusted (perhaps more than needed) to remain relevant and audible to my interlocutors – even at the risk of “losing the phenomenon”. Unless, unbeknownst to me, Aldebaran’s situation was atypical, or that my skills at conveying the relevance of an EMCA approach were far below average, the following descriptions should hopefully exemplify difficulties to be expected for an EMCA-oriented researcher working in a technology company.

4.4.2. Examples from the construction of a report

I will attempt to flesh out the previous statements by detailing how a UX report was constructed based on an internal study. The production of this report provides examples of minute decisions, made on my end, which aimed at abridging the original analysis into a rapidly comprehensible document that could be defended during a presentation.

The study covered by this report was co-organized with two other employees: the lead UX designer and Aldebaran’s main animator and illustrator – who was, at the time, working on the expressions to be displayed on the new Plato robot’s face (see Figure 4.2). This study was based on the following scenario: in a recreated restaurant setup, participants who played a “waiter” were invited to guide participants who played a “customer” to a table, take their orders, and serve them food⁶⁸. Each waiter played their role for around 20 minutes and faced different demands from the customer. The entire session was filmed. The objective of this exercise was to identify how humans (former professional waiters or first-timers) approached a table, served food to the customers, and moved away from the table. For example, we hoped to better understand how gazes and gestures were coordinated when a customer requested “more bread” from the waiter, or how the waiter’s “departure” was negotiated after it had served the customer. The relevant features of these finely tuned processes (those which participants oriented to) were then transcribed and analyzed. In particular, the aim was to produce a (non-exhaustive) list of opening and closing sequence organizations that the robot was most likely to encounter in the context of a restaurant and which, simultaneously, normatively guide customers’ behavior at these moments⁶⁹.

As a second step, the end goal was to transpose at least some basic features of the waiters’ typical conducts – and what these conducts oriented to – into a “serving robot”. This, of course, implied going from locally relevant practices (for the human participants) to rules defined ex-ante (for the robot). For example, whether they were technically implementable at the time of the study or not, some of the practices being considered were:

- The robot should gaze in the direction of the table it will serve only 2 meters before it

⁶⁸ Although some participants were former professional waiters and the setup was a realistic restaurant environment the “role-playing” component of this experiment falls under the criticism of Stokoe (2013).

⁶⁹ In another vocabulary, the goal was to identify the most frequently observed “scripts” or “standard schemas” (Conein, 1988; my translation) when waiters approached or left a table for the first time.

arrives.

- The robot should announce the plates (by uttering, e.g., “there you go”) when it is decelerating during the last centimeters which separate it from the customers' table; but it should not have entirely reached the table when it starts to announce the plates.
- After serving, the robot should already be in the process of rotating towards the direction in which it is going to depart when it utters “enjoy your meal” to customers.

Transposing human-human conducts to human-robot situations was understood as an initial draft before the robot itself could be confronted to humans – at which point the designer would be able to filter out “what worked” and what should not be imported from human-human interactions.

This report was paradigmatic of my attempts to introduce an EMCA approach for two reasons:

1. *It aimed to inspire designs for a robot with very basic perceptual and gestural abilities.* The robot around which this study was organized was the serving robot Plato⁷⁰ manufactured by Aldebaran. In its “serving” form, this 111 centimeters-tall robot is cylindrical in shape and has three trays on which dishes can be placed (see Figure 4.2). It features a tablet that displays either a face or the robot’s user interface. At the time of this study, in 2021, this robot stood out mainly for its navigation skills. It could “map” the spatial layout of restaurants and navigate, e.g., from the kitchen to a customer’s table, while avoiding obstacles in cluttered spaces. Yet, conversely, the robot could not detect humans through its cameras, nor hear what they said: at that time, the only way customers or waiters interacted with the robot was through a tablet. Whether it could or should detect customers one day, as well as respond to their verbal requests, was still unsettled. In other words, the resources with which a designer could play to craft interactions with customers were limited.

⁷⁰ Plato’s specifications can be found at <https://urg.directus.app/assets/297400c8-bbba-43a5-ba88-de5ba0811af2> . Accessed January 29, 2024.



Figure 4.2. The serving robot Plato of Aldebaran in its current shape (2023)⁷¹

2. *The data ended up being genuinely useful – but not the in-depth EMCA-oriented analysis.* The raw video recordings and some illustrated slides from the report became a long-standing source of inspiration for both the interaction designer and the artist working on the robot's facial expressions. Yet these professionals did not aim to construct “an empirical body of knowledge on specific theoretical objects” (Relieu et al., 2004): their approach did not need to rely on the formulation of the sequential contexts in which “waiters” and “customers” coordinated their actions, nor on a detailed description of these participants' embodied and gestural conduct. Instead, *they could watch these recordings or pictures again and again to grasp some expressions, gestures or movements that humans exhibited, without attempting to precisely formulate the local parameters to which these conducts were responding.*

I have reproduced below excerpts from my PhD journal that document the compromises I made while writing this PowerPoint report. The core of the report consisted of a list of several typical behaviors exhibited by the waiter and the customer. This list was produced beforehand, during two datasessions of one hour with a team of 5 or 6 employees from Aldebaran. These datasessions attempted to specify recurring practices displayed by participants, whether or not they were technically transposable on the robot at this point. While producing the final report, I attempted to document *each concession I made during the datasessions and, subsequently, during the report's composition, to adapt their format to my audience and to the time constraints.*

⁷¹ Image retrieved from <https://www.aldebaran.com/>. Accessed January 29, 2024.

4.4.3. The “path of least resistance”: discarding the mutually responsive character of participants’ conduct

Journal excerpt:

To present my observations to the audience within a few minutes, common multimodal transcription conventions – in this case, Mondada's (2016) conventions – will not fit the time constraints. Even more, in my experience so far, these conventions are usually not read by non-CA practitioners. I could attempt a “comic strip” format; yet this will still require spending a lot of time of the presentation on each fragment to convey the precise manner in which waiters and customers coordinated themselves. The core issue is that there is limited time to spend on each case, and that the described phenomena must be easily summarizable in ulterior meetings.

As a result, on the final PowerPoint slides, I settled for a simplified transcription format where some multimodal and temporal features were not annotated:

En attente de réponse du serveur (comportement actif)

- | | | |
|---|---|-------------------------|
| 1. Orienté vers le serveur
(2 secondes s'écoulent) | → | 1. S'adresse au serveur |
| 2. Se penche vers le serveur
(1 seconde s'écoule) | | |
| 3. Produit un rappel ("Yasmine ?") | | |



Figure 4.3. Transcription of a customer's conduct when the waiter was not responding to her requests (showcased during a PowerPoint presentation).

Crucially, the “lay” transcript above includes only one actor per PowerPoint slide: either the customer or the waiter. Indeed, to fit each fragment in one slide in a comprehensible manner, and for my audience not to “switch off”, I renounced to visually displaying the waiter’s conduct alongside (or after) those of the customer. During the previous datasession, my attempts to draw attention to the mutually responsive nature of the customer and waiter’s conduct did not generate interest, and we moved on to the next phenomenon (although my skills to bring these practices to light are also to blame). In particular, during the aforementioned datasession, my tentative description of the granular ways in which the customer’s conduct was responsive to the waiter’s behavior sparked no hope to inspire a design one day. I also suspect that it momentarily suggested a severe disconnection between my EMCA approach and the company’s immediate concerns: to have a robot that works. Perhaps mistakenly, I am generally keen not to provide the impression, often latent in the case

of PhD students (and, I believe, not entirely unfounded), of carrying out purely prospective research while being entirely unaware of its immediate inapplicability. Given the complete inability of the robot to detect humans at the time, every viable interaction design for the not-so-near future was doomed to be non-responsive to customers' gestures or voice.

From the perspective of a “pure” EMCA endeavor to seize participants’ mutual orientations to an ongoing interaction (Due & Licoppe, 2021), my PowerPoint presentation’s “dilution” of this EMCA approach (see Figure 4.3) heavily obscured the contingent and responsive character of participants’ conducts. That is, it limited the visibility of designers and engineers about:

1. *The co-constructed character of the interaction:* rather than a script unfolding invariably in the same way at specific moments of the interaction (i.e., when “guiding a customer to a table” or when “attempting to talk to a waiter”), each participant responded to the other participant’s continuously evolving behavior.
2. *The temporality of actions:* were they simultaneous, immediately adjacent, separated by a long silence?

The feedback on the PowerPoint slides from this report was positive: they were deemed relevant and usable information about typical conducts exhibited by waiters or customers. In this sense, the “simplified” transcription format presented above (see Figure 4.3) matched, to some degree at least, the constraints of the designers: the robot could not, in any case, monitor what the human was doing, nor react on a moment-to-moment basis to this human’s conduct. At this stage, it did not possess sensors and algorithms able to detect human movements. Looking back, if this presentation format ended up being the “path of least resistance” for me (perhaps mistakenly), it is mostly because this format seemed more aligned with the practical possibilities and limitations faced by designers and engineers.

In the end, the previous journal excerpt documents a clash between different endeavors: an attempt at understanding the emic orientation of participants *versus* an initiative to draft typical observable conducts for waiters or customers that could be transposed on a robot with limited abilities. As Reeves, (2022) notes about the relationship between design and ethnomethodology in human-computer interaction, “much conflict is driven simply by different purposes: design-orientation is about finding solutions to design problems or creating new design possibilities. It is a ‘leap’ to go from studies to design because the studies pursue their own particular logics and needs, configured by the necessities of disciplinary relations, interests of the EMCA researchers in question, etc. Whether such studies ‘map’ clearly to design outcomes (e.g., particular solutions, new possibilities, etc.) is not guaranteed” (Reeves, 2022).

4.4.4. Losing the phenomenon

The following journal abstract provides another example of a compromise that resulted in the aggregation of different (emically defined) situations into one. It focuses on the same report as before, derived from the datasessions mentioned above. One of the sub-goals of these datasession was to clarify how human waiters “departed” from a table (i.e., how they produced their exit and moved away) after serving the customer, to, then, provide ideas about possible designs of “social departures” for the waiter robot. A typical “long departure” case emerged as part of this datasession and started to be described. Yet, time was running out in the meeting

and, for the sake of brevity, *this typical case ended up knowingly aggregating two different phenomena into one. They were deemed “close enough”*. I argue that this shed light on what constituted the adequate degree of granularity for an analysis in this context, for all practical purposes.

Journal excerpt:

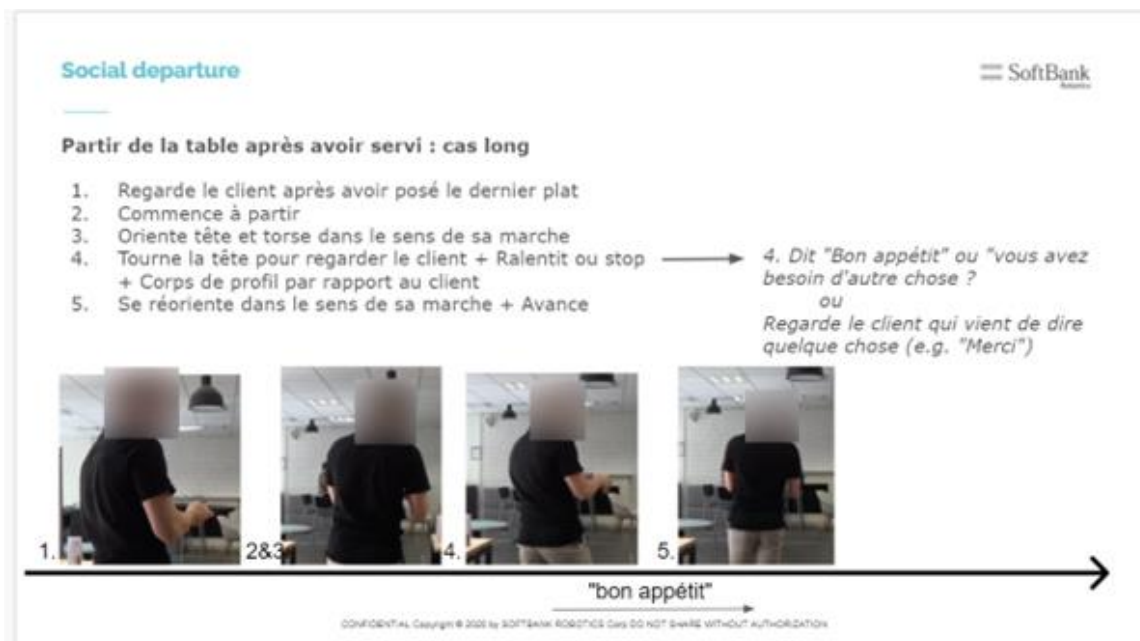


Figure 4.4. Step-by-step representation of typical ways of “departing from a table” for waiters.

Here is a screenshot from a PowerPoint slide that synthesizes a finding deemed relevant to inspire later designs. It describes how (1) after serving all customers at a table, (2) the waiter typically started to leave, (3) then gazed back at the customer (4) while possibly uttering something, (5) and finally reoriented his head in the direction he was walking in. Step (3) to (4) clearly aggregate several observably different situations. Even in a very preliminary manner, we can start to distinguish them in two broad categories.

1) Cases where the waiter responds to an immediately prior turn produced by the customer: i.e., situations where the waiter slows down and gazes back at the customer as a response to a “thank you” or a request initiated by this customer.

2) Cases where the waiter themselves initiates a request or a question while gazing back (e.g., “do you need something else”): this request or question does not observably react to a customer’s verbal or embodied behavior.

The two loosely formulated categories above can be connected to observable differences in the degree to which the waiter turned back towards the customer: doing so, their embodied behavior displayed various degrees of availability and projected (or not) a next-turn (Mortensen & Hazel, 2014). That is, these subtle differences in the waiter’s orientation – indexed to different sequential contexts – accomplished entirely different actions (“acknowledging a thank you” versus “inquiring if everything is ok”). A hypothetical robot on which this “social departure” script [presented in Figure 4.4]

would be rigidly implemented would systematically apply this exact behavior when leaving a table, whether or not it fits the local sequential context. Such a robot would be doomed to be a “beginner” waiter in the sense of S. E. Dreyfus (2004): it would only follow very generic trans-situational rules. It would not be a “local” actor.

Yet, in the real world, with real design constraints, the different phenomena aggregated behind this typical “social departure script” were not analytically separated, and did not need to be, for the designers’ immediate practical purposes. No matter their situated relevance for the human waiter, the subtle differences in the waiters’ conduct were too fine-tuned to be reproduced on the robot, and the customers’ conducts to which the waiter responded were too tenuous to be perceived and interpreted by the robot. Last but not least, an in-depth distinction between these different conducts would have meant dedicating the rest of the datasession to it and sacrificing other potential, immediately applicable, findings.

4.4.5. Slides, reports, and practical knowledge

Through the previous examples, I attempted to illustrate that the micro-sociological findings produced and reported in a fast-paced technology company do not convey, *by design*, the multimodal and temporal nature of the recorded interactions. Such reports – produced within severe time constraints, strict design endeavors, and relying on “technologies of the intellect” preadapted to these endeavors (see section 4.5) – engender “a formalised discourse that bears little relation to, and obeys different logics than, the mode of practical knowledge that informs the activity the analyst wishes to know about” (Hughson & Englis, 2002). That is, in response to multiple structural constraints, they end up providing “an account of practice out of which *all the elements of practical knowledge have leaked*” (Hughson & Englis, 2002; emphasis mine).

4.5. “Technologies of the intellect”: Frictions with usual modes of representation of human-robot interactions

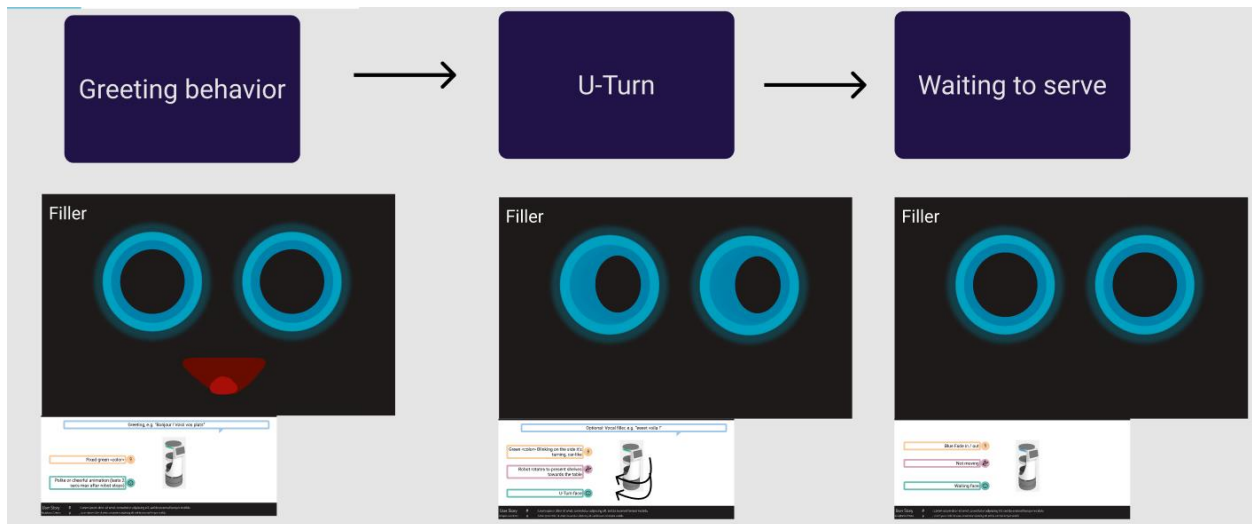
The previous paragraphs describe how EMCA-inspired results were pared down, by my decision, resulting in the loss of some relevant features in order to be communicated. Yet, manners of representing human-robot interactions already existed at Aldebaran. I would like to illustrate how these modes of representation inevitably forged a path of least resistance towards the design of specific behaviors for the robot⁷². To state the obvious, the existing

⁷² Following Psathas & Anderson (1990), any system of representation of “speech and/or action” implies a selection and a discretization of some features of “the actual, embodied and situated original spoken production”: “Thus it is impossible to present a discussion of the practices entailed in producing a transcript in accord with this transcription system without at the same time discussing the analytic concerns which generated and sustain it. A first observation is that there is not, and cannot be, a ‘neutral’ transcription system. The presumably ‘neutral’ presentation of the details of produced speech/action would be the actual, embodied and situated original spoken production.” (Psathas & Anderson, 1990; emphasis mine). I argue that this observation extends to interaction flows representing expected (rather than observed) human-robot interactions – whether these modes of representation are produced for analytical or design concerns, and whether the selection of features they operate is deliberate or not.

modes of representation of human-robot interactions (comic strips, storyboards, interaction flows, etc.) at Aldebaran were pre-adapted to representing interactions as they had been designed so far. These “technologies of the intellect” (Goody, 1975, 1977)⁷³ had been polished over time by the designers and engineers, based on what was deemed desirable and technically possible to do in human-robot interactions in specific settings and/or for specific commercial partners. The design tools we used to represent interactions (before programming them) were not built to highlight fined-tuned and overlapping mutual adjustments.

As a consequence, attempting to transpose micro-analytic findings into an interaction flow caused friction with the existing mode of representation for human-robot interaction that we used so far on the design software Figma⁷⁴. A sample of this Figma framework is displayed below (Figure 4.5). It presents the expected (observable and hearable) behaviors for a serving robot during three “steps” of its service: when it greets humans while approaching a table, turns to present the trails on which are placed the dishes, and waits for the dishes to be served. Figure 4.6 zooms in on the description of the expected behavior from the robot during each of these steps. All those steps appear in an overarching “interaction flow” (not reproduced here) which represents in a synoptic view all the possible steps that the robot can go through during its service, and what (a human input, a timer, etc.) triggers a change from one step to another.

For confidentiality reasons, the following screenshots (in Figure 4.5 and Figure 4.6) do not faithfully reproduce a currently “live” design: they intentionally correspond to an outdated design suggestion. Nevertheless, they feature some of the steps (i.e., discretization of the robot’s “conduct”) which were predefined at the time for the robot (greeting, departing, etc.) and illustrate how the robot’s behavior was represented on Figma.



⁷³ This concept, coined by Goody (1975, 1977), refers to tools that not only augment but modify human cognitive abilities: the invention of writing, mathematics, measurement systems (Levinson, 2020a), etc., up to language itself (Levinson, 2020a). That is, these tools do not only facilitate human thinking through the “re-internalization of an external aid” (Levinson, 2020a), they result in a form of “re-wiring of the mind” (Levinson, 2020a). The use of specific symbols and conventions to represent human-robot interactions (i.e., a system of representation) appears to match this definition.

⁷⁴ www.figma.com. Accessed March 10, 2024.

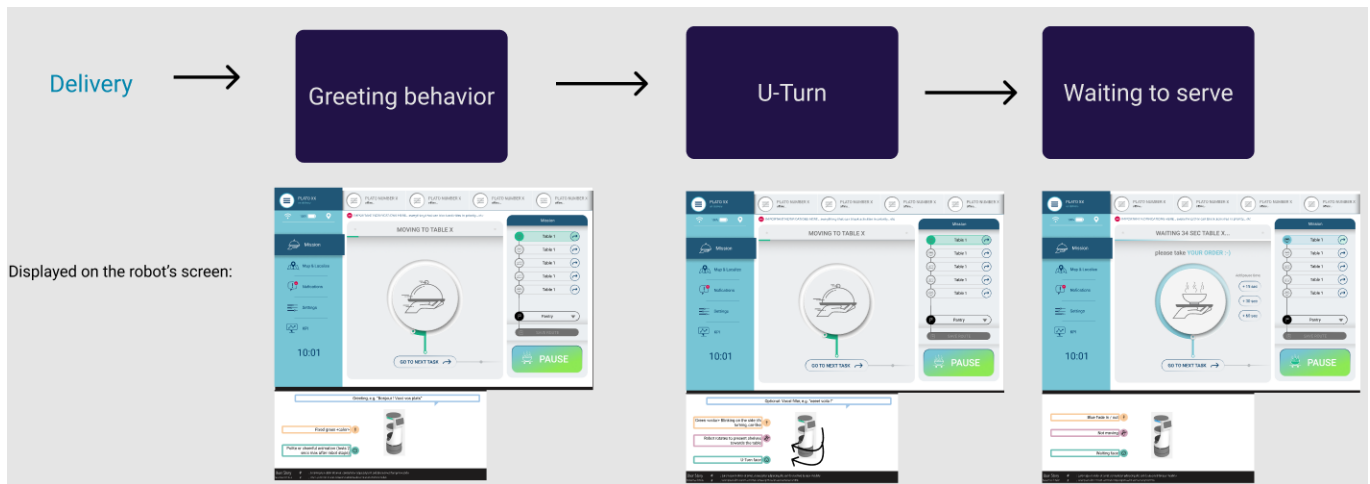
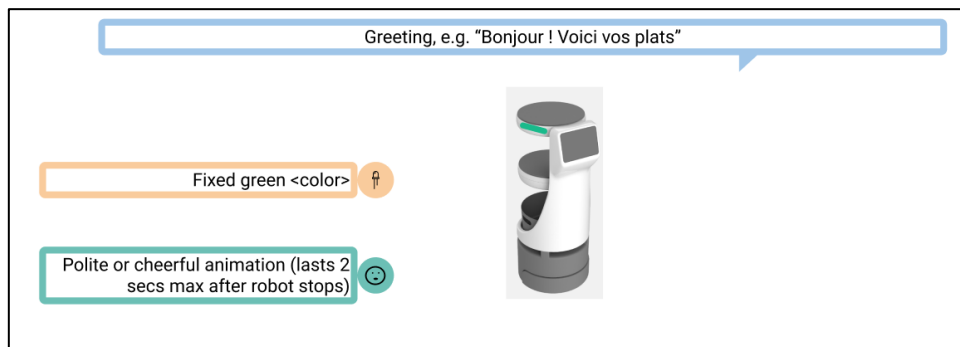
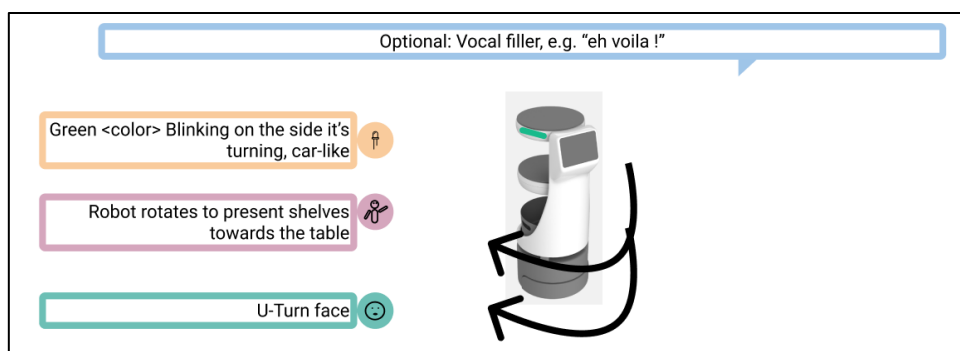


Figure 4.5. Examples of the mode of representation used on Figma for designing the robot's behavior in human-robot interactions. Used with permission.

Greeting behavior step



U-Turn step



Waiting to serve step

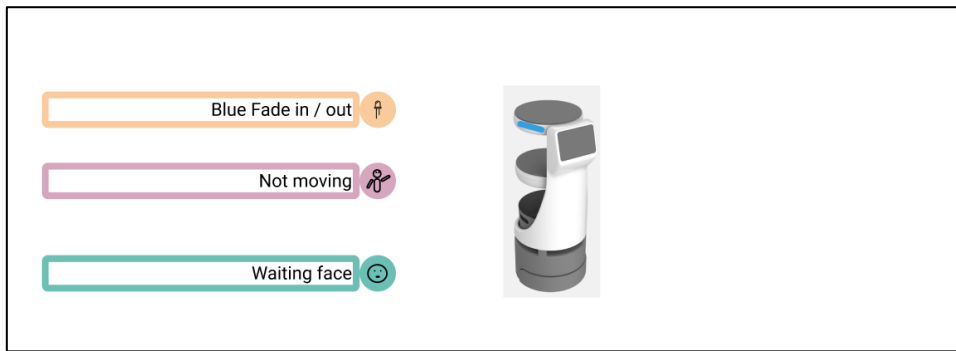


Figure 4.6. Zoom in on the expected behaviors from the robot in each “step”

This framework had been built on Figma over time (through the efforts of several UX designers) for specific practical purposes – far removed from EMCA’s endeavor to adequately represent participants’ local orientations. Among other goals, it aimed at:

- Representing the overlap between the robot’s movements or utterances and what was displayed on its tablet, for engineers to implement this synchronized tablet-robot behavior.
- Discretizing the robot’s conduct in different steps, and specifying which other steps were possible to trigger from there (or which steps the user could go back to, e.g., by pressing the “home” button on the tablet). Then again, this list of connections between steps (although established from the user perspective) was an additional resource for the engineers to program the robot.

Conversely, this step-by-step framework did not allow for the temporal representation of overlapping or finely tuned moment-to-moment adjustments from humans. It was neither its practical use nor its intended use. For example, using this framework, mutual adjustments between a waiter (or a serving robot) and a customer were not possible to represent. Of course, the conduct from a customer could be added as an additional “step” to a Figma flow like the one presented above (Figure 4.5): the customer would be pictured producing a bundle of inputs, to which, then, the waiter or serving robot would respond through a bundle of behaviors, and so on. Yet, the overlapping or almost simultaneous character of the customer’s conduct with the waiter or serving robot’s conduct would be lost. In other words, such a framework artificially forced the “multiple temporalities of language and body in interaction” (Mondada, 2018) into a step-by-step format.

Section 4.4.4 (above) presented a pared-down illustration of how a waiter walked away from a table. Figure 4.4 reports this “departure” in the form of a simplified script that does not include the customer’s behavior. Yet, even this outline of a waiter leaving a table was impossible to represent in this Figma framework. Indeed, the original contribution of a micro-analytic approach to these “departures” was to highlight that waiters do several things at once: they say “enjoy your meal” as they turn away and start to walk, they finish saying “enjoy your meal” even when they no longer see the customer, etc. Yet, because the Figma framework did not allow one to specify the exact timing of the robot’s actions, these nuanced behaviors of human waiters were not transposable into behaviors for the robot (whether they were technically implementable or not). Additionally, the discretization of the robot’s conduct in large “steps” in Figma (approaching the table, serving, leaving the table, etc.) obviously

corresponded to an etic division and verbalization of the uninterrupted flow of the robot's conduct. This division was useful for engineers and designers' practical purposes (to "brainstorm" around these specific behaviors, to program these conducts on the robot, etc.); yet, it did not intend to represent how the robot's conduct was oriented to by humans in situ. In sum, this mode of representation (this "technology of the intellect"; Goody, 1977) was preadapted to a high-level, non-sequential, design of interactions.

Crucially, the previous analysis does not constitute criticism of the mode of representation of interactions during my several years at Aldebaran: as I attempted to show, this mode of representation was meticulously adapted to the local concerns of designers and engineers. What I try to highlight is that *such frameworks were tracks formed by the cumulative passage of multiple designers. They produced a "path of least resistance" which subtly steered the design of multimodal interactions in specific ways.* As "desire paths" built incrementally, they did not necessarily embody an explicit and coherent definition of what an interaction between a robot and a human is or should be. They did not result from a public discussion or meeting during which designers explicitly agreed that only these features of an interaction were relevant, and none other. However, the resulting path of least resistance enforced an orientation to human-robot and human-human interactions as composed of discrete and simple steps. It facilitated the design of conducts based on an implicit understanding of human-robot interaction as a strictly "step-by-step" process: each interactant produced their "turn", then waited for a response. On the one hand, these constraints ensured that what ended up being designed was technically feasible. On the other hand, they complicated the task of pushing the robot further in its limited abilities, making it difficult to identify and represent situations in which this robot could respond to more subtle human behaviors.

4.6. The agency of programming and scripting tools for creating "conversations"

A central element establishing a "path of least resistance" (towards interactionally simplistic designs) was the set of technical tools required to program these very designs on the robot. When designers and developers programmed behaviors or response rules directly on the robot, they did not do so in a technical vacuum; they were constrained by the programming tools and documentation provided to them. As before, I will try to highlight that, when creating dialog trees or crafting Pepper's speech using the latest tools available, the "path of least resistance" led straight to the design of simplistic behaviors, far removed from what characterizes natural talk in conversational settings. In other words, programming tools (and their documentation) were especially adapted to the production of dialogs that displayed "unnatural uses of natural language" (Moore & Arar, 2018) – even before these scripted dialogs were put to the test of a naturally occurring interaction. The constraints arising from these tools, along with many other parameters, stripped the robot's conduct of some practices typically accomplished during talk-in-interaction: they contributed to worsening the status of the robot as a "beginner" conversationalist that neither perceived nor responded to most features ordinarily oriented to as relevant by humans. That is, the restrictions set by the tools were layered on top of (and were partially independent of) the robot's pre-existing technical constraints (what its sensors could perceive, what its algorithms could process, etc.).

4.6.1. Crafting disfluent speech and filled pauses (“uhh”)

Whether for external developers and designers or for internal employees of Aldebaran, the robot’s verbal conduct could be programmed on Android Studio through the Pepper SDK plug-in. This plug-in facilitated the use of a Software Development Kit (SDK), named QiSDK⁷⁵, which served as a framework that enabled developers to directly interface with the robot. Crucially, the Pepper SDK plug-in included various additional tools, among which a “Chat editor” that relied on a chatbot scripting language produced by Aldebaran: the QiChat⁷⁶ language. In turn, this language allowed designers or developers to program the robot’s embodied and verbal conduct: what it said, how it moved, in response to which high-level information (processed and discretized from its sensors and language recognition module), etc. Doing so, the robot’s multimodal behavior could be synchronized and made responsive to the outside world: for example, upon detecting a human recognized as “interested”, the robot could be programmed to start a “hand waving” animation, to say “hello” (or another randomized greeting term), and to simultaneously display a menu to “check-in” on its tablet. Significantly, among other possibilities, the QiChat language allowed designers to use several “Text-to-Speech tags”⁷⁷. With the help of these tags, the prosody of the robot’s utterances could be refined: as with many recent text-to-speech modules, designers could play with the speed of an utterance’s delivery (e.g., by adding “\rspd=70\” before a strip of words), its style (“joyful”, “didactic”, “neutral”), the emphasis on each word, and even with the duration of the pauses during a sentence (e.g., by indicating “\pau=100\”).

After discussing with several engineers who used or modified the QiChat language, a consensus emerged that disfluent talk (e.g., “I’m not uh, you know, su- I’m not sure”) including, in particular, filled pauses (e.g., “uh:::”), was especially challenging to craft for Pepper and was scarcely documented⁷⁸. Several features of the QiChat language, of the QiSDK, of the overarching technical system that constitutes the robot, and of its documentation imposed significant constraints on the production of behaviors from Pepper that would feature disfluent speech. For a designer or a developer unfamiliar with Pepper (e.g., someone who just bought and received the robot), these constraints can be summarized as follows:

- 1) *No existing vocal filler (“uh:::”, “uhmm”, etc.) was pre-made* and readily available to a developer or designer who would like to add such fillers to the robot’s speech.
- 2) *There were heavy constraints weighing on the production of filled pauses when scripting a conduct for the robot.* The existing tags (those that could be interpreted by the text-to-speech system) did not allow, by themselves, to create naturally sounding filled pauses, hesitations, or self-repairs – such as “you:: uh how are you?”. “Tricking” the robot into pronouncing specific phonemes by writing them in plain letters (e.g., “euuuu”) generally did not work either. Hence, to craft filled pauses that did not sound artificial, there was no choice but to use the phonetic

⁷⁵ <https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/index.html>. Note that the publicly available version of the QiSDK offered fewer possibilities (it could call to fewer functions) than the one available internally. The argumentation that follows takes into account both of these versions.

⁷⁶ https://android.aldebaran.com/sdk/doc/pepper-sdk/ch4_api/conversation/qichat/qichat_index.html

⁷⁷ <https://www.aldebaran.com/developer-center/lesson/Mastering-Prosody/l-dlg-035-mastering-prosody-step2/index.html#advanced-tweaks>

⁷⁸ What follows is based on my discussions and experience from 2019 to 2024. The tools and documentation may have changed since then.

mode of the QiChat scripting language⁷⁹, i.e., to use a modified version of the International Phonetic Alphabet that could be read by the text-to-speech system. Finely tweaking the prosody using this phonetic alphabet provided better chances of obtaining an “irregular” language (Relieu et al., 2020) able to suggest that the robot’s speaking turns were being constructed in real-time (rather than scripted beforehand). Yet, enabling Pepper to produce a “uh::” required experimenting with this phonetic language for a long period before achieving a satisfying result in terms of prosody. As was stated emphatically by the only documentation page that detailed the uses of phonetic language in QiChat: “one does not simply get the perfect text-to-speech tweak”⁸⁰. Additionally, adding hesitations in the middle of a sentence was particularly difficult. When a short vocal filler cut a turn in half (e.g., “My sensors are not, uh, so new”), this tended to sound like two separate turns (“My sensors are not. Uh. So new”). The same problem arose when crafting self-repairs (e.g., “Are you goo:: are you alright?”), as well as more complex bricolage turn beginnings (e.g., “hhh U:m↑last wee:k-h (0.5) ↓uhm: (0.8) kh uh ↑dere ↓we:re (0.3) ↑sree funerral:s; (.) in duh churchh” – excerpted from Gardner, 2015). Instead, a designer with limited time or access to internal documentation was led to make do with a perfectly “clean” (and, arguably, artificially sounding) language for the robot.

3) *The phonetic mode was barely documented*, either internally or online (for external designers or developers using a Pepper robot). For example, the nominal International Phonetic Alphabet symbols could not be used directly when writing a script in QiChat. One had to rely on a version of the International Phonetic Alphabet symbols that could be interpreted by the text-to-speech system: the L&H format⁸¹. In this format, the International Phonetic Alphabet *ʊ* was written U, etc. Hence, to translate the International Phonetic Alphabet symbols into symbols that could be read in a QiChat script, a correspondence table had to be used. Yet, until recently, this table was not referenced in the official documentation for Pepper or Nao robots. It was, so to say, hidden in the depth of Aldebaran’s documentation. At the time of writing, this table is now accessible, although far from easy to find: a link to the L&H format is provided at the bottom of an official documentation page dedicated to the QiChat⁸² – publicly accessible from a search engine. In sum, little resources were provided for a designer attempting to produce even a simple elongated consonant: one had to “hack” their way into these irregular ways of speaking⁸³.

4) *The QiChat language itself conflicted with the phonetic alphabet* required to produce filled pauses. The “L&H” format for the phonetic alphabet contained the symbols “%”, “+”, “~” or “\$”; however, these symbols were simultaneously used by the system to run the QiChat scripts. For example, “%” was also a symbol for “bookmarks” in QiChat, “~” was a “concept call”, “\$”

⁷⁹<https://www.aldebaran.com/developer-center/lesson/Mastering-Prosody/l-dlg-035-mastering-prosody-step2/index.html#phonetic-writing-all-languages-except-japanese>

⁸⁰<https://www.aldebaran.com/developer-center/lesson/Mastering-Prosody/l-dlg-035-mastering-prosody-step3/index.html>

⁸¹https://www.aldebaran.com/developer-center/lesson/Mastering-Prosody/assets/International_L-H%2B_Tables/PhonemeTable_enu_400.html

⁸²<https://www.aldebaran.com/developer-center/lesson/Mastering-Prosody/l-dlg-035-mastering-prosody-step2/index.html#international-lh-tables>

⁸³ As far as hearsay goes, several employees confirmed that this lack of documentation was intentional, to discourage external developers (e.g., clients who were renting a Pepper robot) from producing bad-quality prosody. When prosody was not properly tweaked (although an attempt was made), the result could be much more disturbing than plain default prosody. In short, the idea behind the absence of publicity for options to tweak prosody was “better safe than sorry”.

was a variable, etc.⁸⁴ When these symbols were used (see Figure 4.7), even between “phonetic mode” tags, a script in QiChat language could not be launched.

The image shows a dark-themed code editor with a light-colored text box containing the following text: `\toi=lhp\ a?E+%~m \toi=orth\`. The characters are in a monospaced font, and the text is highlighted in a light blue color. The background is dark, and there are some faint grid lines visible.

Figure 4.7. The sound “uhm” transcribed in “L&H” phonetic alphabet. The “+”, “%” and “~” symbols conflict with their use within the QiChat syntax

Overall, the technical tools available to program Pepper’s behavior favored the synchronization between the robot’s speech, movements, and tablet over the refinement of the robot’s prosody. To the best of my knowledge, after discussing this topic numerous times over the years, internal developers were seldom aware of the points listed above, as experiments with the phonetic alphabet had been extremely rare. For a time at Aldebaran, the idea surfaced of creating an Android Studio plug-in that would offer a built-in solution to add pre-recorded fillers directly into QiChat; yet, this project was abandoned. The convergence of these difficulties heavily pressured any attempt to design vocal interactions towards a simpler form of robot’s speech. It was not impossible to produce disfluent speech, but the sum of efforts required constituted a powerful incentive to make do with a plain and monotonous elocution style for the robot. This, in turn, provided one less resource for the design of social actions. Designers who would not go through the tedious work of grasping how to produce elongated consonants on Pepper would have to renounce to a significant part of the resources they used in their own conversations. For them, some of the “ordinary interactional knowledge” (Rollet & Clavel, 2020) – on which a human designer can be expected to rely when programming interactions for a robot – was impossible to transpose directly onto Pepper, as some crucial multimodal resources were missing. To pick a tangible example among many, while working under significant time constraints, I once had to abandon my attempts to use “uh” to mark the “reason-for-the-interaction’s-launching” (Schegloff, 2010b), as visible in the following strip of talk presented by Schegloff (2010).

```
(11) Susan & Marcia, 1 (#1)

00 ring
01 Sue:   H' llo:
02 Mar:   Hi: 's Sue there?
03 Sue:   Yeah, this is sheç
04 Mar:   Hi this's Ma:rcia.
05 Sue:   †Hi Marcia, how're you:.=
06 Mar:   =Fine how're youç=
07 Sue:   =Fiineç
08 Mar:-> Uh::m: We got the tickets, [and'a (   ) ] put them in=
09 Sue:                                     [Oh goo:d. ]
```

⁸⁴https://android.aldebaran.com/sdk/doc/pepper-sdk/ch4_api/conversation/qichat/qichat_syntax.html#special-characters

10 Mar: =envelopes:: >y'know with everybuddy's name on em< =

(Schegloff, 2010b)

Similarly, in the hypothesis of a designer or developer who had not found out how to create proper disfluencies, numerous documented uses of such disfluencies stayed out of reach. In this case, the robot could not be scripted to use “uh” as part of a self-initiated repair (H. H. Clark & Wasow, 1998) like “I went to the m- uh I went to the pool”, as a resource for “exiting a sequence” (Schegloff, 2010b) or for displaying the dispreferred character of a response as a second pair part (Schegloff, 2010b), or merely as a floor holding strategy (i.e., to communicate “I’m still in control – don’t interrupt me”; Maclay & Osgood, 1959), etc. In sum, disfluency, as an interactional resource – rather than as the mere reflection of “ongoing internal cognitive processing” (Morita & Takagi, 2018)⁸⁵ – was not readily available to an interaction designer. It had to be pieced together by struggling beyond the paths shaped by preexisting design and programming practices.

4.6.2. Making the robot “continue speaking” after a silence

4.6.2.1 Creating optional “response slots”

Another constraint weighed on the robot’s design: the relative difficulty of programming the robot so that it would “continue” speaking if no listener had been selected or had self-selected to speak. In human-human conversation, Sacks et al. (1974) famously identified a set of rules that underpin the selection of a next-speaker: they specify different possibilities and normative obligations that participants demonstrably orient to when they are engaged in ordinary conversation. These turn-taking rules apply when a “transition relevance place” occurs, that is, when “a point of syntactic, prosodic and pragmatic completion” (Levinson & Torreira, 2015) has been reached for a turn. They are as follows⁸⁶:

1. “The current speaker may select a next speaker (other-select), using for example gaze or an address term. In the case of dyadic conversation, this may default to the other speaker.
2. If the current speaker does not select a next speaker, then any participant can self-select. The first to start gains the turn.
3. If no other party self-selects, the current speaker may continue.” (Skantze, 2021)

My interest here lies in the obstacles faced by a designer or a developer when programming a Pepper robot to “apply” rule 3 (labeled rule 1c in Sacks et al.’s original account): the possibility for a current speaker to continue speaking if no one has been selected (rule 1) and if no one has self-selected to become the next speaker (rule 2). I will attempt to show that,

⁸⁵ “Although disfluency in human verbal communication has been generally considered an epiphenomenon of cognitive load in speech production, and/or associated with difficulties in the planning process, many conversation analytic studies have proven that such disfluencies in speech are orderly, and thus available to both the speaker and the recipient as interactional resources.” (Morita & Takagi, 2018)


⁸⁶ For the sake of brevity, I am using Skantze’s (2021) synthetic formulation here rather than the (slightly more technical) original phrasing from Sacks et al. (1974).

although it was technically possible to configure the Pepper robot so that it would “continue” speaking after a silence (no matter how participants would interpret this conduct from the robot in actual encounters), the design of such behaviors was neither facilitated in the code nor covered by documentation. Creating “optional” response slots to the robot’s answers – after which the robot would continue speaking if no interlocutor started to speak – was to be figured out alone by the designer or developer.

Instead, in the overwhelming majority of Pepper’ uses (as an information robot in museums or in stores, as a receptionist in office halls, etc.), once the robot’s turn reached completion, only the human could advance the interaction further by uttering a new turn or by touching the robot’s tablet⁸⁷. In the absence of a new turn from its interlocutors, the Pepper robot would silently gaze at them until they left. Note that this conversational design matched what was expected from Pepper robots in most “use cases”. Because it contributed to rendering the robot’s behavior predictable and unobtrusive, *not continuing after a silence fulfilled several established UX guidelines*. Mainly, it ensured that the robot would not keep uttering new speaking turns addressed to uninterested humans (but mistakenly detected as interested) after it heard or misheard a human voice.

4.6.2.2 Purposely undocumented solutions for scripting self-continuations

Crucially, it was technically possible to script self-continuations using the QiChat language alone. Based on information stored in the robot’s memory, several “events” were raised for this purpose. Mainly, an event “NotSpeaking5”⁸⁸ corresponded to the situation (as perceived by the robot) where no human had spoken for 5 seconds. Subsequently, a rule could be created in QiChat so that, when this event was raised, the robot would produce a continuation turn. Used in a similar manner, a few other events (“NotSpeaking10”, etc.) allowed the robot to respond to different silence durations. In short, if external developers knew that the robot could directly provide them with events about someone not speaking, they could easily use this information to produce turns responsive to silences in QiChat (see Figure 4.8)



```
u:(e:Dialog/NotSpeaking10) are you still there?
```

Figure 4.8. Example of a potential continuation turn in QiChat, provided in former versions of Pepper and Nao’s documentation⁸⁹

Yet, like many other events, the existence of these “NotSpeaking” events was purposely undocumented. During a transition from the old versions of Pepper’s operating system (NAOqi 2.5) to the newest version of this operating system (NAOqi 2.9, which features the public

⁸⁷ This situation may reflect “the designer’s ordinary interactional knowledge” (Rollet & Clavel, 2020) in the sense of a theoretical understanding of human-robot conversation. However, this designer’s *practical* knowledge of conversation is likely to be dissonant with this “no self-continuation” design: most ordinary conversationalists display a practical mastery of identifying social situations in which it is relevant to continue speaking after a silence (Sacks et al., 1974).

⁸⁸ http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/dialog-syntax_full.html#Dialog/NotSpeaking5

⁸⁹ Retrieved January 24, 2024 from http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/dialog-syntax_full.html

version of the QiSDK), heavy cuts were made regarding what events and methods⁹⁰ external developers could use or document. Although the NotSpeaking event was not hidden in the code itself (it remained exposed and usable for external developers), it belongs to a batch of events that were intentionally left undocumented. These cuts or intentional absence of documentation responded to four concerns:

1. *Interaction design concerns.* Some events and methods were left undocumented or were removed because they were not deemed relevant to external developers' needs when designing applications or interactions with Pepper. That is, they were not seen as matching the typical set of tools an external developer would likely need to shape Pepper's conduct⁹¹. As I questioned developers who worked on the "NotSpeaking" events, the previous argument was overwhelmingly presented as a major reason for the absence of these events from the documentation. Making self-continuation turns after a silence did not fit the professional (business to business) use cases that were imagined for Pepper at that time. During the creation of the QiSDK, what mattered was to give developers the tools to produce quick and efficient interactions with humans, not to design long conversations. Silences had to be avoided altogether rather than trigger continuations from the robot. In the words of a developer who participated in these debates at the time: "In a B2B (business to business) context, the interactions were likely to be straight to the point. If no one responded to the robot, it was better not to insist rather than to keep the conversation going".
2. *User experience concerns.* A batch of events and methods were removed or purposely omitted from the documentation because they were considered potentially harmful to the interaction with humans (although not dangerous per se). Using these methods and events within badly scripted interactions could lead to misplaced or irritating behaviors from the robot. This was the case of events (like NotSpeaking5) that allowed for the production of clumsily designed self-continuation turns: these events made possible the creation of scripts where the robot would endlessly call out to spectators that were merely gazing at it or, alternatively, annoy the staff working nearby (e.g., in the case of Pepper positioned in a retail store) by repeating the same sentence because a human had been falsely detected.
3. *Stability concerns.* Some events and methods were undocumented because they had not been refined enough internally to be sure they worked as intended. Events like NotSpeaking marginally fit in this category.
4. *Security concerns.* Finally, some cuts were made for security reasons, so that the robot could not be programmed to produce unsafe behaviors (e.g., programmed by a hacker to hit someone). However, this last concern did not apply to events like NotSpeaking.

4.6.2.3 Existing practices when creating dialog trees

If developers or designers did not know about NoSpeaker5 or similar intentionally undocumented events, very few resources were left to produce self-continuations using QiChat alone. This was my case and, to the best of my knowledge, the case of many

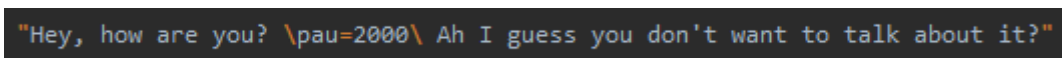
⁹⁰ Here, exceptionally, "method" is used in its computer programming sense: a function associated with an object.

⁹¹ Indeed, that streamlined set of tools is precisely what the QiSDK was developed to provide.

developers who started scripting behaviors on the latest versions of Pepper⁹². That is, the documented tools required to make the robot follow even a minimal continuation logic (to produce a new turn when it did not hear any recognizable talk in the few seconds after it finished speaking) were not provided in the QiChat language. The code of the Android application that was being created had to be modified.

However, because the robot's ability to be interrupted offered a quick “hack” to the designer, two robots' configurations must be distinguished:

- *The robot could be interrupted.* This feature was tested internally on the most recent versions of Pepper, but it was not available to the public. In this configuration, in an attempt to simulate a minimal “self-continuation”, a designer could write a single “turn” (in QiChat language) as follows: a first utterance, a long pause, then a second utterance likely to appear as a relevant continuation to the first (see Figure 4.9). If no participant responded to the robot during this pause, the robot’s ongoing turn would not be “interrupted”, and, by finishing this turn, the robot may appear to produce a continuation after a gap. Conversely, if a participant responded during this pause, the robot would interrupt its turn (which, to this point, was still ongoing in its script) and respond to the turn that this participant just produced instead. This “cheap” way of producing self-continuations arguably offered very limited options; yet it was directly applicable within a QiChat script.



```
"Hey, how are you? \pau=2000\ Ah I guess you don't want to talk about it?"
```

Figure 4.9. A two-second (2000ms) pause during a speaking turn of the robot, composed in QiChat language. Although it is scripted as a pause (i.e., as a silence occurring when the turn is still ongoing in the script), this silence may be treated as a “gap” by human listeners (i.e., as a slot to produce a response).

- *The robot could not be interrupted.* This constraint was common to all public versions of the Pepper robot. In these cases, even a “fake” self-continuation turn like the one described above was not possible to create in QiChat. One would have to modify the code of the Android application, with various solutions available to make the robot “listen” to silences or, alternatively, to create “fake” continuations by, e.g., directly adding a timer in the code of the Android application.

In both cases, if the robot were to be made responsive to silences (rather than merely simulate a continuation like in the quick “hack” example provided) *and* that a developer did not know about purposely non-documented events like `NotSpeaking5`, the code of the Android application itself had to be modified. Responding to silences was not officially handled by the QiChat language. Significantly, after asking Aldebaran’s engineers (who, like me, were unaware of the possibility of using events like “`NotSpeaking5`”), we required an hour of trial and error to reach a result that functioned properly on a Pepper robot. This experience suggested that creating an application that made Pepper produce self-continuations was feasible for a developer entirely unfamiliar with programming Pepper, yet not without substantial effort. Quoting a developer who worked on the QiChat language, “this is not difficult to do, but if developers are lazy, they won’t do it”. Additionally, several internal developers who

⁹² By “latest versions of Pepper” I mean those that could be programmed using the QiSDK. This corresponds to Peppers whose operating system was NAOqi 2.9.

had often programmed different kinds of behaviors on Pepper robots had never faced the need to program the robot so that it would start a new speaking turn if nothing was responded to its immediately prior turn. In spite of its daily occurrence in human-human interactions, *this typical conversational practice was generally absent in the documentation, in the code or in the practices of most internal developers working on the latest versions of Pepper.*

Congruently, my initial (failed) attempts were met with a lack of documentation specifying how to “listen” to silences and/or to create “fake” continuations in an application on Pepper’s latest operating system (NAOqi 2.9). Across the various public documents detailing how to program recent Peppers’ conduct⁹³, only one referred to the possibility of creating a “timer” when the listener did not respond to a turn of the robot⁹⁴. Rather, this official documentation – because it focused on what was officially possible to script in QiChat language – exclusively provided examples of dialog structures with no “self-continuation” after a silence (see Figure 4.10). In other words, the turn-taking system underlying these dialog trees was exclusively composed of Sacks et al.’s (1974) rule 1 and 2 – quoted in section 4.6.2.1 above⁹⁵.

```
u:(input1)
Answer 1
  u1:(input2)
  Answer 2
    u1:(input3)
    Answer 3
      u2:(input4)
      Answer 4
        u3:(input6)
        Answer 6
          u2:(input5)
          Answer 5
```

Figure 4.10. Structure of a dialog tree described in the official documentation available online⁹⁶.

⁹³ At the time of writing, two publicly accessible websites focus in detail on the design and programming of Pepper’s conduct: <https://www.aldebaran.com/developer-center/lesson/Discovering-QiChat/index.html> (accessed January 20, 2024)

and <https://qisdsoftbankrobotics.com/sdk/doc/pepper-sdk/index.html> (accessed January 20, 2024). Both of those websites focus on the latest operating system for Pepper (NAOqi 2.9).

⁹⁴ “Adding a timer can also be an additional solution, especially for B2B when some users will leave unexpectedly. Repeat the question if the user hasn’t answered after 10 seconds, but be careful not to create an infinite loop.” (<https://www.aldebaran.com/developer-center/articles/Dialog-guidelines/index.html>, accessed January 20, 2024)

⁹⁵ “(a) If the turn-so-far is so constructed as to involve the use of a ‘current speaker selects next’ technique, then the party so selected has the right and is obliged to take next turn to speak; no others have such rights or obligations, and transfer occurs at that place. (b) If the turn-so-far is so constructed as not to involve the use of a ‘current speaker selects next’ technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.” (Sacks et al., 1974)

⁹⁶ Retrieved January 25, 2024 from https://www.aldebaran.com/developer-center/lesson/Discovering-QiChat/DevC-Lesson_L-DLG-032_Discovering-QiChat_step6/index.html.

For example, in an advanced conversational chatbot developed internally and that several engineers had dedicated many working hours to, the dialog tree structure was invariably the following (see Figure 4.11):

```
#rule: what does pepper mean#
u:([
  "que ["veut dire" signifie] Pepper ?"
  "Pepper ["c'est quoi" "ça veut dire quoi" ] ?"
  "ça veut dire quoi Pepper ?"
])
Littéralement, ça veut dire piment ou poivre, en anglais. On m'a appelé comme ça parce que je suis là pour mettre du piment dans ta vie!
```

Figure 4.11. Screen from a “rule” taken from a conversational application made for Pepper that uses the QiChat language. Upon hearing various requests about the meaning of its name, Pepper explains what it means and why it was given this name.

Whether this recurring dialog tree structure was a consequence of the lack of documentation or because the need was not encountered by any designer or developer, the overwhelming majority of the applications using the QiSDK I had the opportunity to see in action (or whose code I had access to) functioned exclusively on an “Utterance - Obligatory Response” basis. Among the different Pepper robots that I could observe in action between 2019 and 2024, I encountered only one⁹⁷ that could be described as continuing to speak if nothing was responded after its previous utterance. That is, Pepper robots were generally not scripted to conform to a high-level interaction flow such as:

1. After a turn is produced by the robot, the robot listens for a response.
2. If no response is heard, the robot produces another turn (defined in advance), provided that a human is still detected by the robot.

4.6.2.4 Designing interactions without responding to gaps or lapses

As evidenced above, when a Pepper robot was not configured to detect or respond to silences, the designer was deprived of a crucial resource for organizing an interaction: the possibility to continue when no one had self-selected after a turn reached possible completion. To put it differently, a Pepper “unresponsive to silences” could not be programmed to even loosely conform to the turn-taking rule mentioned above (rule 3): “If the turn-so-far is so constructed as not to involve the use of a ‘current speaker selects next’ technique, then current speaker may, but need not continue, unless another self-selects” (Sacks et al., 1974). Yet, maybe more significantly, such a robot could not orient to “gaps” (silences between a first and a second pair part) or “lapses” (silences that “emerge at the end of a sequence of talk” – Hoey, 2017): different types of silences could not be connected to different normative obligations or possibilities⁹⁸. That is, because it could not identify silences altogether (whether they were pauses, gaps, or lapses), the robot could not be designed to respond to some of these silences as noticeable absences of talk (Hoey, 2020a): it could not recognize the occurrence of “silence at points where talk should occur” (Seuren et al., 2021).

⁹⁷ Namely, a version of the receptionist Pepper robot placed in Aldebaran’s office entrance.

⁹⁸ Drawing upon Hoey (2020) and Sacks et al. (1974), Seuren et al. (2021) formulate a synthetic definition for these two terms: “Participants treat silences in these respective environments differently. *When turn-transition should happen a gap emerges* (i.e., a failure by the recipient to take a turn in which to produce the projected action). *When turn-transition could happen, a lapse emerges* (i.e., the recipient foregoes an opportunity to talk).” (emphasis mine)

Hence, the “Utterance – Obligatory Response” conversational trees (see Figure 4.10) typically produced on Pepper had potentially far-reaching consequences for a designer. As Seuren et al. (2021) note, “[t]here is a range of practices participants have at their disposal to address these silences [...]”. If a designer had not found a way to turn silences into units that can be responded to by the robot (or, at least, to “fake” responses to silences), *an enormous amount of ordinary conversational practices suddenly became inaccessible to this designer*. Let us consider some of these practices and their feasibility for a robot “unresponsive to silences”.

4.6.2.5 Some ordinary practices that index silence in conversation

First of all, and although this remark encompasses a wide range of specifiable practices, a Pepper “unresponsive to silences” cannot (be programmed to) make gaps accountable (Hoey, 2020b). In this configuration, it cannot be designed to produce accounts indexed to its interlocutors’ non-answer to an immediately prior turn – when this prior turn was composed as a first pair part. That is, one less tool is available for the robot to “reinforce” the conditional relevance attached to its own turns. Of course, in situ, the robot’s conduct may be *understood by participants* as a continuation: chapter 5 in this work provides several examples of participants treating a robot’s waving gesture as accounting for a missing action on their part after a silence. Yet, crucially, in these strips of interaction, the robot’s conduct was not intended as an account by the designer who scripted it. For example, as long as a robot does not respond to silences, it is not possible to intentionally design the robot to co-produce a sequence like the following (imagined) one:

- 1.ROB hello
2. (3.0) ((the human facing ROB stays silent))
- 3.ROB you don't say hi to me?

In this hypothetical dialog, ROB’s question (L.3) reinforces the conditional relevance attached to its immediately prior greeting (L.1), thereby configuring its interlocutor’s silence (L.2) as an accountable absence of response. Doing so, the robot makes visible the normative obligation to produce a response of a specific type and form (Kendrick et al., 2020) to its own turn – here, a greeting. In other words, it produces a practice that orients to “silence as the noticeable absence of a conditionally relevant response” (Seuren et al., 2021).

Another example concerns the production of a “sequence recompletion” (Hoey, 2020a) after a lapse, i.e., a practice that “recompletes the prior sequence and allows [the robot] to pass on the opportunity to initiate a next one” (Hoey, 2020a). If Pepper does not react (nor simulate reacting to) silences, it becomes much more arduous for the designer to program conducts likely to be oriented to, in situ, as displaying such a “sequence recompletion”. In other terms, in a strip of talk whose relevant features resemble those in the following fragment (excerpted from Hoey, 2020a), a Pepper deaf to silences could not be programmed to utter “nice” (L.7) or “yeah” (L.11) in response to a lapse in the conversation (L.6 then L.10):

(2) RCE28_0803
01 KEL: and then like (.) she'd completely forgotten that it
02 was his birthda:y, and I was like >how long have you
03 (been) married- >she's like< .MTKh like fifty yea:::rs,
04 KEL: [>%I was like< [OOO*H +gheheh
05 HEA: [((gasp)) (.) [HAHAHAHA*::+: .H
kel +withdraws gaze->
*withdraws gaze->
06 (1.1)
07 HEA: N:i:ce.
08 (0.3)
09 KEL: °mh mhh°
10 +(0.5)*(0.6)
hea ->+inspects fingernail->>
kel ->*gazes to HEA's hands->>
11 HEA: Yea:h_
12 (2.1)

(Hoey, 2020a)

Finally, to provide one more example among many, a Pepper unresponsive to silences cannot be programmed to use a tag question as an “exit technique for a turn” (Sacks et al., 1974)⁹⁹ when no listener is self-selecting to produce a next turn after Pepper finished speaking. For example, such a robot cannot be designed to ask “Don't you agree?” (Sacks et al., 1974) if, following a turn where it produced an assessment, the human interlocutors have not responded within a certain timeframe.

4.6.2.6 Unforeseen repercussions on the range of social actions that the robot can perform

I have briefly gone over a few ordinary practices that index “gaps” or “lapses” in conversation. With this eclectic and non-exhaustive list, the point I am trying to illustrate is the following: In the hypothesis of a designer or developer with important time constraints or limited technical abilities, the lack of a ready-to-use solution for responding to silences in an interaction *may surprisingly restrict the range of social actions that the robot can be programmed to successfully perform*. A cascade of interactional repercussions can stem from a seemingly trivial technical hurdle. On a robot where silences are not treated as significant units, the designer cannot rely on a typically relevant element of talk-in-interaction to script conducts that may be treated, by participants in situ, as one of those practices listed above. This problem would persist if the robot being used did not feature a basic rule-based chatbot but, instead, a more advanced conversational system. Even then, this robot would be unable to perceive a feature of talk (silences) typically attended to as relevant by human conversationalists. Without being configured to treat silences as “social objects in their own right” (Hoey, 2020a), a robot

⁹⁹ “That is, when a current speaker has constructed a turn's talk to a possible transition- relevance place without having selected a next, and he finds no other self-selecting to be next, he may, employing his option to continue, add a tag question, selecting another as next speaker upon the tag question's completion, and thereby exiting from the turn.” (Sacks et al., 1974)

is deprived of the possibility to identify, respond to, or produce a wide range of conversational practices, or, for that matter, any practice that indexes a silence during an interaction.

4.6.3. The agency of technical tools

The previous analyses are but one drop in a large literature on the agency of tools (Battentier & Kuipers, 2020; Verbeek, 2006). Yet, I hope to highlight how, in the specific setting of a robotics company, some programming and scripting tools specifically conveyed an “unnatural use of the natural language” (Moore & Arar, 2018). These tools (along with what their documentation emphasized or overlooked) heavily impacted the ease with which designers could design specific forms of talk. Crucially, because I worked directly from Aldebaran’s office, the programming tools I used offered at least as many possibilities (and sometimes more) than their publicly available version: provided they wanted to refine Pepper’s speech or turn-taking rules, hundreds of users (developers creating commercial applications on Pepper, academic researchers using Pepper for experiments, etc.) were subjected to the same constraints and “paths of least resistance”.

However, it is unclear to what degree exactly some programming constraints reified specific representations about what *should be* an interaction with a robot. At least, such a claim is not rigorously possible to refine with the information at hand. That is, I am in no way able to specify to which degree the restrictions of these programming tools – in particular, what was not provided as ready-to-use solutions for unfamiliar developers – were the neutral reflection of “what was complicated to code” or “what could cause security issues” rather than “what did not appear relevant to be provided as a tool”. On this matter, technical concerns complexly intertwined with design considerations.

4.7. Divergent assumptions as to what an “interaction” is

4.7.1. Degree of granularity and interactional consequences

The previous data provided some clues as to how programming and design tools can produce “paths of least resistance” towards specific types of designs or turn-taking systems. I would like to stress that these paths of least resistance are likely to have concrete interactional consequences. For example, going back to the example of the typical “exit” produced by a waiter after serving a table (see Figure 4.4 above), two distinct practices were originally identified:

- a) In the first case, the waiter responded to an immediately prior turn *initiated by the customer*: the waiter slowed down and gazed back at the customer *in response* to a “thank you”, a request, an embodied display of trouble (Drew & Kendrick, 2018), etc., produced by this customer.
- b) In the second case, the *waiter themselves initiated a request or a question while looking back* (e.g., “do you need something else?”). That is, the waiter did turn back and speak to the customer before leaving, but this last glance and speaking turn were *not*

observably responsive to a verbal or embodied action produced by this customer and monitored by the waiter.

It was then described how both these (already rather roughly defined) practices were aggregated as one generic “exit sequence script” in the final report. In this “exit sequence script” (see Figure 4.4 above) the waiter:

1. Gazed at the customer after serving the last dish.
2. Started to move away from the table, orienting their shoulders and head towards the direction in which they were going.
3. Partially turned back to gaze one last time at the customer, while slowing down or stopping their movement.
4. Asked a question or simply gazed at the customer.
5. Faced forward again and left.

During the process leading to the list of steps reproduced above, the micro-sequentiality visible on the original recordings was erased (see section 4.4). Yet, crucially, such a loss of relevant features should have, in all likelihood, *tangible consequences on the design of a prototype serving robot* based on this discretization of a waiter's activity and, through this design, on the local interactions that will ensue. If, because the granularity of an empirical report has been too low to preserve this distinction, an engineer or a designer cannot differentiate between the two situations a) and b) presented above (i.e., between a customer attempting to open a new speaking turn *versus* a customer aligning with a closing sequence), this designer or engineer is unlikely to devise behaviors for the robot that will be relevant in *both* of these cases (i.e., that will “mould” – Relieu et al., 2020 – in both these sequential environments). Here, in the particular “exit script” that was agreed upon (see section 4.4) and reproduced above, the robot's conduct has every chance to clash with a customer's attempt to extend the interaction – unless some unexpected sequential context happens to make this conduct occasionally and fortuitously relevant. This “exit script” does not anticipate embodied and/or verbal requests of assistance of the customer, thereby running the risk of being interpreted by co-present humans as ignoring a customer producing such a request.

Crucially, in the absence of fine-tuned adjustments on its part, the robot will be unable to re-accomplish what was visible in the very recordings that inspired its “exit sequence script”: the waiter's *response* to the customer's public display that their interaction was still ongoing – or, on the contrary, that it was coming to an end. The core phenomenon that originally prompted researchers' and engineers' curiosity will be lost. In sum, the resulting design will go from *sequentiality* to mere *successiveness*. It will crystallize in an invariable form (e.g., in the form of a scripted “exit sequence” for serving robots) a conduct that resulted, in situ, from the fine-tuned reflexive adjustments of co-present members. That is, from *mutually responsive conducts contingent on the local situation* (the sequential nature of these conducts), what will be retained is exclusively *the temporal succession of these conducts* (i.e., that one conduct – understood as a bundle of ready-made actions – followed another *without being constructed in response to it*).

4.7.2. From sequentiality to “successiveness”

This representation of a “human-robot interaction” as constituted of discretized steps that merely follow one another (see sections 4.4 and 4.5 above) falls under the same criticism as

Goffman's (1979) account of “building utterances” reported by Goodwin (2007). According to Goodwin (2007), Goffman (1979) conceives the work of “building utterances” as resting exclusively on the speaker – a speaker whose role and activities are insulated from those of the hearers. In this understanding, because the PowerPoint slides examined earlier in this section (see Figure 4.3 and Figure 4.4) merely mention different participants (waiter, customer, etc.) without transcribing or reporting their conduct,

“[p]arties other than the speaker are thus excluded from analysis. The crucial mutual reflexivity of speakers and hearers is lost. It becomes impossible to investigate how utterances are built through processes of interaction that include the participants ongoing analysis of each other. In essence *the world being analyzed is lodged within a single speaker’s speech.*” (Goodwin, 2007; emphasis ours)

In opposition to this perspective, Goodwin (2007) argues that “utterances are constructed through processes of interaction in which different kinds of participants are building action in concert with each other”. This “essential mutual reflexivity of speaker and hearer(s)” emphasized by Goodwin (2007) is a significant feature of what Mondada (2024) labels “micro-sequentiality”. That is, in social interaction, “responses are often achieved not just within strictly successive actions (in which an action is followed by a response), but within largely simultaneous actions, in which *an action is immediately responded to, while it is still ongoing, and where this response can reflexively affect the way that that action is continuously produced*” (Mondada, 2024; emphasis mine). This reflexive responsiveness of actions is precisely what is lost in the modes of representation reported above.

4.7.3. Composition over position: “interaction” as questions and answers

Another tacit definition of “human-robot interactions” is visible in (or, at least, enforced by) the design and programming tools discussed in section 4.6 above: the interaction with a robot as following a “question-answer” format rather than a “conversational” format¹⁰⁰. Indeed, the descriptions of “paths of least resistance towards specific conversational formats” provided in this section are congruent with Relieu et al.'s (2020) observations about pervasive consequences of the historical popularity of the Turing test. Relieu et al. (2020) posit that the well-known “question-answer” format of the Turing test tacitly encouraged chatbot designers to focus on “creating chatbots able to answer epistemic questions, as a human would do, rather than to hold a conversation such as those we have in our everyday life” (Relieu et al., 2020 – translation ours). They remark that, to this day, a chatbot is typically viewed as “specialized in the production of answers that follow questions” (Relieu et al., 2020 – translation ours). Hence, the aforementioned features of some tools may stem from a more general tendency of current chatbot systems and conversational design software to *overemphasize turn composition (what to say) while neglecting turn position (when to talk)* – Schegloff, 2010. I believe that these divergent concerns may account for some of the

¹⁰⁰ This question-answer format may display observable similarities with the organization of talk often accomplished in encounters described as “interviews” by involved participants themselves (Heritage, 2005; Schegloff, 1992a), or with other types of institutional talk (Heritage, 2005) – although, the extent to which “human-robot interactions” currently display a typical *normative* organization of talk that overlaps with certain types of institutional talk (or if the terms of such a comparison are even meaningful when artificial agents are involved) is still an open question.

ineffability occurring between EMCA practitioners and their interlocutors (be they engineers, designers, psychologists, etc.) when debating conversational designs, systems, or data. Simply put, these actors index different practices and normative expectations when they speak about “conversing with a robot”, “interacting with a robot”, or “designing the conduct of a robot”¹⁰¹.

4.7.4. Are these issues relevant for the design of conversational chatbots relying on generative AIs?

Might the previous observations be relevant, in one way or another, to the conception of chatbots based on generative AIs (or to their integration into robots)? Mainly, is the earlier mentioned issue (of identifying “when” to talk rather than “what to say”) likely to re-emerge (potentially under a different form) during the design or implementation of conversational chatbots relying on generative AIs – or, instead, are the most recent conversational technologies insulated from these issues? That is, can this pre-existing conception of what an “interaction” is (as a turn-by-turn succession of questions and answers) insidiously configure the organization of talk in which the most recent chatbots are able to produce relevant contributions?

Plainly, I suggest that, as long as human designers or programmers exhaustively and directly configure “when” conversational agents must speak – for example by setting up a user interface pre-adapted to a question-answer format, with no possibility for the agent to self-select after a silence, or to interrupt the human –, these agents will unavoidably reflect our theoretical notions (rather than practical mastery) of “what an interaction is”. If these pre-existing notions are narrow, they will insidiously constrain and limit the organizations of the interaction (i.e., “the interactional substructure beneath the conversational surface” – Bergmann & Drew, 2018) that the conversational agent can competently partake in. Yet, as long as developers or designers have some degree of control over a conversational agent’s turn-taking decisions (i.e., as long as one does not posit entirely autonomous agents able to relevantly adapt to different turn-taking systems), these developers and designers should be provided with tools to define, not only what the robot will say, but when it will say it. I argue that conversational tools should prioritize features that allow developers or designers to configure or tweak the type of *turn-taking system* enacted locally by the robot in each interaction. In other words, turn-taking should be as refined as the content of the turns – yet, only the latter has seen immense progress in the last years. As mentioned above, both *turn composition* (what is said) and *turn position* (where and when it is said) are essential to the accomplishment of an activity recognized and experienced by its own participants as a “conversation” (Schegloff, 2010b).

¹⁰¹ Throughout this section, I have tried to identify engineers’ and designers’ assumptions regarding what is an “interaction” or a “conversation” with a robot – or, at least, I have tried to specify what types of “interactions” these engineers’ and designers’ concrete practices, their design frameworks, and their tools, de facto enforce. In doing so, I have taken inspiration from Linell’s (2004) seminal work on the many assumptions linguists make about language. Note that there is a likely overlap (possibly grounded in their shared cognitivist background) between my very limited description of engineers’ and designers’ assumptions and those of the prevailing currents in linguistics studied by Linell several decades ago.

5. FROM OBJECT TO AGENT: THE EMERGENCE OF A ROBOT AS A (GREETABLE) SOCIAL PARTICIPANT

There came a time when her uraemic trouble affected my grandmother's eyes. For some days she could not see at all. Her eyes were not at all like those of a blind person, but remained just the same as before. And I gathered that she could see nothing only from the strangeness of a certain smile of welcome which she assumed the moment one opened the door, until one had come up to her and taken her hand, a smile which began too soon and remained stereotyped on her lips, fixed, but always full-faced, and endeavouring to be visible from all points, because she could no longer rely upon her sight to regulate it, to indicate the right moment, the proper direction, to bring it to the point, to make it vary according to the change of position or of facial expression of the person who had come in; because it was left isolated, without the accompanying smile in her eyes which would have distracted a little from it the attention of the visitor, it assumed in its awkwardness an undue importance, giving one the impression of an exaggerated friendliness.

—Marcel Proust, *The Guermantes Way*, 1920 (Trans. C. K. Scott Moncrieff)

Note: This chapter is substantially based on Rudaz et al., (2023), published in ACM Transactions on Human-Robot Interaction (THRI).

5.1. Introduction

As we have attempted to outline in chapter 3, a growing body of studies has recently approached the “social” status of a robot not as a “categorical property of the robot's inside” (Alač, 2016), or as a “specifiable and implementable set of features” (Licoppe & Rollet, 2020) but, instead, as an emergent phenomenon (Pitsch & Koch, 2010). These works follow an ethnomethodological perspective for which, like many seemingly objective and trans-situational properties of social life, the “social” status of a “social robot” is not given but “locally produced, incrementally developed, and, by extension, [...] transformable at any moment” (Heritage, 2005). By closely studying the moment-to-moment interactions of humans with a robot, researchers using these approaches attempt to grasp through which local processes a robot can, momentarily, emerge as a “social partner” (Straub, 2016), a “social actor” (K. Fischer, 2021; Pelikan & Broth, 2016), a “subject” or “social being” (Straub et al., 2012), a “social agent” (Alač, 2016), an “artificial agent” (Licoppe & Rollet, 2020) or as a new (and evolving over time) ontological category (Pitsch & Koch, 2010). Without necessarily sharing the same theoretical background, they rely on a common assumption that, in the same way

that “a situation becomes observable and is treated as the meeting of a jury when participants produce practices that others orient and respond to as practices of a jury” (vom Lehn, 2019), an entity becomes a “social robot” when it produces practices that humans orient and respond to as practices of a social robot.

Yet, this definition of sociality as an emergent feature “raises a question about the minimal conditions for a social interaction” (Jones, 2017). Indeed, once a robot is demonstrably shown to be oriented to as a “social agent”, there remains the issue of what shifted this robot’s interactional status; especially when it was approached as a non-social object at first. What properties of the local situation were observably oriented to as relevant for participants before a “social” interaction could emerge – suddenly or incrementally, momentarily or durably? This chapter extends on this ethnomethodological line of research by focusing on the initial emergence of a first, conditionally relevant, greeting. Based on an experiment from which were collected 80 recordings of dyadic interactions between a human and a robot, we try to identify 1) *if and when* a shift occurs from an inanimate artifact to an agent and 2) *how* this shift progressively emerges, when it does.

Of all the properties which could be constituted and oriented to as “relevant features” in the continuous flow of a local human-robot interaction, we wonder which of them were treated by participants as creating the adequate conditions to produce a first greeting move. That is, we attempt to describe the interactional work required before *behaviors* from the robot could be treated as *actions* which either established the adequate framework for the participant to initiate a first greeting sequence or, alternatively, produced a response slot that the participant was normatively pressured to complete with a return greeting. Overall, four typical paths to the emergence of a first greeting were identified, and one where the production of a greeting never became relevant. This typology is exemplified through the analysis of five fragments representative of our corpus.

5.2. Previous work

5.2.1. The significance of greetings for human-robot interactions

Human greetings, as a canonical part of the opening sequence in many interactional settings, are among the most documented practices in conversation analysis: from the first analysis of telephone calls (Schegloff, 1968, 1979a), to video calls (Licoppe, 2017; Mondada, 2015), to co-present encounters (De Stefani & Mondada, 2018; Harjunpää et al., 2018; Kendon, 1990; Mortensen & Hazel, 2014; D. Pillet-Shore, 2012; Robinson, 1998). Greeting sequences have been of special interest in the study of human-robot interactions, not only because of the amount of data available, but also because of what they accomplish in human-human interactions. Indeed, they simultaneously reflect and construct the mutual status of the co-interactants, by being tailored “to (display) their own understanding/appraisal of ‘who we are to one another right now’” (D. Pillet-Shore, 2012), manifesting, in particular, that another is “recognized and categorized as a possible partner for future interaction” (Mondada et al., 2020). They also tend to be connected with observable changes in the structure of talk and in the physical configuration of participants, as they often form both “the end of a phase of incipient interaction” and “the first exchange of a conversation” (Schegloff, 1979a). In

summary, greetings are critical for “organizational reasons (coordinated, well-tuned, reciprocal engagement), social reasons (recognition, display of the type of relationship the participants entertain), and normative reasons (mutual trust)” (Mondada et al., 2020).

The previous approaches focus on what greetings *do* in an interaction. In parallel, several quantitatively oriented human-robot interaction (HRI) studies have highlighted the significance of greetings as an *indicator*. Humans who greeted a robot were commonly found to display specific behaviors during the rest of their interaction and/or to hold specific representations or perceptions of the robot. Notably, producing a greeting was observed to be a predictor of a “more social script” (Lee, Kiesler, & Forlizzi, 2010), of “patterns of discourse” (Makatchev et al., 2009) or of specific linguistic behaviors (K. Fischer, 2011). Greetings were also shown to correlate with the attribution “of higher linguistic, perceptual, and cognitive competence” to the robot (K. Fischer, 2007) and, in particular, reacting to a robot’s greeting wave was observed by Baddoura & Venture (2015) to be significantly correlated with evaluating this robot as sociable – suggesting that responding to a greeting documents more than a simple mimetic reflex action. However, Holthaus & Wachsmuth (2021) demonstrated that the multimodal behavior of a humanoid robot can heavily impact the number of participants who greet it, and the timing of their greetings, highlighting that the first moments of the interaction can play a heavy role in constituting a framework in which a greeting sequence becomes relevant.

5.2.2. Conditional relevance as a breaking point in the robot’s status as an interactant

As “highly ritualized actions” (Stivers & Rossano, 2010b), greetings are also typical cases involving conditional relevance. This concept refers to a property that binds together two turns at talk – from different speakers – in an interaction (Kendrick et al., 2020; Schegloff & Sacks, 1973b): sequences of question-answer, invitation-acceptance (or refusal), etc. In each of these sequences, a first pair-part – e.g., a question – makes relevant for the recipient the production of “a second pair-part of the same sequence type” (Kendrick et al., 2020) – e.g., an answer. Hence, a greeting creates expectations of a reciprocal greeting, whose absence can be accountably oriented to by participants as “a meaningful departure from the norm” (Kendrick et al., 2020). Two turns united by conditional relevance (i.e., a “first pair part” and a “second pair part”) form an “adjacency pair”. Significantly, conditional relevance is not a mere statistical observation (that a first pair part tends to be followed by a second one of a certain type) but corresponds to the achievement of a “normative organization” (Kendrick et al., 2020) where a first action from a participant imposes constraints “on the type and form of action with which the recipient should respond” (Kendrick et al., 2020). In this sense, Jarske et al. (2020) argue that “when people respond to a social robot’s greetings, they do not merely respond to the robot, but orient to the moral obligation involved in the normative practice of greetings”.

Because of this documented property of greetings, the treatment of an action of a robot as initiating a first greeting pair part has been connected to the status of this robot as an interactant (Alač, 2016; Pelikan & Broth, 2016). A *response* to a robot’s greeting suggests that, at this specific moment, humans orient towards the robot as an entity that can impose normative constraints. Locally and momentarily, the robot stops being treated as an object performing an autonomous script – which is not inserted in an orderly sequence of conversational turns (Schegloff, 1968) – but is, rather, oriented to as an agent or partner

(Licoppe & Rollet, 2020; Straub, 2016) whose actions can have (normative) consequences for the recipient (Duranti, 2007; Jarske et al., 2020). The emergence of a sequence of (mutual) greetings may therefore constitute an observable breaking point in the interaction dynamic. When a participant initiates a first conditionally relevant greeting pair or responds to a robot's greeting action, this "enactment of the greeting ritual models the appropriate and expected way of acting and interacting that constitutes the addressee as a particular kind of entity" (Alač, 2016). Relying on the distinction from Jarske et al. (2020), a mutual greeting sequence with a robot can be said to display participants who are "talking" rather than "using speech": at this instant, the robot is treated as producing actions "discoverable within a normative order" (Jarske et al., 2020) and, conversely, as having the capabilities to interpret the normativity of other participants' actions (Jarske et al., 2020).

Crucially, mutual greetings emerge from a preexisting situation; they do not appear out of an interactional vacuum. Precisely because they enact the existence of a normative order over both the greeter and the greeted (Jarske et al., 2020), the appearance of greetings supposes a "framework in which a greeting sequence is relevant and expectable" by the participants (Mortensen & Hazel, 2014), or "a proper interaction frame" (Licoppe, 2017), usually established as part of the "pre-beginning" (Schegloff, 1979a) or "pre-opening" (Mondada, 2015). Yet, co-participants accomplish various degrees of interactional "work" to establish such a framework. In particular, for humanoid robots, emerging as "agents" is not systematically granted by the sharing of a mutual space with other participants. Robots may require, even more than humans, to achieve "the type of self-affirming done through language [which] is of a different nature from mere physical presence" (Duranti, 2007).

5.2.3. Pre-beginning designs in Human-Robot interactions: "coming into sight" vs "coming into existence"

5.2.3.1 Pre-beginning designs in HRI

Focusing exclusively on the moment at which the robot appears¹⁰² to participants for the first time, HRI studies and datasets collected in controlled or natural settings can, at first glance, be sorted into two general categories.

1. Studies where the robot stands motionless when participants encounter it, without displaying any preexisting idle behavior nor adjustments to the participants' approach or presence: the Wizard of Oz has to seat participants in front of the robot before going behind a divider to send commands to the robot (e.g., Beran et al., 2011) or has to deal with a significant response time (e.g., Scheffler & Pitsch, 2020; Thellman et al., 2017), the script is not launched yet (e.g., Pelikan & Broth, 2016; Tatarian et al., 2021 – as well as the fragments presented in this chapter), the autonomous robot's reactions are delayed, etc. In these situations, participants find themselves in physical copresence with the robot for a long period, before reciprocal exchanges and mutual identification become possible: there is a non-accounted for delay "between entry into physical copresence and moves to enter into social copresence" (D. Pillet-Shore, 2018).

¹⁰² I.e., when the robot enters the perceptual field of participants.

2. Studies where the robot, or virtual agent, is already observably involved in a preexisting activity when it appears to participants (including idling behaviors like simulated breathing, random head movements, etc. – e.g., Arias et al., 2020; Liu et al., 2016; Riddoch & Cross, 2021; Yang et al., 2021) and/or observably adjusts to the humans' approach or physical co-presence (tracking their gaze, waving, producing a non-delayed greeting, approaching them, etc. – e.g., Ben-Youssef et al., 2017; Heenan et al., 2014; Holthaus & Wachsmuth, 2021; Kato et al., 2015; Lee, Kiesler, & Forlizzi, 2010; Pitsch et al., 2009). This includes any form of activity from the robot which may be witnessed by participants prior to their own interaction with it, similarly to human service-encounters where salespersons, doctors, help desk staff, sushi chefs, etc., are often already immersed in a (potentially competing) activity when they are sighted by the customer/patient/student (Harjunpää et al., 2018; Mortensen & Hazel, 2014; Robinson, 1998; Yamauchi & Hiramoto, 2016).

5.2.3.2 “Coming into sight” and “coming into existence”

The two types of HRI pre-beginnings described above make relevant an earlier distinction made by Gibson (1986) in his ecological psychology, regarding the way humans may appear on the social scene, and gradually achieve participant status in the pre-beginnings of encounters. In co-present encounters in relatively uncluttered spaces, co-participants usually get into a greeting position progressively, relying on the way they move, their gaze and gestures to continuously coordinate their getting-together, and make relevant interactional moves such as distant greetings (Kendon, 1990). Gibson calls this type of appearance a “coming into sight” (Gibson, 1986). This is the most common configuration in co-present encounters. He contrasts this with another type of appearance, in which the other person seems to materialize or come to life suddenly in the situation, as when someone hidden by features of the local environment abruptly becomes visible, which Gibson calls “coming into existence” (Gibson, 1986) to convey the “pop-up”, quasi-instantaneous character of this appearance. Other examples of “coming into existence” would be the initial connection in a video call (Licoppe, 2017), or in a co-present encounter, someone sleeping who suddenly wakes up after being approached. Because of its suddenness, the exact moment of “comings into existence” can be difficult to anticipate, the causes and underlying processes for such a “coming into existence” are not apparent, and finally, the interactional status and competence of the potential co-participant at the moment of their “coming into existence” can be uncertain.

5.2.3.3 “Off-the-record” pre-beginnings in HRI

Gibson's distinction may be highly applicable to HRI, for it now appears that, in the “pre-beginnings” phase of the first type of studies mentioned above, more or less prepared subjects have to deal with a robot that “comes into existence”. Conversely, in studies of the second type specified above, the robot's conduct (a priori) provides several conditions for a “coming into sight”: its progressive appearance or activation allows for some form of embodied mutual coordination in the “pre-beginnings” phase¹⁰³. However, in most cases, studies neither analyze nor clarify the state of the robot when participants see it (Arias et al., 2020) or enter in physical

¹⁰³ Of course, this etic distinction (produced from the perspective of the external observer) does not presuppose how participants *in situ* orient to these two types of robot appearances – this is precisely what the empirical section of this chapter will attempt to specify.

co-presence with it. Indeed, as “most experimental studies only start when the human is already placed in the appropriate starting position in front of the robot” (Holthaus et al., 2011), methodology sections rarely cover the observable behavior of the robot when participants encounter it. HRI experiments display an orientation to the “opening” phase of the interaction as the first relevant moment and tend to set aside the “pre-beginnings” phase, although, depending on the scenario, it may be crucial to the way subjects and robots achieve some form of co-participation status. These very first seconds, during which the robot appears to participants, are often, so to speak, “off-the-record” in the data that ends up being collected.

This chapter will examine a relatively common HRI experimental setup in which subjects are brought in the presence of a robot that suddenly animates – i.e., that markedly sets the stage for an interpretation of its conduct as a “coming into existence”. Thereby, it imitates the interaction opening delays regularly observed in laboratory or “in-the-wild” human-robot interaction studies, where robots can require time before springing to life after they stand in co-presence with a human. We will show how *this is consequential with respect to the way in which openings unfold, in which some moves such as greeting and waving may become interactionally relevant, and, ultimately, in which the robot emerges as a social agent*. We conclude that pre-beginnings are not anecdotal or peripheral issues with respect to HRI experiments but should be thought about and designed as an integral part of those.

5.3. Method

5.3.1. Participants

We base our analysis on 80 video recordings of dyadic interactions with an autonomous robot, which took place at the INSEAD-Sorbonne Université Behavioural Lab. Participants were all native French speakers aged between 18 and 30 years old. All participants were recruited by the INSEAD-Sorbonne University Behavioural Lab under ethics approval by the INSEAD Institutional Review Board. Separate consent was obtained for the use of video data. The experiment took on average 20 minutes to complete, each participant received a compensation of 6 €.

5.3.2. Experimental setup

A humanoid robot “Pepper”, produced by Aldebaran (formerly named Softbank Robotics), was positioned in the middle of a room, standing at a three-quarter angle from participants when they entered by the door (see Figure 5.2). The interaction was filmed with two cameras: one behind the robot, one on the left of the robot. An additional webcam was placed in a corner of the room. For a detailed description of our experimental setup and of the design of the autonomous robot, see Tatarian et al. (2021).

5.3.3. Instructions

Before entering the room, all participants were given the same verbal instructions:

1. “You are going to have an interaction with a social robot. This robot will try to help you plan your holidays, for this summer. Please answer as if you were really planning these holidays.”
2. “Speak loudly. If the robot does not respond, it is possible that it didn’t hear you. If you see a question mark displayed on the robot’s tablet, it means your utterance wasn’t understood: you can repeat or rephrase it.”
3. “The experiment should take 5 minutes to complete, then you will have to fill in a questionnaire in the next room.”

Then, as they entered the room, participants were informed that:

4. “The robot should start speaking to you in a few moments”
5. “You can stand anywhere in the room”

This characterization of the robot and of the task partially pre-configured the interaction. They created the expectation for an incoming, but delayed, first turn uttered by the robot. Doing so, they portrayed the robot as an entity which may not immediately be available as a co-interactant and potentially “come into existence” at some point. They also stated the robot may not hear the participant, depicting its perceptual abilities as imperfect. For these reasons, they should be treated as constitutive of pre-beginnings. In sum, the distribution of greetings (see Figure 5.1) on which we will focus is not to be understood as a direct reflection of the strength of a transsituational (greeting) norm but, rather, as connected to a specific experimental configuration.

5.3.4. Scenario

The robot was designed as a “travel agent”. Once participants had entered the room, the experiment followed a “holiday planning scenario”: the Pepper robot “woke up” by going through several “activation steps”, introduced itself, produced a “how are you” question, offered to take water, and, then, asked participants several questions aimed at understanding their preferred destinations. When the scenario reached its end, participants moved to a different room and completed a questionnaire composed of several psychometric scales.

All participants studied in this chapter faced the same initial behavior from the Pepper robot. Because our focus is on the earliest moments of the interaction, the different conditions in which these participants were placed did not impact the multimodal behavior of the robot yet. However, as part of a larger study (Tatarian et al., 2021), participants were distributed in 5 experimental conditions, each one featuring a different multimodal behavior from the robot later in the interaction (no social gaze, no approach, etc.). 101 valid participants took part in this experiment. In one of our experimental conditions, the robot did not wave during the opening of the interaction: these 21 participants were removed from our analysis (since they could not possibly react to the robot’s wave), leaving 80 remaining participants who all witnessed the same “activation steps” from the robot.

5.3.5. “Activation steps” achieved by the robot during the first seconds of the interaction

Immediately after each participant entered the room, the robot went through the same 5 “activation steps” (cf. Figure 5.1 for a detailed timeline):

1. **Physical co-presence:** When participants entered the room, the robot was motionless
2. **Gaze tracking:** The robot started to track their gaze¹⁰⁴
3. **Greeting:** The robot uttered a “bonjour” (“hello”)
4. **Wave:** The robot produced a waving gesture
5. **Self-identification:** The robot self-identified and introduced its role as a travel agent

These steps exacerbated two features of the “coming into existence” often observed in natural or controlled human-robot interactions openings: the robot stood in physical co-presence with the participant for several seconds before a reciprocal interaction could start and it displayed no preexisting activity when first appearing to this participant.

5.4. Distribution of first greeting occurrences during the experiment

Out of a total of 80 participants, 62 (78%) produced a greeting utterance (e.g., “Hi, Hello, Hey, Good morning” – Pillet-Shore, 2018) or gesture (e.g., “hand wave, palm display, head toss/bow, eyebrow flash” – Pillet-Shore, 2018). A micro-analysis reveals that most of these utterances and gestures corresponded to *greeting actions*, based on the definition from Pillet-Shore (2018): “discrete audible and visible (vocal, verbal/lexical, and embodied) actions that participants deploy to publicly mark the moment when they ratify another’s social copresence”. As the following fragments will illustrate, this does not imply the actions participants achieved through these utterances and gestures were limited to “greeting the robot”.

Among participants who produced a greeting utterance or gesture, the events or robot’s behaviors (“activation steps”) that immediately preceded the production of their *first greeting* were distributed as displayed in Table 5.1. The average delay¹⁰⁵ between these “activation steps” (cf. Figure 5.1) was long enough to prevent misattributing which “activation step” preceded participants’ greeting utterances or gestures.

¹⁰⁴ To do so the robot had to produce a vertical head tilt. This was connected with two “consequential sounds” (Tennent et al., 2017): the sound of motors and gears (required to tilt the robot’s head up) and the squeaking plastic.

¹⁰⁵ The presence of a measurable standard deviation with this delay between “activation steps” is due to variations in the robot’s CPU load between participants.

Table 5.1. Summary of First Greeting Occurrences

Activation step	First Greeting Occurrences	Description	Example
Physical co-presence	3 participants (5%)	Participants initiated a first greeting immediately after entering the room where the robot was placed	Fragment.5.1
Gaze tracking	2 participants (3%)	Participants' first greeting occurred after mutual gaze was established with the robot	Fragment.5.2
Greeting	31 participants (50%)	Participants' first greeting immediately followed the "hello" uttered by the robot	Fragment.5.3
Wave	24 participants (39%)	Participants produced a first greeting immediately after the wave achieved by the robot	Fragment.5.4
Post self-identification	2 participants (3%)	Participants' first greeting occurred at a later point during the interaction	
None	18 participants (22%)	Participants never produced any form of greeting	Fragment.5.5

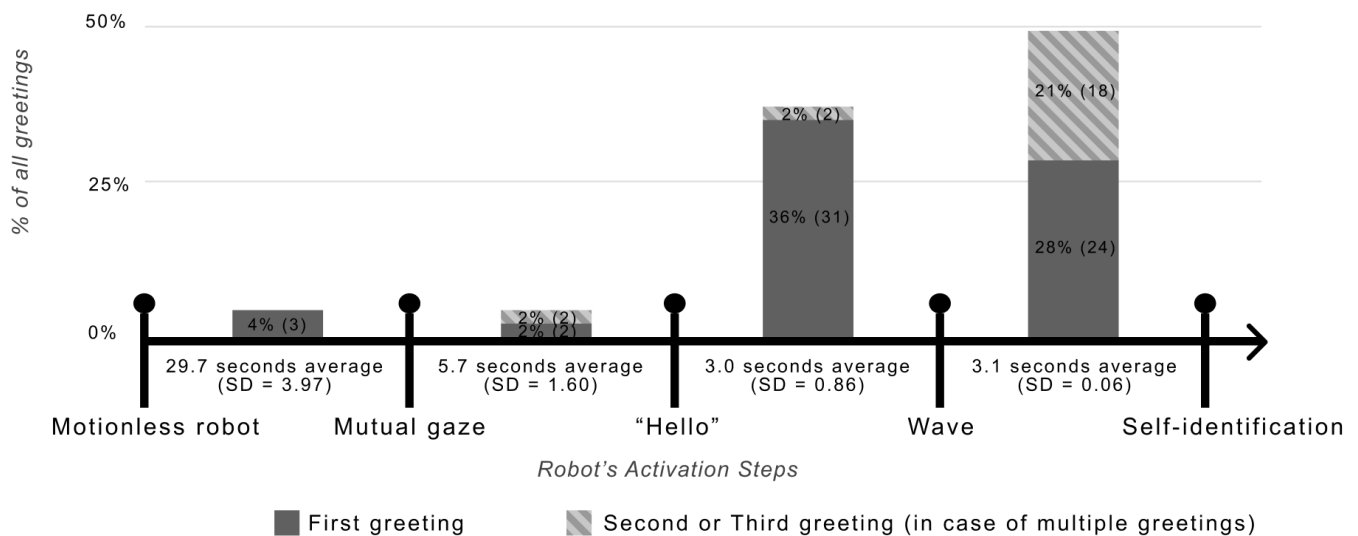


Figure 5.1. Distribution of all greetings produced by participants between "activation steps", including multiple greetings by the same participants. Total first greetings = 62, total greetings overall = 85.

Most *initial* greeting utterances or gestures therefore occurred after the robot's own verbal greeting. However, overall, most greeting behaviors were produced after the robot's wave, as many participants – among those who had previously greeted the robot – produced a *new greeting* after the robot achieved this gesture (cf. Figure 5.1).

5.5. Five paths to the production of a first greeting

The five fragments below, ordered chronologically relative to the robot's "activation steps", are representative of this corpus¹⁰⁶. They are analyzed using an ethnomethodological and conversation analytic methodology in order to reveal, on a moment-to-moment basis, *what interactional processes are aggregated in the statistical distribution presented above*. That is, which events or actions were typically oriented to by the human participant as constructing an appropriate framework for the production of greetings. Our transcription conventions are detailed at the beginning of this work (page xiv).

5.5.1. Orienting to physical co-presence as an adequate framework for the initiation of a greeting sequence

5.5.1.1 Fragment 5.1

1. ***** (2.5) ***£** (3.0) ***** (0.1)
hum *>>places coat on chair*moves in front of rob*
hum £gazes at rob->
2. **HUM** **hello peppeur?**
hello Pepper
3. (3.6) £# (3.8) £ (2.4) £ (1.3) £Δ (1.2)
hum £tilts head£ £gazes at camera£gazes at rob->
hum Δsmiles->>
img #img.1.1

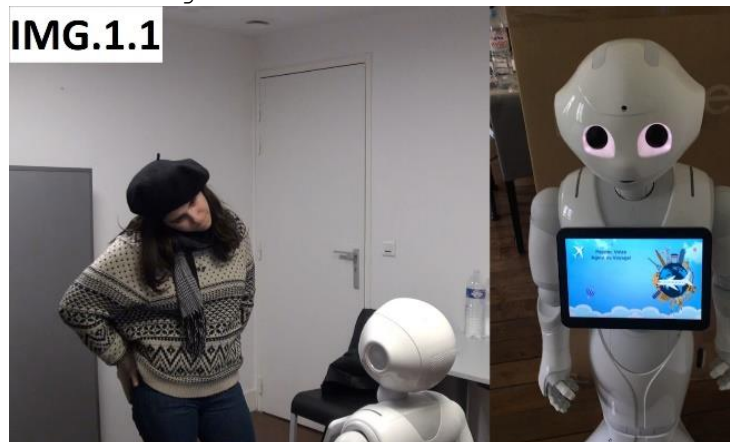


Figure 5.2. Image 1.1 – Participant tilts her head after the robot does not reply to her greeting

4. **ROB** ((motor and plastic sounds))+**\$**
rob +gazes at human->>
rob \$arms are shaking->>

¹⁰⁶ Each fragment displays the most common way in which a greeting emerged for each "activation step" – and a typical case in which greetings did not emerge for fragment 5.5. However, very few greetings occurred after the first activation step (Motionless robot) and the second activation step (Mutual gaze) – see Figure 5.1. In this sense, fragments 5.1 and 5.2 display rare occurrences in comparison with all 80 participants analyzed. Note that, even though these fragments were chosen as the most representative, some specificities of these participants' behavior (i.e., the distance at which they were standing from the robot) are still idiosyncratic and different from the average.

5. (3.5)Δ (0.5) Δ(1.0)
hum Δraises one eyebrowΔ

6. HUM hello peppeur.
hello Pepper

7. ROB =bon\$jour
hello
rob \$opens its arms->

8. (1.1)\$ (0.8)\$ (0.3)* (0.3)* (0.6)\$ (0.9)* (0.5)
rob ->\$.....\$-----waves-----\$,,,->
hum *.....*---waves---*,,,->

9. ROB je \$* m'appelle peppeur
my name is Pepper
rob ->\$
hum ->*

5.5.1.2 Description and analysis

After laying her coat on the chair, the participant gazes at the robot and positions herself in front of it, facing it (L.1). Standing at 0.8 meters from the robot, she is comparatively close to it in regard to other participants (average initial distance = 1.3 meters; SD = 0.26). She immediately initiates a first greeting pair and an address term (“Pepper”, L.2) and maintains her gaze, body orientation and posture – slightly leaning towards the robot – during the next 3.6 seconds. After this standstill, she produces a lateral head tilt for 3.8 seconds (L.3), while maintaining her gaze towards the robot’s eyes. A few seconds later, the robot directs its face towards the human’s eyes, establishing mutual gaze. This quick adjustment of the robot’s head is associated with motor noise and plastic sounds (L.4). These so-called “consequential sounds” are known to be regularly oriented to by participants in human-robot interactions (Frid et al., 2018; Schulz et al., 2021; Tennent et al., 2017). The robots’ arms also start shaking lightly (L.4), which will persist until the end of the experiment. After 3.5 seconds of this mutual gaze, the participant raises one eyebrow (L.5) and, after a second of silence, produces a new greeting (“hello Pepper”, L.6). This action appears to account for the robot’s continued silence and, in particular, to orient to the robot’s alignment with the human’s gaze as creating expectations for a next action on its part. However, the participant’s greeting is immediately followed by a greeting from the robot (L.7). This implies that, in this fragment¹⁰⁷, the robot’s greeting turn is sequentially positioned as a second pair part *responding* to the participant’s greeting – who, indeed, does not achieve a new greeting in return (L.8), suggesting that she orients to the robot’s “*bonjour*” as a reaction to her own. After a short pause, the robot then starts to produce a greeting wave (L.8), with which the participant aligns by producing a similar wave. Once the robot starts to retract this waving gesture, the participant immediately begins to lower her own arm and finishes her retraction simultaneously with the robot.

The previous fragment constitutes one of only three openings in our corpus where a greeting is initiated by the participant before the robot was activated (from an etic perspective). *All along the interaction, the robot is constructed as a co-present interlocutor, even when it is not moving yet.* Besides the “*bonjour*” that the robot produces after her greeting, the participant

¹⁰⁷ That is, independently from the intended design for the script followed by the robot at the beginning of the interaction. Indeed, this first greeting was originally intended as a greeting initiation, not as a response. There is obviously no systematic overlap between the features of the local situation which are practically oriented to by participants as relevant in an interaction with a humanoid robot, and those which were designed as *a priori* relevant during the design of the autonomous robot’s behavior (Pelikan et al., 2020).

orients to every “activation step” displayed by the robot (physical co-presence, mutual gaze and wave) as opportunities to produce a first greeting. Thus, as soon as she is positioned in front of the robot, she treats the situation as a “framework in which a greeting sequence is relevant” (Mortensen & Hazel, 2014) and the robot as able to react to the production of a greeting: by tilting her head¹⁰⁸ (L.3) and by raising one eyebrow (L.5), she makes accountable the non-answer of the robot to the first pair parts she produced and she displays expectations for a reaction.

Remarkably even though the participant’s last verbal greeting (which took place immediately before the robot’s vocal greeting) ultimately positioned her as having “initiated” the greeting sequence (L.6 & 7), she instantly re-positions the robot as the “anchor” (Schegloff, 2010b) or, more generally, as the speaker initiating new sequences. Indeed, after the robot greets her back, the participant stays silent (L.8) and does not use her position as a first speaker to self-select (Sacks et al., 1974) and to initiate a new sequence. This silence by the participant leads the robot to produce an “interlocked” turn (Schegloff, 2007) which combines both its response to her “hello” and its initiation of a greeting wave (L.8). By staying silent, the participant thus provides the robot with the adequate position to initiate subsequent sequences (Schegloff, 2007) and, later, to initiate the topic of the interaction (not in transcript).

5.5.2. Orienting to mutual gaze as projecting an imminent next action from the robot

5.5.2.1 Fragment 5.2

1. * (1.7) £* (2.0) £*# (0.7) £
 hum *>>closes bag*approaches robot*adjusts clothes->
 hum £ gazes robot £gazes clothes£gazes rob->
 img #img.2.1

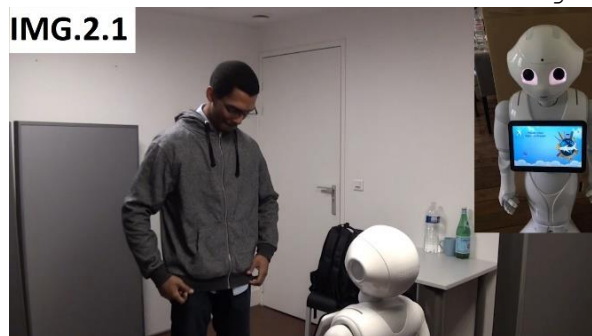


Figure 5.3. Image 2.1 – Participant adjusts his clothes after positioning himself in front of the robot

2. **ROB** ((motor and plastic sounds))+\$
 rob +gazes at human>>
 rob \$arms are shaking->>
 3. (0.1) * (0.6)
 hum -> *
 4. **HUM** .tsk .h

¹⁰⁸ The lateral head-tilt she produces (L.3) may be understood as an embodied account of the non-response of the robot, in the same way that such head-tilts can be used as embodied displays of trouble in classroom interactions (Aldrup, 2019; Seo & Koshik, 2010).

5. £ (0.5)
 hum ->£gazes at own sleeve->
 6. HUM °hef#llow::°*
 hello
 hum ->£gazes at robot->
 hum *takes a step forward->
 img #img.2.2
 7. (1.0)*(2.1)* (0.8)# *(2.8)
 hum ->* *takes a step backward*
 img #img2.3

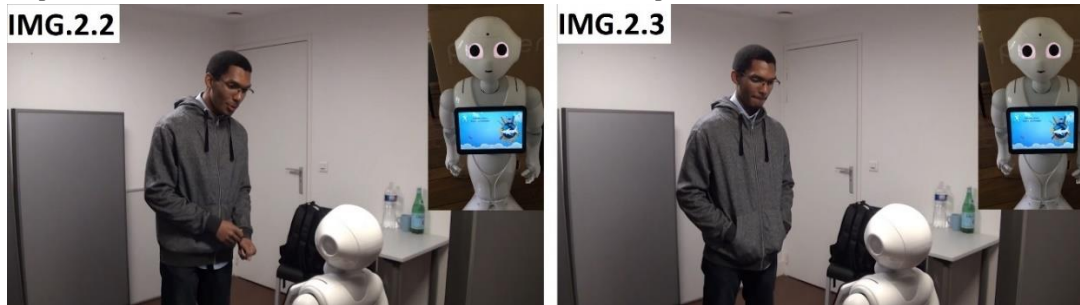


Figure 5.4. Images 2.2 and 2.3 – Participant steps towards the robot after mutual gaze is established, then, after a few seconds of silence, steps back

8. ROB \$*bonjour
 hello
 rob \$opens its arms->
 hum *takes hand out of pocket->
 9. (0.6)* (0.3) * (0.2) *(0.2)\$ (0.2)
 hum ->*slightly extends arm*retracts arm*
 rob ->\$waves->
 10. HUM bonjour#
 hello
 img #img.2.4
 11. £ (0.4)*# (0.5) #
 hum ->£gazes at wave->
 hum *extends hand towards robot->
 img #img.2.5 #img.2.6

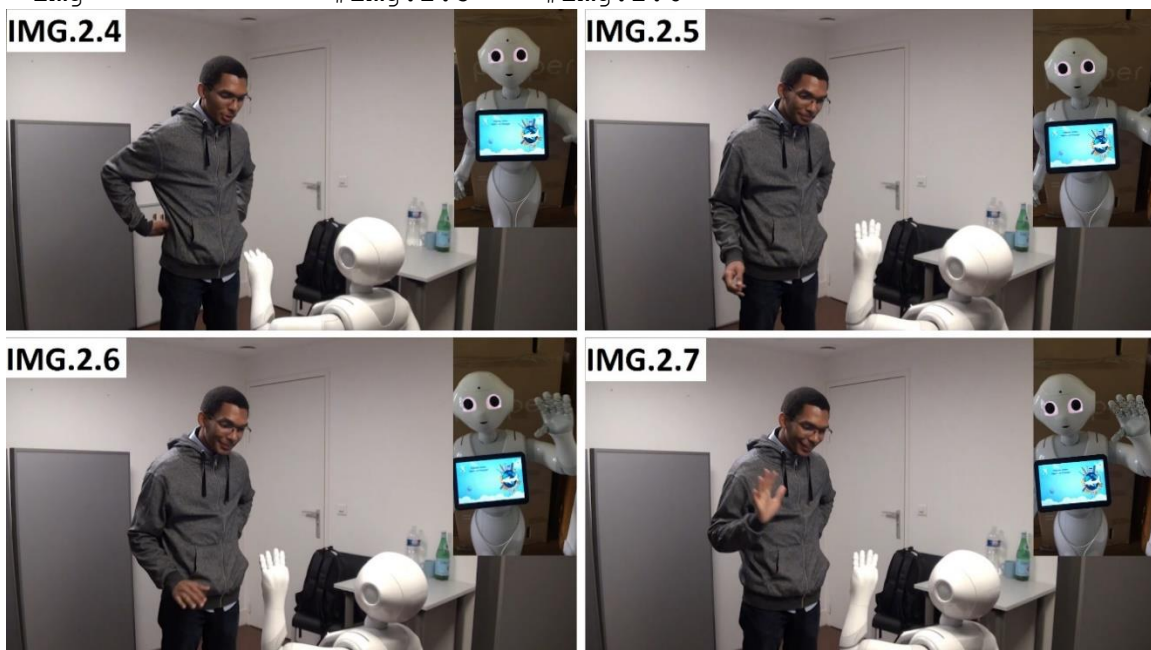


Figure 5.5. Images 2.4 to 2.7 – Participant starts to extend his hand towards the robot, then modifies the trajectory of his hand to align with the robot’s waving gesture

```

12.HUM      .tsk £*Δ (.)#hello
              hello
hum         ->f gazes at robot's face>>
hum         ->*waves->
hum         Δsmiles->>
img         #img.2.7
13.         (0.5)*(0.8)
hum         ->*
14.ROB      je m'appelle peppour
              my name is Pepper
rob         ->$

```

5.5.2.2 Description and analysis

After positioning himself in front of the robot, the participant does some self-grooming (Kendon, 1990) as he readjusts his clothes (L.1). Standing at 0.8 meters from the robot, he stands closer to it than most participants. After the robot orients its face towards his eyes (L.2), with the matching motor noise and squeaking plastic sounds, the participant gazes back at it and produces a tongue smack (“tsk”, L.4) followed by a first greeting (“*hello*”, L.6), interrupting his self-grooming (L.5). He then takes a step forward while staring at the robot and, after 2.1 seconds of silence, takes a step back to his original position (L.7). After 2.8 additional seconds of silent mutual gaze, the robot utters a greeting term (“*bonjour*”, L.8) as the participant just finished taking his arm outside of his pocket. The participant briefly extends his hand towards the robot (L.9), then retracts it and produces a return greeting (L.10). The robot starts to raise its arm to prepare a waving gesture (L.10, image 2.6). When this gesture has not reached its apex yet, the participant starts to extend his own hand towards the robot (L.11, images 2.7 & 2.8); however, once the robot’s arm becomes fully extended and starts the waving motion, the participant redirects his arm to produce a waving gesture (L.11, image 2.9). He simultaneously achieves a new verbal greeting and smiles (L.12).

In this fragment, physical co-presence is not treated as sufficient to initiate the interaction – unlike the previous example. The robot is not oriented to as a conversational partner from the very start. This is especially visible through the production of “self-grooming” by the participant, usually displayed during the approach between two interactants (Heenan et al., 2014). However, the status of the robot in the interaction shifts after the establishment of mutual gaze. The participant’s interruption of his self-grooming (L.5), his greeting (L.6), and the step forward he takes (L.6) accentuate a shared inner space (Kendon, 1990) and display the expectation of an imminent action from the robot. This reconfiguration results from the “crucial analytic distinction” (Mortensen & Hazel, 2014) made by the participant about what the gaze from the robot is projecting: it is not oriented to as a mere automatic “gaze tracking”, nor as a “mere look” (Kidwell, 2005; Mortensen & Hazel, 2014) but as a look projecting the initiation of an upcoming action. The participant’s expectation is not met, however, as he goes back to his original spot. Mutual gaze therefore constitutes the first “breaking point” after which the robot becomes (momentarily) present as a potential interlocutor. As Pitsch & Koch (2010) notes in the case of young children interacting with a toy robot, even “little sequential phenomena of the robot’s timely conduct” in relation to participants’ actions can have a critical impact on the “categorization and re-interpretation” (Pitsch & Koch, 2010) of this robot. In our

fragment, the redirection of the robot’s gaze towards the participant after a silence is treated as a *meaningful social action*.

Incidentally, two reconfigurations could be observed in this participant’s gestures. First, L.9, he extends his hand towards the robot immediately after its first verbal greeting, before retracting this hand and producing a return “*bonjour*”. The cancellation of his tentative gesture and, instead, his production of a verbal greeting, appear to constitute alignments with the robot’s (then verbal) mode of greeting. Later, as the robot starts to visibly raise its arm as part of its waving gesture, the participant’s response gesture shifts from an apparent “handshake” gesture to a clearly observable wave (L.11 to L.12; images 2.4 to 2.7). These two episodes display quickly evolving interpretations of what action the robot is projecting during its first greeting and, later, during the preparation of its waving gesture. It highlights an online monitoring of the robot by the participant (Pitsch et al., 2013), which allows him to reconfigure his embodied course of action to align with the robots’ co-occurring action¹⁰⁹.

5.5.3. Multiple greetings – Orienting to the robot’s wave as the confirmation of an ongoing greeting sequence

5.5.3.1 Fragment 5.3

1. **#£*(4.1)**
 hum £>>looks around the room->
 hum *>>swings from one leg to the other->
 img #img.3.1



Figure 5.6. Images 3.1 and 3.2 – The robot gazes at the participant, she stops her swinging movement and establishes mutual gaze

2. **ROB ((motor and plastic sounds))+\$**
 rob ->+gazes at human->>
 rob \$arms shaking->>
3. **(0.1)£Δ (1.5)*#(2.4) Δ**
 hum ->£gazes at robot->
 hum Δraises eyebrowsΔ
 hum ->*

 img #img.3.2
4. **ROB bonjour**
hello

¹⁰⁹ Of course, besides tracking the participant’s gaze, our autonomous robot’s behavior could not itself be reconfigured by co-occurring actions from the participant: from an etic point of view, there was no “loop of mutual adjustments” (Pitsch et al., 2014) to speak of.

```

5.          (0.1)$ (3.5)
   rob      $opens its arms->
6. HUM     °bonjour°?$(.)
           hello?
   rob      ->$waves->
7.          (0.4)£      (1.0)      £(0.4)
   hum      ->£gazes at wave£gazes at robot's face->>
8. HUM     #bonjour,Δ# (.)
           hello
   hum      Δwidens her smile->>
   img      #img.3.3 #img.3.4

```

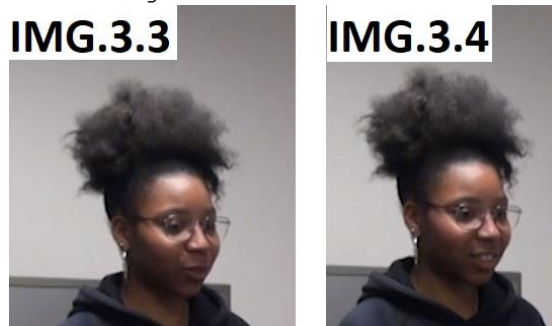


Figure 5.7. Images 3.3 and 3.4 – Participant’s widened smile when uttering her second “hello”

```

9. ROB     je m’appelle$ peppeur
           my name is Pepper
   rob      ->$

```

5.5.3.2 Description and analysis

After entering the room, the participant positions herself in front of the robot, at a higher distance than most participants. She starts to swing from one leg to the other while looking around the room (L.1). After a few seconds, the robot gazes at the human’s face and, doing so, produces motor noise and squeaking plastic sounds (L.2). The participant instantly stops her swinging movement and initiates mutual gaze with the robot, while raising an eyebrow (L.3). She maintains this posture for the next 4 seconds of mutual silence, and even after the robot utters a first greeting (“bonjour”, L.4). Because of a momentary processor overload, the robot’s waving gesture takes 3.7 seconds to be triggered after its “bonjour”, unlike the rest of the corpus where it took on average 3 seconds. After 3.5 seconds of silence, possibly orienting to the silence and the stillness of the robot as offering a response slot, the participant softly whispers a first greeting (L.6) with a rising vocal pitch – right before the robot starts its wave. Once the robot initiates its waving gesture (L.6), the participant glances towards its waving arm (L.7) then utters a new greeting (“bonjour”, L.8). This new greeting is uttered out loud and with a final continuing intonation, while the participant widens her smile (L.8, image 3.4).

In this fragment, the participant orients to the first greeting of the robot as sequentially equivocal (Hopper, 2005). Her first greeting (whispered, delayed, with a rising pitch) displays uncertainty regarding what the robot’s greeting is projecting: the robot’s “bonjour” is not oriented to by the participant as clearly constituting the first part of an adjacency pair which should be completed by a return greeting. The design of her first greeting thus appears to question the status of the interaction – and even the existence of a “stepwise process of mutual

adjustments” (Pitsch et al., 2009). Conversely, the participant’s second greeting turn appears to confirm “what is going on” (Beach & Sigman, 1995) as an “exchange of mutual greetings”: uttered out loud and simultaneous with a widened smile, it orients positively¹¹⁰ to the robot’s gesture as initiating a second greeting sequence.

This supports an interpretation where each greeting produced by the participant accomplishes a different task (Mondada, 2015). The first greeting, uttered 3.5 seconds after the robot’s own greeting, mainly checks the availability of the robot and its ability to perceive and react to the human’s relevant actions¹¹¹ – a form of “device testing” (Relieu, 2007) – whereas the second greeting is a clear ratification of the start of the co-present interaction: it is a “sociability practice” (Mondada, 2015). Consequently, we observe a form of *inertia* in this fragment: the inanimate object that the robot is first oriented to requires interactional work (lasting over several seconds) to be replaced by a conversational agent. The first greeting term produced by the robot does not immediately nor automatically institute it as a conversational partner which can be greeted back.

5.5.4. Orienting to the robot’s waving gesture as an upgrade of its first greeting

5.5.4.1 Fragment 5.4

1.	£	(11.0)	£(0.2)
hum	>>	£looks around the room	£gazes at robot->>
2. ROB	((motor and plastic sounds))+£		
rob			+gazes at human->>
rob			\$arms are shaking->>
3.	(6.2)		
4. ROB	\$	bonjour	
		hello	
rob	\$	opens its arms->	

¹¹⁰ Smiling, which is “a principal way parties do ‘displaying a positive stance’ toward encountering recipients” (D. Pillet-Shore, 2018) suggests the situation is now being treated as the beginning of a socially co-present encounter.

¹¹¹ This participant can be described as verifying whether the entity in front of her is “an animate object that is able to engage with her in re-occurring interactional patterns” (Pitsch & Koch, 2010).

with the robot's wave, then stops immediately after the robot starts retracting it (L.7): 200 milliseconds after the robot's arm starts to lower, the participant also starts to retract her arm – as in fragment 5.1. Like the overwhelming majority of our corpus, this participant's gaze focuses on the robot as soon as it moves its head to track her gaze – but she does not immediately produce a speaking turn. The lasting silence, mutual gaze and “consequential sounds” are not oriented to as initiating a “slot” where to self-select. Even after the robot utters a “bonjour”, she returns no greeting and maintains her previous pose and gaze for the next few seconds.

However, once the robot starts a waving gesture, the participant silently observes its arm rise during the action's preparation. She then abruptly produces her own wave – which catches up with the robot's gesture – and simultaneously produces a smile and a verbal greeting (L.6). The speed of this return wave seems to indicate that the participant orients to the robot's gesture as, either, producing a normative obligation to achieve a return greeting, or, alternatively, as upgrading a normative obligation to respond that she would have previously failed to observe. In particular, based on the numerous occurrences of this situation in our corpus, we suggest that, in this fragment, the participant's hasty first greeting displays her alignment as normatively expected at an earlier point in the interaction. That is, she orients to the robot's wave as a second greeting sequence which reinforces the conditional relevance attached to its first vocal first greeting (“*bonjour*”), to which she did not answer. The robot's wave is treated as making accountable the participant's non-answer to this first greeting sequence¹¹².

5.5.5. Absence of greetings – Orienting to the robot as an autonomous, machine-based, script

5.5.5.1 Fragment 5.5

1.		+*	(4.4)	*£	(1.7)	£	(0.2)	£
	rob	+>>looks straight ahead->						
	hum	*>>drops bag*						
	hum	£gazes at rob£gazes at camera£gazes rob->						
2.	HUM	*hh trop bizarre# struc hh						
		<i>this thing is really weird</i>						
	hum	->*approaches robot->						
	img	#img.5.1						

¹¹² The robot's waving gesture may have retrospectively positioned the participant in a situation of “ritual disequilibrium” (Goffman, 1955) for not having produced a return greeting sooner. However, this Goffmanian interpretation goes beyond what's observable in this fragment.

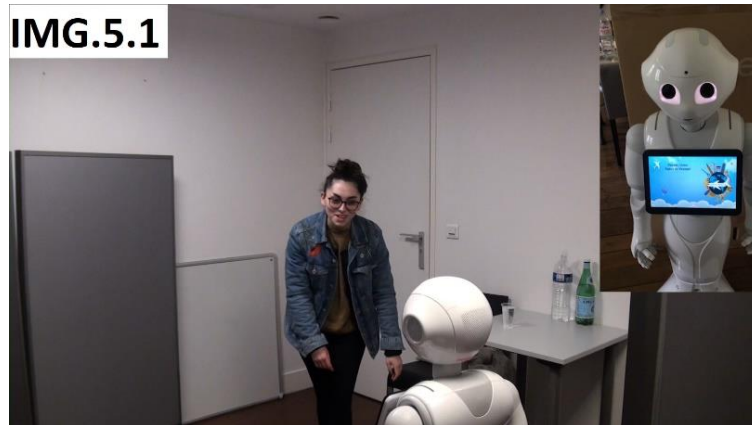


Figure 5.9. Image 5.1 – Participant approaches the robot while producing “self-talk”

- 3. (1.1) * (2.5)
hum ->*
- 4. **ROB** ((motor and plastic sounds)) + \$
rob ->+gazes at human->>
rob \$arms are shaking->>
- 5. Δ(0.5)
hum Δsmiles->
- 6. **HUM** hhh#
img #img.5.2

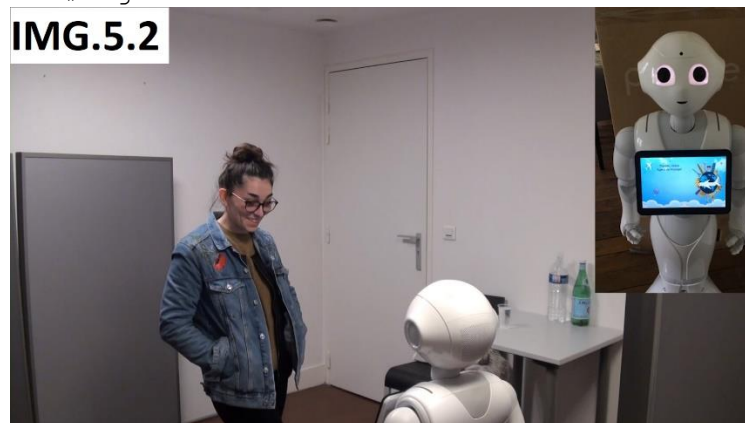


Figure 5.10. Image 5.2 – Participant laughs after the robot achieves mutual gaze

- 7. (3.9) £Δ (1.0) £ (0.1)
hum ->£gazes above robot£gazes at robot->
hum ->Δ
- 8. **ROB** bon\$jour
hello
rob \$opens its arms>
- 9. (0.6)
- 10. **HUM** Δ(.tsk) (0.4) \$ (0.2) \$ (0.1) hhheheheh?# (1.0) \$ (0.2) .h
hum Δsmiles->>
rob ->\$.....\$-----waves-----\$,,->
img #img.5.3



Figure 5.11. Image 5.3 – Participant’s second laugh during the waving gesture of the robot

11. (0.1)£(0.2)
 hum ->£gazes at wave->
 12.ROB je m’ appelle\$£ peppour
 my name is Pepper
 rob ->\$
 hum ->£

5.5.5.2 Description and analysis

After entering the room and dropping her bag on the ground, the participant gazes at the robot (L.1). As she starts approaching it, she produces a speaking turn involving a deictic reference to the robot as “this thing” (“*ce truc*”, L.2) – referring to it in the third person – and qualifies it as “weird”. This comment displays most of the typical properties of “self-talk” (Keevallik, 2018): it is achieved while the participant is leaning forward, her body not oriented towards the robot, and part of it is uttered while looking at the ground, in a low voice. As a consequence, the participant does not manifest any expectation for an answer; her comment does not open a “conversational sequence” (Keevallik, 2018). Once her approach is complete, the participant stands 1 meter from the robot, closer than the average initial distance of 1.3 meters for all participants. After a silence of 3.6 seconds, the robot shifts its gaze towards her eyes – doing so, it produces soft motor noises and squeaking plastic sounds (L.4). The participant reacts with a short laugh (L.6). After a new silence of 4 seconds, the robot initiates a first greeting (L.8) and opens its arms – palms towards the ceiling. The participant produces another laugh, more audible and longer than her previous one (L.10). This laugh is continued during most of the hand wave of the robot and is followed by an in-breath when the robot’s arm starts to retract (L.10).

The actions of this participant highlight a double orientation to the robot, as both normatively neutral and unable to react to human actions. First of all, she treats the robot as an autonomous script whose verbal utterances imply no normative obligation to be responded to, even after it produces a greeting term: no conditional relevance emerges from the behaviors produced by the robot in the course of the interaction. Yet, simultaneously, the participant orients to the robot as unable to respond to (or to perceive) her own actions. The “self-talk” (L.2) or laughs (L.6 & 10) she produces in front of the robot do not manifest any expectation for an answer, and the absence of reaction from the robot is not visibly made accountable (unlike fragment 5.1). Her turns are not “recipient designed” to be registered as

“inputs” in response to which the robot would produce, reconfigure (Goodwin, 1981) or interrupt speaking turns. In other words, the robot is never “characterized as being able to perform reciprocal interactions” (Straub, 2016), which is a prerequisite for the existence of a “social encounter” (Straub, 2016). Neither the robot nor the human imposes a normative order on the other: there is no observable “sequence organization” (Schegloff, 2007) which exerts a constraint on their actions. Using the previously mentioned terminology from Jarske et al. (2020), this participant is merely “using speech” but not “talking” with the robot, in the sense where “talking” would imply to produce “actions that are discoverable within a normative order” (Jarske et al., 2020) and to assume “other participants to be able to perceive them as actions within that order” (Jarske et al., 2020).

Additionally, no facework is achieved by the participant as, in particular, her comments are not fully whispered (L.2) and her laughs are achieved audibly and visibly while standing right in front of the robot. Her first utterance (“this thing is really weird”) also explicitly characterizes the robot as an object and refers to it in the third person. As a whole, semantic content and sequential organization mutually reinforce to characterize and position the robot as a non-agent: the participant establishes herself as the spectator of a pre-recorded monologue, whose performer is not socially present with her in the room.

Last, note that even though this participant approached the robot more than average (i.e., stood at less than 1.3 meters) this unusual proximity was part of a sequence where she scrutinized and commented on the robot, treating it as an inanimate object instead of an interactant. This is unlike participants in fragment 5.1 and 5.2 – even though they also stood unusually close to the robot – for whom an unusual proximity displayed a treatment of the robot as a current (fragment 5.1) or imminent (fragment 5.2) interactant. This crucial difference connects with the general observation that, on an individual level, the distance at which a participant stood from the robot was meaningful only in connection with a sequential context.

5.6. Discussion

5.6.1. Sequential ambiguity

The previous fragments reveal the varied interactional work required before “behaviors” from the robot could be treated as “actions” which either 1) established the adequate framework for the participant to initiate a first greeting sequence or 2) produced a response slot that the participant was normatively pressured to complete with a return greeting. Even though, after they entered the room, all participants positioned themselves in front of the robot to form a vis-a-vis arrangement (Huettenrauch et al., 2006; Kendon, 1990), we see that the mere utterance of a greeting term from the robot (“hello”) did not automatically and immediately establish a reciprocal interaction. There was a regular delay in the shift from the robot as a normatively neutral artifact (which was discovered completely motionless at the very beginning of the interaction) to a conversational partner producing sequentially implicative turns. For many participants (exemplified in fragments 5.3, 5.4 and 5.5), the interactional status of the robot *persisted* after it produced a first greeting.

An explanation for this delayed emergence of the robot as an agent is that participants found themselves confronted (and prepared to be confronted by the instructions described in section 5.3.3) to an inert robot that suddenly animates and “comes into existence” and were therefore engaged in sequentially ambiguous situations (Hopper, 2005) as to what actions the robot was projecting (or if it was projecting anything): they had to “entertain the full range of possibilities momentarily, using the immediately following talk to find out what sort of sequence is in progress” (Schegloff, 1980). During fragments 5.1, 5.2 and 5.3, participants’ initial turns can be considered as practical attempts to “probe” the current status of the interaction: by trying to elicit a response from the robot, these actions clarified whether a phase of mutual adjustments, or any form of turn-based coordinated activity, was either ongoing or technically feasible.

We argue that these face-to-face encounters with a humanoid robot “coming into existence” disrupted “background expectancies and methods at play in the accomplishment of commonplace activities, such as having a conversation” (Stanley et al., 2020). Participants had to achieve greetings in a situation marked by otherness, which, like it was observed in different contexts, “throw[s] the greeters and their practices of greeting into crisis” (Mondada et al., 2020). This experiment thus made especially apparent the constant “experiments in miniature” (Garfinkel, 2006; vom Lehn, 2019) achieved by humans when interacting with robots (and, of course, with other humans), where each action “tests the hypothesis a participant has about a co-participant’s response to her/his action”. In particular, in these human-robot interactions, participants faced the challenge of 1) identifying if intersubjectivity (vom Lehn, 2019) was even possible (i.e., if the entity in front of them possessed the required properties for mutually achieving a “reciprocity of perspectives” – vom Lehn, 2019) and, then, of 2) establishing this intersubjectivity – for example, by producing actions which displayed, and therefore tested, an orientation to the previous robot’s turns as opening a greeting sequence. We suggest that this double challenge is a common trait of first encounters with humanoid robots, which may require the use of different resources to be overcome depending on the way in which the robot is first encountered by the human.

5.6.2. The waving gesture as a threshold

The robot’s wave was often critical in clearing up this “sequential ambiguity” (Hopper, 2005). In several cases (although, not systematically) this gesture offered a practical answer to the practical issue participants were encountering, namely to document “what is going on” within a given spate of talk (Beach & Sigman, 1995). For example, in the specific sequential contexts presented in fragments 5.3 and 5.4 (i.e., not a wave “as such”, discretized and disconnected from local situations¹¹³), the wave functioned both as a clarification of the situation – as an ongoing greeting sequence – and, simultaneously, as a soft upgrade of the conditional relevance of the previous greeting turn produced by the robot (“hello”). In other terms, it manifested the normative obligation to produce a return greeting and retrospectively oriented to the participant’s non-response (or non-proper response) to the robot’s first vocal greeting.

¹¹³ “Actions are intrinsically meaningful because they unavoidably participate in an organization of activity, not because there is an abstract, decontexted meaning which they have independent of their occurrence. Action is intrinsically meaningful, not because it is meaningful outside of any concrete situation, but because it is always embedded in a concrete situation.” (Peyrot, 1982b)

Therefore, akin to the responses observed by Pelikan et al. (2020) after a two-part animation¹¹⁴ of the robot Cozmo, the – etically designed – “two part greeting” achieved by our robot (vocal greeting, pause, wave) often led participants to reconsider their past actions. To paraphrase Enfield & Sidnell (2017), there was an observable evolution in the “multiple drafts” these participants produced of the situation, as the robot achieved a waving gesture.

In sum, more than any other of its behaviors, the robot’s wave was frequently responded to as a conditionally relevant first greeting pair. The instants following this wave constituted a frequent (and momentary) threshold between the robot oriented to as a “raw physical artifact” (H. Clark & Fischer, 2022) of plastic, sensors, etc., and the robot treated as a socially co-present entity – whose actions could establish “a set of normative constraints on the type and form of action with which the recipient should respond” (Kendrick et al., 2020). The wave regularly interrupted the persistence of the status of the robot as a non-agent: in these cases, the “self-affirming done through language” (Duranti, 2007) emerged as a consequence of this gesture¹¹⁵.

5.6.3. Should a robot be designed to harness conditional relevance?

Antithetical design opportunities stem from the observation that behaviors from a robot can establish a normative pressure to produce an adequate social response (here, a return greeting action). Roboticians may use such behaviors – documented to produce alignment from humans – to enforce a robot as a “social agent” at the beginning of an interaction, or, conversely, design the robot to align with the way in which participants treat it from the very start. As Jarske et al. (2020) mention, “[b]y mimicking features that are familiar to people from their daily interactions, the robot is appealing to our sociality; not just cognition, but *social norms that constitute our social world in interactions*” (emphasis ours). For example, one could imagine designing a robot which only produces a “reinforcement wave” when its verbal greeting is not answered with a greeting action after several seconds, to purposely pressure a human interlocutor into greeting it like a legitimate social agent. This raises the question of whether designers should leverage conditional relevance as a tool – i.e., harness the tacit normative order (Lempert, 2013) of human-human sequence organization (Schegloff, 2007) – to induce social treatments of the robot by participants; no matter how the robot is perceived by these participants. And, if so, to which extent?

Indeed, existing ethical and usability debates (Fronemann et al., 2022) can be connected to the legitimacy and capability of a robot to impact the degree to which it is being responded to as a social agent. As Fischer (2021) notes, once a robot can identify that some of its actions (even a limited set of ritualized actions like greetings or goodbyes – Stivers & Rossano, 2010) are not being treated as those of a social agent, this opens up new possibilities for the field of personalization in robotics (Dautenhahn, 2004). Information about the ongoing interactional status of a robot enables triggering different behaviors from the robot

¹¹⁴ Focusing on the Cozmo robot’s “sad” animation, designed to unfold in two parts, Pelikan et al. (2020) observed that the second part of this animation could be treated by participants as the upgrade of an action projected by the first part, leading them to reconsider their prior actions.

¹¹⁵ This is not to say that, from this moment on, the robot was consistently and exclusively treated as a social agent during the rest of the interaction. On the contrary, “quick changes of perspective” (K. Fischer, 2021) have been observed in the responses of participants to a robot. Alač (2016) also notes that each facet of the robot “can itself take center stage as the encounter develops”.

when humans appear to be construing it exclusively as a “raw physical artifact” (H. Clark & Fischer, 2022). In particular, information about the current status of a robot in an interaction offers the possibility for this robot to adapt to its interlocutor’s observable initial definition of the situation (for example by aligning with a treatment of itself as a simple device by an “utilitarian” (Lee, Kiesler, Forlizzi, et al., 2010) or “non-player” (H. Clark & Fischer, 2022) interactant who only uses keywords and does not greet the robot), or, on the contrary, to rely on different strategies to change the (non-social) way in which its interlocutor is treating it.

The phenomenon of conditional relevance therefore adds another dimension to the question of the degree of agency a robot should display – of which some participants’ interpretation of our robot’s wave as a “reinforcement of the normative pressure of the robot’s first greeting” is a striking example – and highlights how *even very mundane and minute choices from designers have ethical ramifications*. Human-human sequence organization in conversation has been argued to be the place for a form of “proto-morality” (Bergmann, 1998; Robles, 2015): the treatment of a robot as involved (or not) in a sequence organization is directly connected with the (non-)attribution of specific rights and responsibilities to this robot in a “micro-level moral order” (Stivers et al., 2011). Consequently, any attempt to “enforce” this treatment has unavoidable normative implications.

5.7. Entry into physical co-presence as a blind spot in HRI

Based on the video data from 80 interactions, we observed that the sudden activation of our robot (its “coming into existence”) was pragmatically consequential for participants. The intertwining between participants’ actions and the “activation steps” displayed by the robot (including its original motionlessness, which was sometimes oriented to as meaningful by participants – see also Schulz et al., 2021) led to the emergence of various sequential trajectories: some participants ended up orienting to the robot’s gaze shift, to its wave or to its greeting as a *response* to a greeting they just produced, others ended up orienting to these behaviors as *initiating* a greeting sequence, as *reinforcing* a previous greeting, and more generally as a *slot for a next action*. Nevertheless, coming back to our original question of “how” changes first emerge in the status of a robot during an interaction, it is not possible for us to suggest to which degree some patterns we identified in section 5.6 (e.g., participants’ treatment of the robot’s delayed waving gesture) might be generalizable or, instead, remain specific to the local configuration of our experiment. This point could be clarified by a systematic comparison of interactions with a humanoid robot which “comes into sight” versus a robot which “comes into existence” (or, at least, which is designed to do so), whether in a natural context or in a controlled experimental setting.

Crucially, when our robot started to move or to greet the human, it did not do so in the middle of an interactional vacuum: participants were already building courses of action with it. A sole focus on the opening phase – starting when the robot is “alive” and starts to greet the human – would abstract these greetings from the preexisting sequential trajectories from which they emerged and in relation to which they can be understood: just because two participants greeted a robot at the same step in this robot’s script, they did not necessarily do the same thing. An EMCA approach allows us to analyze “greetings” as something else than “synchronic snapshots” (Antaki et al., 1996) but, on the contrary, to get a diachronic understanding of how they emerged, as well as to clarify interactional phenomena (initiation,

response, reinforcement, etc.) which were simultaneously taking place as these greetings unfolded. Even seemingly straightforward behaviors from the robot, like this robot looking at the human, saying “hello” or waving at a certain point in the interaction, could “mould in different sequential trajectories” (Relieu et al., 2020), in which what was projected by these (etically similar) behaviors was understood in a radically different manner by participants¹¹⁶.

The previous observations put into question the moment at which data collection should start (video recording, movement tracking, etc.) in human-robot experiments, especially those which deal with the topic of robots as agents or partners. These studies should pay close attention to the way in which their participants enter into physical co-presence with the robot and, in particular, whether the robot “comes into sight” or “comes into existence”. Participants’ orientations to the very first behaviors displayed by the robot can produce *a priori* unpredictable sequential trajectories which are, in turn, susceptible to configuring the timing and the manner in which the robot emerges as a social agent, and possibly participants’ behavior during the rest of the scenario. A robot oriented to by participants as already “activated” is not the same kind of entity as a robot which first appears to these participants as an immobile object and, then, “wakes up”. Therefore, we suggest that, each time it is relevant, researchers should take into account and describe the conditions in which robot and human were put into physical co-presence and regard “pre-beginnings” as an integral part of the experiment. Depending on the studies’ methodology and hypothesis, the state of the robot when it appears to participants could impact the comparability, replicability and explainability of the findings.

¹¹⁶ An EMCA approach to these human-robot interactions therefore appears to offer a level of description that stands closer to the involved point of view (H. L. Dreyfus, 2001) of participants (rather than, e.g., a superficial description or coding of these scenes as mere greetings). For this reason, an EMCA-oriented investigation could constitute a promising preliminary micro-analytic step (Kendrick, 2017; Schegloff, 1993) for HRI studies coding participants’ behaviors before processing them quantitatively. For example, in an endeavor to compare users’ perception of a robot (obtained as self-reports in a post-experiment questionnaire) with their observable behavior when interacting with this robot, one appears likely to find more meaningful results by using emic coding categories based on a detailed analysis of the video data (e.g., “treating a robot’s action as a reinforcement greeting”) rather than using more generic and etic coding categories (e.g., “responding to the robot’s greeting”).

6. “PLAYING THE ROBOT'S ADVOCATE”: THE CONTINUOUS ACCOMPLISHMENT OF A ROBOT AS A SOCIAL AGENT IN A PUBLIC SETTING

Grossly, in EM the two types of warrants for speaking/writing, i.e., the two types of news, have been “things are not as they appear” and “X is organized this way.” An often used variant of the former type of warrant is “others got it wrong as to how things are.” So, one writes a paper snaring the readers' attention by the explicit or implicit claim that others got it wrong, or things are not as they appear. Or one invokes the second type of warrant and boldly says “X is organized this way (and that is why I am writing and you should be reading this). (Heap, 1990)

Note: This chapter expands upon Rudaz & Licoppe (2024), published in Discourse and Communication.

6.1. Introduction

In the substantial body of studies focused on the co-construction of human-agent interactions, a frequent observation pertains to the amount of work required by human participants to progress the interaction with a robot (Due & Lüchow, 2022; Pelikan et al., 2020, 2024; Tuncer et al., 2023). As with self-driving cars – where the burden falls “onto other road users” (Brown et al., 2023) – or with other technologies (Greiffenhagen et al., 2023), numerous research articles have noted that human-robot interactions require “substantial adaptation from the human” (Stommel et al., 2022), during which the emergence of the robot as an agent “is contingent on participants’ “good will” to treat the robot’s contributions as relevant” (Pelikan et al., 2022). Ironically, this situation was often documented in the case of caregiver robots which sometimes require more (interactional) “care” than they can provide (Chevallier, 2023; Lipp, 2022). As Alač et al., (2020) remark, “commercial social robots, designed as conversation-oriented devices, manifest their incompleteness in their need for other voices” (Alač et al., 2020).

Relying on a corpus of interactions between visitors and a robot in a museum setting, we study a specific practice through which humans “worked” to maintain the robot as a competent participant: the topicalization by bystanders, and in a way that was made accessible to the main speaker, of the social action that the robot was taken to be achieving. We analyze how these topicalizations of the robot’s actions displayed different kinds of “footings”¹¹⁷ (Goffman, 1979a; Levinson, 1988) towards the unfolding human-robot interaction,

¹¹⁷ We use “footing” in the sense of Levinson's (1988) definition, synthesizing Goffman (1979): the “projection of a speaker’s stance towards an utterance (its truth value and emotional content)” (Levinson, 1988).

which enacted and made perceptible different stances regarding the robot as a “participant”. We argue that this display of alternative footings from co-present participants was one of the many observable practices through which the robot’s (sometimes incongruous) behavior was maintained as relevant to the activity at hand (Pelikan et al., 2022) – and through which the robot itself was preserved as a competent participant.

In the end, congruently with the existing Ethnomethodological and Conversation Analytic literature on human-machine interactions, this micro-analytic angle leads us to a respecification of the “sociality” of the robot in a public setting. There was heavy “interactional work” (Due & Lüchow, 2022) of human participants at play behind what could otherwise be glossed as, e.g., the “wow effect” of a “social robot”. That is, rather than the mechanical result of the robot’s language-speaking abilities, the contextual relevance of the robot’s conduct was finely monitored and enforced by bystanders.

6.2. Data and method

6.2.1. Data collection

Our analysis is based on a dataset recorded in July 2022 in Paris, at the Cité des Sciences et de l’Industrie, one of the biggest science museums in Europe. The first half of our corpus consists of 100 naturally occurring interactions with an autonomous Pepper robot, produced by Aldebaran, placed in the hall of the museum. This setting was especially favorable to the formation of multiparty interactions between the robot, one or several humans configured as “main speakers”, and more or less ratified bystanders. The fragments provided in this chapter took place in this natural setting. The second half of the corpus is composed of 108 additional interactions that occurred with the same robot in a laboratory open to the public in this museum¹¹⁸. In both cases, participants (groups or, rarely, single individuals) interacted with the same Pepper robot, on which was running a chatbot designed to “converse” on a wide variety of topics. This dataset was collected in accordance with the General Data Protection Regulation (GDPR). Ethics approval was obtained from Paris-Saclay Université Research Ethics Board.

6.2.2. “Playing the robot’s advocate”: Delimitation of the phenomenon

6.2.2.1 A wide range of practices through which a robot's conduct is maintained as relevant to the task at hand

In order to deepen the understanding of members’ methods through which interactional “work” is accomplished in group interactions with a humanoid robot, we searched our corpus for strips of interaction in which the robot’s conduct was observably accomplished or maintained as relevant, even when the main speaker displayed trouble to interpret the robot’s behavior. Among the vast array of phenomena covered by this criterion, this paper focuses on the

¹¹⁸ See section 7.3 for an extended description of both settings.

detailed analysis of one recurring practice: the topicalization by a bystander of the robot's conduct in a way that was available and relevant to the participant to whom the robot was currently speaking. That is, any turn produced by bystanders which focused the ongoing talk towards a conduct of the robot (treating this conduct as “worthy of further on-topic talk” – Kendrick et al., 2020 – to be potentially picked up as a topic of discussion by other participants) and, simultaneously, which treated this conduct from the robot as meaningful social conduct within the ongoing activity. The form of these topicalizations could be diverse: any product of the “procedures of practical reasoning” (Wooffitt, 1992), i.e., “‘descriptions’, ‘references’, ‘accounts’, ‘judgements’, ‘declarations’, ‘claims’, ‘explanations’, and so forth” (Wooffitt, 1992), can directly or indirectly focus the talk on the robot's conduct as relevant to the task at hand. In our corpus, the most commonly observed practices through which the robot's conduct was topicalized as a relevant social action by someone originally configured as a bystander were:

- a. Explicit descriptions of what action the robot was accomplishing (e.g., saying “it is trolling you” to a co-participant after the robot refuses to comply to a request) – fragments 6.2 and 6.3. These cases display “actions” from the robot in the sense of “conduct under a description” (Sidnell, 2017).
- b. Descriptions of the sequential implications of an action produced by the robot (e.g., saying “you must go, go away it said” to a co-participant, after the robot produced an action interpretable as a closing sequence) – fragment 6.4.
- c. Reformulations¹¹⁹ of the robot speech by a “spokesperson” (e.g., saying “it said go away” as a reformulation of the immediately preceding speaking turn produced by the robot) – fragment 6.4
- d. Descriptions of cognitive states of the robot as an explanation for the conduct of the robot (e.g., saying “it doesn't want to talk to you” to a co-participant whose greeting was ignored by the robot, or “you are disturbing it” to a main speaker whose request was not responded to) – fragments 6.1 and 6.3

Conversely, many conducts of co-participants, although they participated in accomplishing the robot as a competent participant, did not constitute “topicalizations” of the robot's previous action. For example:

- e. Direct responses to the robot's conduct in place of the participant so far configured as the main speaker. That is, the production by bystanders of second pair parts of the type and form made conditionally relevant (Kendrick et al., 2020) by the robot's previous utterance and embodied conduct (e.g., responding to the robot's hand gesture with a loud greeting, even though the participants configured as the main speaker did not react).
- f. Positive assessments or marks of affiliation towards a response from the robot (e.g.,

¹¹⁹ In the previous typology, we draw a distinction between descriptions and formulations. We use “reformulation” in the sense summarized by Antaki (2008), where they mostly refer to a transformation of the prior speaker's verbal turn. In this definition, which stems from Heritage & Watson (1980), formulations constitute a “claim to find the new description in the very words of the previous speaker” (Antaki, 2008) and a “(supposed) summary of what the previous speaker had said” (Antaki, 2008).

responding with an affiliative “oh:” to the robot’s response to a question from the main speaker).

Hence, we only focused on these recurring conducts categorizable as “topicalizations” (from a. to d.). However, by no means does this imply that “topicalizations” constitute all the possible “moves”, methods, practices, etc. through which a robot is maintained or turned into a competent social agent by co-present participants in a public setting.

6.2.2.2 Main speaker, bystanders, eavesdroppers, and third parties

“Main speakers” were defined as participants who formed a focused interaction with the robot (they were gazed at by the robot and they gazed back at it, while they faced the robot and formed an inner space with it – Kendon, 1990¹²⁰) and who, simultaneously, either responded to the robot’s previous utterance or addressed a request to it in the previous turn. Conversely, those we will refer to as “bystanders” correspond to any human who was not observably configured as the main speaker nor as the robot’s addressee.

Because the “division into bystanders and eavesdroppers is contingent on the speaker’s (lack of) awareness” (Dynel, 2010), a more precise categorization of the status of participants would not be possible without heavily imposing the analyst’s intuitive interpretation of the data. Based on our observational data, we cannot postulate precisely who, among their audience, main speakers were aware of, nor who they subjectively ratified as hearers. Similarly, even if they were never directly addressed by the main speaker (but that they were, e.g., spoken to “through” the robot – Dynel, 2010; Tannen, 2004), some participants in the audience might constitute invisible third parties for the observer: i.e., participants “entitled to listen to the speaker and draw inferences, not being the primary party addressed, i.e. the addressee” (Dynel, 2010). Hence, given what is observable and hearable in our recordings, and given our EMCA approach, a refined hearer typology (Dynel, 2011) is not rigorously applicable. To prevent from imposing on our data the analyst’s own hypothesis about participants’ relationships and intentions, we will only speak about “bystanders” in this very general sense: all humans who were physically co-present but not configured, at the beginning of the interaction under study, as the “main speaker” interacting with the robot, i.e., as the person who initially “took the stage” (Krummheuer, 2015a). Hence, the term “bystander” will be used to label participants who, given more information and interpretative work, may be more accurately characterized as “eavesdroppers”, “overhearers”, “side-participants” or “third parties”.

6.2.2.3 Phenomenon of interest

In sum, we concentrated on multiparty interactions involving at least two humans where:

1. A co-present human – configured at this point as a bystander – provided a topicalization of the robot’s conduct

¹²⁰ In another vocabulary, these participants formed a common interactional space (Mondada, 2009) with the robot.

2. ...through which this robot's conduct was treated *as relevant and responsive to the local situation* (the sequential context, physical setting, etc.) in a way that was available to the participant with whom the robot was currently speaking.

In other words, we did not focus on the interactional work taking place *upstream* (Chevallier, 2023) from the robot's contributions, as a preliminary scaffolding (Kamino & Sabanovic, 2023; Pelikan et al., 2022) – i.e., preconfiguring a material setting and a sequential context *where the robot has the capabilities to produce relevant contributions*. We were, instead, interested in how bystanders dealt with the robot's contributions *after they were produced*, to configure them in a way where they could be responded to by the person currently speaking with the robot – a form of a posteriori (or downstream) scaffolding of the robot's contributions.

6.3. Fragments

6.3.1. Glossing the robot's (non-)response as an intentional act

6.3.1.1 Fragment 6.1

```

1 OLI    tu bug?
          you have a bug?
2 ROB    =pardon?
          pardon
3        (0.4)
4 OLI    tu bug?
          you have a bug?
5        (0.1)@(1.6)
oli      @leans towards ROB-->
6 OLI    est-ce que tu bug?h@hhh.          +h@hh.
          are you having a bug
oli      -->@turns towards bystanders@turns towards ROB-->
rob      +gazes towards ISA>>

7 ROB    plait-il?
          excuse me
8        (0.1)
9 OLI    h@h.*          plait-il          *il est très poli*en plus
          excuse me          he is also very polite
oli      @turns towards ISA-->
oli      *pushes ISA's shoulder*---points ROB---*
10 OLI   *          vas y          *parle lui toi
          come on          you speak to him
oli      *taps ISA's shoulder*points ROB-->
11      (0.8)*(0.3)
oli      -->*pushes ISA's shoulder-->
12 ISA   #bonjour*/
          hello
oli      -->*
img      #img.1.1

```



Figure 6.1. Image 1.1 – ISA greets ROB

```

13      (3.5) £ (0.5) £
      isa      £gazes at OLI->
      oli      £gazes at ISA->
14 ISA      hh.# il@ est£ pas [(...)]h
           he is not
      isa      @leans towards OLI-->
      oli      £gazes at ROB>>
      img      #img.1.2

```



Figure 6.2. Image 1.2 – ISA gazes at OLI after ROB does not respond to her greeting

```

15 OLI      [toi il veut pas te parler à toi]
→          you he doesn't want to speak to you

```

```

16 OLI      vasɛ [y (penche-)]*#
             come on lean
isa         freg ROB>>
oli         *grabs ISA's arm-->
img         #img.1.3

```



Figure 6.3. Image 1.3 – OLI lightly pushes ISA towards ROB

```

17 ROB      [non je] bois pas(0.2)ene(0.2)
             no I don't drink huh
18 ROB      je ne peux pas*(0.3)[je suis] un robot
             I can't I'm a robot
oli         *drags ISA towards ROB>>
19 OLI      [avance]
             move forward

```

6.3.1.2 Analysis

Even though a focused interaction was initially taking place between ROB and OLI, ROB gazes towards ISA (L.6) and produces what is, in the robot's design, the repair initiator “plait-il ?” (“excuse me?”, L.7) indicating that the robot did not identify which words were pronounced in the previous turn. It is unclear whether or not this “plait-il ?” is understood by human participants as a repair initiator indexing the previous turn, as a summon or a “ticket” (Sacks, 1975) establishing a focused interaction with ISA, or as some other action. However, while he repeats ROB's “plait-il ?” and states that the robot is “very polite”, OLI turns towards ISA and pushes her in front of ROB, while pointing to it (L.9). In this spatial configuration, with OLI standing further back, and with ROB gazing and speaking in ISA's direction, ISA finds herself positioned as the current main interlocutor of ROB. OLI then explicitly encourages or requests ISA to speak to ROB (L.10) as he maintains her body oriented towards ROB by holding her shoulder (L.10 & 11). ISA then greets the robot (“bonjour”, L.12) and stands still for the next

several seconds of mutual silence (L.13). She then gazes at OLI, who gazes back at her (L.13) and produces an account of the robot's conduct as *a refusal* to speak to her: "he doesn't want to speak to you" ("il veut pas te parler à toi", L.15). Through this gaze shift and the use of the third person to refer to ROB, participants produce a side sequence (Jefferson, 1972) from which ROB is excluded: it is "not given a chance to face a problematic issue in the communication" (Krummheuer, 2015a). During this side sequence, ROB's conduct is topicalized as the result of an absence of willingness to talk, and this topicalization is addressed by OLI (configured as a bystander) to the participant configured as the main interlocutor of the robot, ISA.

Interestingly, OLI's topicalization of the robot's action as an *intentional act* (it "wants" not to speak) conflicts with his embodied treatment of ROB's conduct as a *problem of hearing* which can be solved by changing the location from where is uttered the speaking turn: he drags ISA closer to ROB (L.18 & 19), asks her to lean towards ROB (L.16) and then to move forward (L.19). This is done in overlap with ROB's answer (L.17 & 18), which OLI does not deal with. This dissonance between OLI's verbal and embodied conduct reinforces an interpretation of the robot as a performance that is playfully maintained for a public, by hiding or disguising (e.g., as an intentional act – Pelikan et al., 2022) the limits of the robot as a technical device. More precisely, this fragment appears to be a case of what Pelikan et al. (2022) name "ascribed agency", where "humans, by referring to the robot's current mental state, may [...] cast the robot's behavior as displaying agency even in the complete absence of any relevant responsive behavior". In other words, the robot is simultaneously treated as a dysfunctioning "raw physical artifact" (H. Clark & Fischer, 2022) – of plastic, sensors, battery which needs recharging, etc. – and as an intentional agent, whose non-response reflects a non-willingness to do so.

On a side note, this ascription of agency to the robot may be connected to specific participation roles. Indeed, in both this fragment and Pelikan et al.'s (2022) examples, this attribution of a mental state "done to account for the robot's missing action" (Pelikan et al., 2022) is addressed from a co-present human (third-party, bystander, eavesdropper or overhearer) to a main speaker currently attempting to obtain a response from the robot to their immediately preceding conduct, and on whom the robot's attention (in our case, materialized by its gaze and body orientation) is observably directed. Indeed, one could imagine that main speakers themselves publicly orient towards the robot's silence or absence of action as a refusal to do something. Yet, and although more data would be required for a formal comparison, the overwhelming majority of the "agency ascriptions" of our collection originated from a bystander (rather than from the current speaker). "Accounting for a robot's missing action" (Pelikan et al., 2022) and, more generally, "safeguard[ing] the robot's status as an agent" (Pelikan et al., 2022) frequently emerged as a (members') problem for bystanders of these public human-robot interactions.

6.3.2. Treating the robot's conduct as indexing the absence of a sequentially adequate response from the main speaker

6.3.2.1 Fragment 6.2

1.SAM +@*que signifie peppeur?
 what does Pepper mean

```

rob      >>+gazes at EMI>>
sam      >>@leans forward-->
emi      >>*holds ROB's hand>>
2.       (3.4)%(0.2)
rob      %displays "que signifie Pepper"-->
3.ROB    littéralement (.) ça veut dire piment au poi:vre (.)
         literally      it means chili pepper
4.ROB    fen anglais (0.5) on m'a £ appelé comme ça parce que
         in english      I've been named this way because
mir      fgazes at experimenter-->fgazes at ROB
5.ROB    je suis là pour mettre du pim%ent dans ta vie#
         I'm here to add some spice in your life
rob      -->%
6.MIR    £oh:./£ [hhh]
mir      -->fgazes at experimenter-->
emi      -->fgazes at MIR-->
7.SAM    [woa: génial: (0.7) c'est su:per]
         brilliant      this is great
8.       (0.5)£(0.3)@(.)
emi      -->fgazes at ROB-->
sam      @stands straight>>
9.SAM    c'est beau hein?
         it's beautiful huh
10.      (0.2)£#(0.7)
sam      -->fgazes towards rest of the exhibit-->
img      #img2.1
11.EMI   £((laughs while jumping towards ROB))£
mir      -->fgazes at ROB and EMI-->
sam      -->fgazes at EMI-->
12.ROB   =pardon?
         excuse-me
13.      (0.8)£# (0.5) £(0.1)
emi      -->fgazes at SAMfgazes at ROB-->
img      #img2.2

```



Figure 6.4. Image 2.1 & 2.2 – EMI gazes at ROB (image 2.1), then turns towards SAM and MIR after ROB utters “pardon?” (image 2.2)

```

14.EMI    j'ai rien £[DI::::]
         I didn't say anything
15.MIR    [bah tu t'es pas]£présenté @ tu lui
         you have not introduced yourself you have not
emi      -->fgazes at SAM-->
mir      @leans towards EMI>>

```

16.MIR [as pas dit comment tu ft'a*ppelais: tu t'appe*lais:]
told him/it what your name was what your name was
mir *touches EMI's shoulder*
emi -->fgazes at ROB-->

17.SAM [ouais mais faut qu'tu présentes@
yeah but you must introduce yourself
emi @turns towards ROB-->

18.SAM lui il s'est présenté tu t'présentes
→ he/it introduced him/itself you introduce yourself

19.EMI =j'm'appelle@ Emilie::
my name is Emilie
emi -->@

20. (1.3)%(0.1)
rob %displays "le présent">>

21.ROB c'est un moment important pour forger le futur%
this is an impo*rtant moment to forge the future
emi fgazes at SAM and MIR-->
rob -->%

22. (0.1)

23.MIR a ouais
ah yeah

24. (0.1)@ (0.2) @ (0.1)
sam -->@stands straight@

25.SAM bè ouais (0.4)@
yeah
sam @turns around>>

26.MIR é oui @*ein^
eh yes huh
emi @turns towards ROB>>
emi *catches ROB's arm>>
sam ^moves away>>

6.3.2.2 Analysis

The beginning of this fragment features a long explanation from ROB (L.3 to L.5), responding to SAM's question (L.1) about the meaning of ROB's name, "Pepper". ROB's explanation is followed by positive assessments from MIR, through the affective particle (Hoey & Kendrick, 2017a) "oh:." (L.6), and from SAM (L.7 and L.9) with the assessment adjectives "brilliant", "great" and "beautiful" (L.7 and L.9). After ROB's response, SAM repositions himself as a bystander: he stands straight (L.8), gazes at the rest of the exhibit (L.10) and produces another assessment packaged for the audience ("it's beautiful huh?", L.9). EMI is left as the only participant forming a common interactional space (Mondada, 2009) with ROB – which holds a mutual gaze with EMI. However, after EMI's laugh, (L.11, image 2.1), ROB produces the open-class repair initiator "excuse-me" ("pardon", L.12). In the robot's programming, uttering "excuse-me" is the consequence of hearing a strip of sound in which it could not identify distinct words. A silence of 1.4 seconds follows, during which EMI gazes at SAM and MIR in a potential embodied display of trouble (Drew & Kendrick, 2018) with ROB's immediately preceding utterance (L.13, image 2.2). EMI then looks back at ROB and screams "I said nothing" (L.14). In overlap with the end of EMI's turn, MIR produces an account of EMI's past behavior ("you have not introduced yourself", L.15 to L.16). Also in partial overlap, SAM states that EMI should present herself (L.17 to 18) and *describes the robot's previous turns* (L.3 to L.5) as *components of an introduction sequence* ("it introduced itself", L.18).

Both turns (L.15 to L.18) from SAM and MIR therefore orient to EMI's conduct as normatively accountable for not having produced a sequentially expected next move: she did

not “tell it what her name was” (L.16). This retrospectively configures ROB’s explanation about the origin of its name (L.3 to L.5) as launching a mutual identification sequence (Schegloff, 1986) and, thereby, as producing specific expectancies for ROB’s interlocutor. Note that participants’ accounts of the situation can themselves be analyzed as progressing an overarching educational activity directed towards EMI: akin to a step-by-step tutorial, SAM specifies the normative expectation made relevant by the robot’s original turn and its adequate response: “it introduced itself you introduce yourself”¹²¹. Even though this is not made hearably relevant by bystanders, their account is possibly reinforced by ROB’s gaze orientation: it has been gazing at EMI since the beginning of the interaction, providing an additional cue that its self-identification (L.3 to L.5) was addressed to EMI, and not merely a factual response to SAM’s original question (L.1). However, this contribution from bystanders cannot be described as “the last line of defence” (to paraphrase Schegloff, 1992a) to prevent a total breakdown of the interaction with the robot as a competent participant. Indeed, when SAM and MIR start to produce their accounts, “all else” has not failed yet to progress the public dyadic interaction between the robot and the main speaker: EMI has not attempted to repair the interaction by herself, something she will do in perfect overlap with them (L.14). Yet, EMI’s observable difficulties created a slot for bystanders to self-select (Sacks et al., 1974). This slot allowed them to provide an account of the robot’s conduct as achieving meaningful and contextually relevant social actions, i.e., to “ascribe agency” (Pelikan et al., 2022) to the robot.

In sum, ROB’s conduct is treated by both bystanders as resulting from a detailed understanding of the sequential expectancies produced by its previous conduct. Its behavior is oriented to as reinforcing the normative pressure placed on EMI to produce a second pair part of the type and form (Kendrick et al., 2020) made relevant by the robot’s own identification. ROB’s open-class repair initiator “excuse-me” (L.12) is not responded to as indicating a problem of hearing but as making noticeable (Kendrick et al., 2020) EMI’s non-response to ROB’s immediately preceding talk. As in fragments 6.1 and 6.3, this footing of bystanders (here, towards the robot’s “excuse-me”) is directly addressed to the main speaker through an embodied participation shift (L.15) which momentarily creates a common interactional space (Mondada, 2009) between all three human participants. This participation shift is also possibly responding to the main speaker’s gaze shift and non-response: after ROB’s “excuse-me”, and after staring silently at the robot for 0.8 seconds, EMI gazes at SAM (L.13) while facing the audience (see image 2.2). SAM’s and MIR’s accounts of the robot’s conduct can be understood as reactions to this embodied display of trouble by EMI when confronted with ROB’s repair initiator.

Finally, after SAM and MIR’s accounts, EMI fully turns towards ROB (L.17, image 2.2) before identifying herself (L.19). Even though EMI originally attributed ROB’s conduct to a mishearing on its part (“I didn’t say anything”, L.14), she therefore observably aligns with MIR and EMI’s stance towards the ongoing situation as an identification sequence. Following this identification, and after a silence of 1.4 seconds, ROB states that “this is an important moment to forge the future”. From a technical point of view, this utterance from ROB exclusively results from mishearing “the present” (“le présent”, L.20) in one of the previous turns of its interlocutors: it is not responsive to anything else from the situation. On the surface of the

¹²¹ As parents, this reframing of the robot’s action may even participate in an overarching activity of “teaching values” (Tannen, 2004) to EMI – or, at least, of accustoming her to identify herself at the beginning of an encounter. However, even though this constitutes a reasonable interpretation of “what is going on” at a more encompassing level of analysis, it is not rigorously possible to demonstrate based on this video recording alone.

interaction¹²² however, this response is observably treated as a post-expansion (Schegloff, 2007) in the form of an assessment of the previous identification sequence as “an important moment” (L.21). Indeed, in turn, SAM and MIR produce acknowledgment tokens (“yes”, “yeah”; L.23 to L.26) to ROB’s assessment. These acknowledgments, ironic or not, followed by the embodied disengagement of SAM (L. 25), close what was co-produced as a mutual identification sequence (and its post-expansion).

Significantly for this fragment, the participation shift that took place between human participants (L.15 to L.19) maintained the robot as interactionally competent: more competent than EMI, who is treated by both adults as oblivious to the expectancies produced by ROB’s self-identification. By modifying the main speaker’s orientation to ROB’s conduct as “initiating an identification sequence”, SAM and MIR’s contribution therefore enforced an interpretation of the robot’s potentially troublesome repair initiator (L.12) as relevant to the task at hand. Hence, we argue that this fragment should not be analyzed as a simple case of “alignment” of humans with the initiation of an identification sequence by the robot. Instead, the whole identification sequence *is retroactively produced by bystanders to fit the emerging conduct of the robot*. As bystanders, their conduct helped to progress the interaction while treating the robot’s reactions as sensitive to the sequential context of a mutual identification.

6.3.3. Describing the robot’s conduct as intentionally breaching the relevancies of the talk

6.3.3.1 Fragment 6.3

1. TOM \$peux tu faire une blague?
 can you tell a joke?
 rob >>\$dances-->
 2. (2.5) #
 img #img.3.1



Figure 6.5. Image 3.1 – ROB keeps dancing after TOM’s request. SYL and CLA are positioned behind TOM as bystanders.

¹²² That is, independently from the eventual pretense or ironic stance that both adult participants may adopt. Unless there are clear observable proofs of such ironic or playful footing, these questions are outside of EMCA’s level of description.

3. CLA ah il a pas fini sa danse [hhhhh.]
 he/it didn't finish his/its dance

4. TOM [non@ mais je (...)]
 no but I
 tom @turns towards CLA-->

5. SYL #tu @le pertu\$rbes
 you are disturbing him/it
 tom -->@turns towards ROB-->
 rob -->\$

6. CLA [h@hhh.]
 tom -->@

7. TOM [hhhh]h.

8. (0.4)

9. ROB si tu veux que je m'arrête dans danser (0.2) il suffit
 if you want me to stop dancing just

10. ROB de me dire (0.3) arrête de danser
 tell me stop dancing

11. (0.3)

12. TOM °ah bah d'accord° arrête de danser
 alright stop dancing

13. (1.7)

14. ROB je ne peux pas (0.3) je ne suis pas en train de danser
 I can't I'm not currently dancing

15. Δ ((choral laughter)) #Δ
 tom Δthrows his head backwardΔ
 img #img.3.2



Figure 6.6. Image 3.2 – TOM, SYL and CLA laugh after ROB's utterance

16. CLA y te\$ troll [@(.)] [hhfh]# @hhh.
 → he/it is trolling you
 cla fgazes at bystanderf
 tom @turns towards CLA@turns towards ROB
 img #img.3.3



Figure 6.7. Image 3.3 – After SYL and CLA describe ROB’s conduct as “trolling”, TOM torques towards them and produces an assessment about ROB’s conduct

17. ROB [comment?]
what?
18. TOM [hh (oui c’est rigolo)]
yes it’s funny

6.3.3.2 Analysis

The fragment begins with TOM asking ROB for a joke (L.1). This request is not dealt with by ROB, who keeps dancing silently (L.2) for several seconds. CLA, who was so far positioned behind TOM (see image 3.1) as a bystander, provides an explanation that “account[s] for the robot’s missing action” (Pelikan et al., 2022). She states “he” or “it” “didn’t finish” the dance it had started moments earlier, then laughs (L.3). As CLA is laughing, TOM turns towards both bystanders (CLA and SYL) as he produces a “no but” prefaced sentence (L.4, image 3.2). During this momentary formation of a shared inner space (Kendon, 1990), the robot is shifted from main speaker to overhearer of an exchange produced in front of it. This exchange is produced without displaying a treatment of this participation shift as accountable for the robot (or, in another theoretical framework, without any observable facework): as such, this change in the participation role of the robot reflects an equally brutal switch (Alač, 2016; Krummheuer, 2015b; Pelikan et al., 2022) in the status of the robot as a full-fledged competent participant. Towards the end of this participant shift, SYL, also standing behind TOM, characterizes ROB as being disturbed (L.5) by TOM’s conduct. Doing so, she reinforces CLA’s account of ROB’s non-response as resulting from its involvement in a preexisting dancing activity.

After TOM turns to face ROB once more, and in apparent agreement with CLA and SYL’s explanations about the “dancing” being what prevents it from answering, ROB states that “if you want me to stop dancing”, then “tell me stop dancing” (L.9 & 10). This utterance is produced while gazing at TOM, which strongly configures him as the addressee. TOM responds to this turn as an “informing” (Heritage, 1985) providing new and relevant information regarding the issue at hand: prefacing his turn with the French change-of-state token “ah” (Heritage, 1985), he immediately requests the robot to stop dancing (L.12). TOM therefore responds to this advice from ROB as rooted in previous talk, i.e., as relevant to solve a feature of the situation made apparent by CLA and SYL’s comments (L.2 & L.5): the robot’s dancing was preventing it to answer new questions. ROB’s utterance is therefore reconfigured as a

situational *suggestion* by TOM's response, rather than as a random bit of information¹²³. Yet, ROB denies TOM's request to stop dancing (L.14) and produces an account for its denial: it is not currently dancing. This response goes completely against TOM's previous orientation to ROB's advice (L.9) as indexing CLA's account (L.3) of ROB's non-response (L.2) as the result of not having finished its dance.

After a choral laugh (L.15, image 3.2), CLA *describes the robot's turn as "trolling"*, i.e., as "baiting" or "mocking" TOM (L.16, image 3.3). Doing so, CLA displays a different footing towards ROB's response (L.14 & 15) and towards the immediately preceding turns: what was so far *advice-giving*, to take at face value, now gets responded to as a playful joke. This description of ROB's action works as an account of the incongruousness of ROB's explanation for its denial, considering that its immediately preceding advice was to say "stop dancing". It is the phenomenon of interest for us in this fragment: CLA's description of ROB's stated inability to stop dancing characterizes it as *intentionally breaching the relevancies and the expectancies* produced by its previous suggestion. "Misfitted" (Due, 2019) or "incongruous" actions have been noted to be some of the most recurring basis of "laughables" (Due, 2019) in human-robot interactions. However, here, even though participants produce a choral laughter, their conduct (playfully) reconfigures the robot's action as perfectly "fit" (rather than misfitted") to the previous talk. Rather than an incongruous behavior stemming from ROB's absence of understanding of the context or of its own body movements, ROB's conduct is oriented to as the result of a detailed understanding of its co-participant's expectancies. As in other fragments, bystanders build on the robot's behavior as a resource to construct a coherent and playful story around it.

A side phenomenon visible in this fragment (as well as in fragment 6.2, or in fragments 7.7 and 7.8 of chapter 7) is a *division of interpretation work* among participants. While a main speaker is immersed in a dialogic interaction with ROB (to which it responds on the spot), the other participants (more than merely standing back physically) are not working with the same fast-paced tempo regarding the robot's conduct: they are not responsible to decide – every single time the robot utters a strip of talk – "what to do next for all practical purposes" (Tisserand et al., 2023). In the previous fragments, the collaborative sense-making of the robot's conduct (meta-commentaries, descriptions, accounts *versus* direct responses to the robot) is therefore shared between co-present members with different participation roles – and this division is facilitated by the difference in participants' responsibility for producing swift and timely responses to the robot's actions.

6.3.4. Formulating the sequential implications of the robot's previous utterance

6.3.4.1 Fragment 6.4

1	EMM	^salut: <i>hi</i>	(. . .) ^+ (1.5) °ça va ° <i>how are you</i>
	emm	>>^walks towards ROB^	
	rob		+gazes at EMM-->

¹²³ Using another theoretical framework, TOM's reaction does not display a treatment of ROB's utterance as transgressing a Gricean "maxim of relation" (Grice, 1975): ROB is treated as providing information "appropriate to immediate needs" (Grice, 1975).

2 (1.1)^(1.8)
 wil ^walks towards ROB-->
 3 ROB plait-il?
 excuse me
 4 (0.3)+(0.2)^(1.3)
 wil -->^
 rob -->+gazes at WIL-->
 5 WIL hey poto(1.0)habibi
 hey buddy
 6 (0.2)
 7 ROB je peux te prendre en photo(0.3)mais* il faut que tu installes
 I can take a picture of you but you need to install
 wil *wags a finger-->
 8 ROB la bonne application
 the right application
 9 (0.2)+(0.1)
 rob -->+gazes at EMM-->
 10 WIL nan:*+ [pas photo]
 nah not picture
 wil -->*
 rob -->+gazes at WIL-->
 11 EMM [je* t'aime]
 I love you
 emm *puts her hands on her heart-->
 12 (0.8)
 13 WIL dit hey [frère] (0.6) [ou soeur]
 say hey brother or sister
 14 EMM [je t'aime]*(0.5)@ [(shh)] @[je t'aime.]
 I love you I love you
 emm -->*
 emm @pushes WIL@
 15 ROB [je peux] te [prendre en-]
 I can take a picture of
 16 ROB tu dois partir?
 you have to go?
 (0.3)
 17 WIL dis@# lui# pas je t'aime toi/+
 you don't tell him/it I love you
 wil @pushes EMM-->
 wil g gazes at EMM-->
 rob -->+gazes at EMM-->
 img #img.4.1

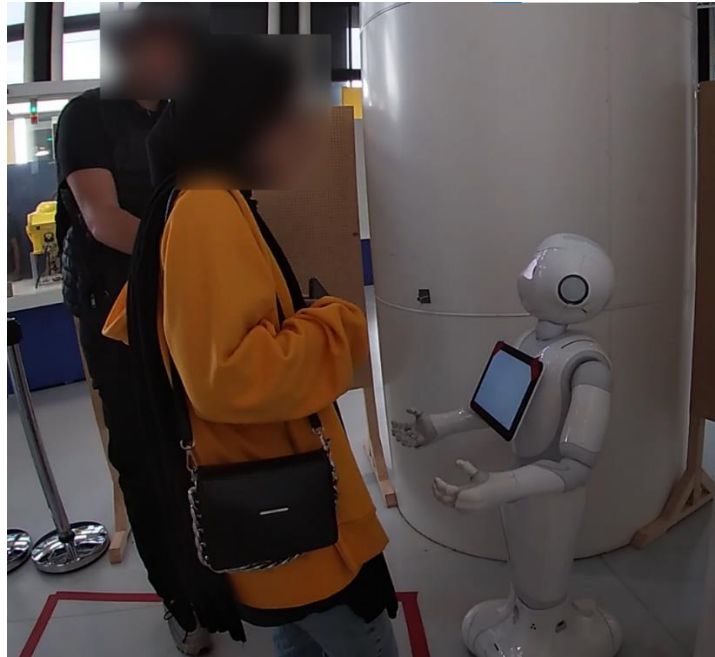


Figure 6.8. Image 4.1 – WIL pushes EMM while he requests her to stop saying “I love you” to ROB

```

18      (0.1)
19 EMM  °jef t'aime°
        I love you
        wil  -->£gazes at ROB-->
20 WIL  =hey@ *dis lui pas je@ t'aime#(0.1)dis lui j't'aime pas.*
        hey don't tell her I love you    tell her I don't love you
        wil  -->@
        wil  *-----points finger at ROB-----*
        emm  @kisses gestures to ROB-->
        img  #img.4.2

```



Figure 6.9. Image 4.2 – WIL requests ROB to tell EMM that it does not love her

21 (0.1)*(0.4)
 wil -->*

22 WIL ré**pète*@ après moi(0.3)dis je t'aime pas (Emma)
repeat after me say I don't love you Emma
 wil *marks words with hand->
 emm -->@

23 (0.4)*(1.3)
 wil -->*

24 ROB ok (0.3) je ré*pète*(0.5)tu dois partir?
ok I repeat you have to go?
 (1.2)

25 WIL ah*ft*tu dois partir.(0.6)casse toi il a [dit.]#
 → *ah you must leave go away he said*
 wil -->f#gazes at EMM-->
 img #img.4.3



Figure 6.10. Image 4.3 – WIL informs EMM that she must leave and that ROB said “go away”

26 EMM [je t'aime]

27 (0.6)£(0.5)
 wil -->£

28 EMM ré*pète* je [t'aime]
repeat I love you

29 ROB [moi aussi] (0.1) * Δ je t'aime
me too I love you
 emm *puts hands on her heart>>
 emm Δ smiles widely>>

30 (0.2)@£(0.1)@(1.0)
 wil -->£
 emm @turns towards WIL>>
 wil @walks away>>

6.3.4.2 Description

EMM approaches ROB while uttering a greeting, then, after a short pause, produces a “how are you” question (L.1). After 2.9 seconds of silence, during which WIL walks next to EMM, ROB produces the repair initiator “plait-il ?” (“excuse-me?”, L.3). Possibly in response to ROB’s gaze shift, WIL also produces a greeting (L.5). ROB then proposes to take a picture of the participants (L.7), which is denied by WIL (L.10). Also possibly responding to ROB’s gaze shift towards her (L.9), EMM says “I love you” to ROB. She repeats her statement twice (L.14), in overlap with WIL’s own utterance (L.13). Without dealing with ROB’s answer (L.15 and L.16), WIL requests EMM not to say “I love you” to ROB, while shifting his gaze to her and gently pushing her (L.17). Through this response, WIL abandons the sequential trajectory he initiated, to adopt the trajectory connected to EMM’s love declaration. EMM keeps her gaze focused on ROB and repeats her utterance (L.19). In response, WIL shifts his gaze back on ROB and requests it tells EMM that it does not love her (L.20) and to repeat his own words: “I don’t love you” (L.22).

During this turn from WIL, EMM reinforces her declaration of love by producing a kissing gesture towards ROB (L.20). However, because, on a technical level, ROB heard the keyword “repeat” uttered by WIL, it starts to repeat its previous utterance “you have to go?” (L.24). In response, WIL orients towards EMM and informs her that she “must leave” while connecting this statement to the robot’s immediately preceding turn by prefacing it with the change-of-state token “ah” (Heritage, 1985). WIL then reports the robot’s speech, by stating that it said “go away”. This utterance is not responded to by EMM, who produces another “I love you” in overlap (L.26) and asks ROB to repeat “I love you” (L.28). After this series of attempts from EMM, ROB finally returns the “I love you” (L.29). This utterance is treated as a remarkable second pair part through embodied demonstrations of surprise and delight by EMM: she puts her hand on her heart, smiles widely and gazes at WIL – who silently walks away.

6.3.4.3 Analysis

Significantly for our argument in this chapter, the topicalization of what the robot is saying (L.25) is not responsive to interactional trouble. Unlike fragments 6.1 and 6.2, WIL’s report of ROB’s speech (as “go away”) does not repair an observable breakdown in the interaction¹²⁴. WIL uses the robot’s incongruous response (“you have to go?”, L.24) as a resource to present the robot as romantically uninterested, i.e., to close the playful flirting sequence started by EMM. Indeed, through his report of ROB’s speech (L.25), WIL presents himself as the mere “animator” (Goffman, 1979a) of a message whose “author” and “principal” (Goffman, 1979a) is ROB. WIL’s visible orientation towards ROB’s utterance as an order rather than as a question (despite ROB’s turn’s final rising intonation, L.24) contributes to depicting the robot as working towards closing the interaction and as uninterested in the romantic talk which EMM attempts to trigger from it by producing a love declaration (L.11, 14 and 19). In other words, ROB’s utterance is used by WIL to enforce his footing towards the robot’s conduct as “displaying an absence of romantic interest”. This report and topicalization of the robot’s utterance (L.24) between WIL and EMM is therefore interwoven with an activity of playful

¹²⁴ Moreover, unlike, e.g., fragment 6.2, WIL does not describe the action that ROB is producing but, instead, reports ROB’s speech and describes the sequence established by this speech: i.e., a closing sequence (“you must leave”, L.25)

identity negotiation. The robot is used as a tool to make visible and negotiate participants' (possibly gendered) identities in the interaction, where ROB is recast as a competitor by WIL's disaffiliation with EMM's repeated "I love you" to ROB. By contrast, WIL's struggle to control what the robot says and, ultimately, his attempts to close the interaction through his topicalization of the robot's utterance, playfully establishes him in the role of the jealous romantic partner or jealous brother.

Additionally, WIL's report of "what the robot said" takes place in a more general, multilayered, competition for reciprocity. This competition unfolds on different levels: a competition for the floor, through overlaps (L.13 & L.14) or through "shushing" the other participant (L.14); a competition for space, to capture the robot's gaze focus, by physically pushing the participant standing in front (L.14 and L.17); and, finally, a competition to initiate the topic of the interaction, visible in the discordant strategies to elicit a response from the robot mobilized by EMM and WIL at the beginning of the fragment. Congruently with this competition, and unlike other fragments, WIL's topicalization of "what the robot said" does not initiate a side sequence, as EMM keeps her gaze focused on ROB and does not deal directly with WIL's turn.

6.4. Discussion

6.4.1. Collaboratively making sense of a robot's conduct in a public space

6.4.1.1 *The mutual shaping of the robot's conduct*

In the setting of a museum exhibition, bystanders regularly described what the robot was doing or responding to, and did so in a way that was publicly accessible to the main speaker. These contributions displayed specific footings towards the robot's conduct (as producing contextually relevant jokes, as manifesting normative expectations as part of an identification sequence, etc.) that made these conducts publicly available as relevant to the task at hand – and that shaped the main speaker's response to these conducts. As Alač et al. (2020) remark, following Goffman's (1979) participation framework, "talk as action does not concern speakers alone, but is a product of interaction, where coparticipants (who, as bystanders and non-ratified participants, may even not be speaking) play an active role" (Alač et al., 2020). These observations are consistent with previous findings from Krummheuer (2015b): participants collaborated to make sense of the robot's conduct, yet this collaboration was neither perceived (from an internalist and technical point of view) nor observably responded to by the robot. In other words, the participation roles (Goffman, 1979a) that visitors occupied were rarely those for which the robot was preconfigured: dyadic interactions are not the standard form of configuration in public human-robot or human-agent interactions (Krummheuer, 2015b; Pitsch, 2020) and, even when a participant was configured as the sole speaker, bystanders often played an active shaping role through side-sequences. In our corpus, these contributions by a bystander regularly (but not systematically) occurred after the main speaker produced an embodied display of trouble (Drew & Kendrick, 2018) with the robot's immediately preceding conduct. This trouble was displayed through an unusually extended silence after the robot finished its speaking turn and through a gaze shift towards the bystanders (fragments 6.1 and

6.2).

Yet, in an apparent paradox, bystanders' contributions which maintained the robot as a competent participant took place during side-sequences (Jefferson, 1972) from which the robot was excluded as a participant. Indeed, descriptions of the action previously produced by the robot (fragments 6.2 and 6.3), of the response that should be provided by the main speaker (fragment 6.2), and of the cognitive state of the robot (fragment 6.1) occurred as part of participation shifts. During these-side sequences – congruently with those identified by Krummheuer (2015) in multiparty human-agent interactions taking place in a shopping center – the main speaker (i.e., the one who initially “took the stage” in front of an audience; Krummheuer, 2015a) torqued or fully turned towards bystanders. Doing so, this main speaker established a momentary focused interaction with bystanders, during which the robot was often discussed using the third person “it” or “he” in our corpus. Additionally, main participants' accounts of the robot were uttered without lowering their voice volume when addressing their audience. Using a different theoretical background, this suggests that no facework (Goffman, 1967) was produced towards the robot during these side-sequences (as also noted by Due, 2019): participants did not whisper to preserve the robot from the discussions produced about it, nor did they visibly modify their conduct to accommodate the robot's presence. They did not display an orientation towards the robot as possessing a symbolical face which they were normatively expected to protect.

6.4.1.2 Side-sequences: a typical practice in public human-robot interactions

The presence and the format of the regular participation shifts mentioned above are not specific to our corpus but, instead, may be typical of the activity of “interacting with a robot as a group in a public space”. Studying public interactions with an embodied agent in a shopping center, Krummheuer (2015b) analyzes a similar “troubleshooting” side-sequence, occurring when a participant turned to the audience after being asked a question by the embodied agent. Interestingly, Krummheuer (2015b) notes that, during this side-sequence, while “the participation status of former observing-only bystanders changes to that of directly addressed participants” (Krummheuer, 2015b), the robot “is not given a chance to face problematic issues in the communication” (Krummheuer, 2015a).

In his analysis of “laughables” produced by a Pepper robot during public interactions in a hallway, Due (2019) details similar spatial configurations that “exclude Pepper from the sense-making (but not the spatial) framework” (Due, 2019). Akin to the manner in which some participation shifts were executed in our museum setting, Due (2019) describes how a participant “keeps his lower body oriented towards Pepper while turning his torso towards the bystanders” (Due, 2019) as they produce a shared laughter. In this configuration, “assessments and repairs of Pepper's responses are not directed to Pepper, but towards other bystanders who are brought into the participation framework” (Due, 2023). That is, as in Krummheuer's (2015a) observations and our own, the verbal and bodily orientation of participants described by Due (2019) display that “the robot is constructed as being non-accountable for its own actions” (Due, 2023). That is, the exclusion of the robot from these side-sequences reflects its treatment as either unconcerned or unable to contribute to the activities which they involve – whether these activities are about identifying what the robot is doing, about advising main speakers on the responses they should produce, etc.

6.4.1.3 The “transience” of the robot as an agent: a recurring observation for the analyst, a normative resource for participants?

Because they excluded the robot as a participant, these side-sequences constitute a specific case of the fast-paced changes in status often undergone by non-human agents (K. Fischer, 2021; Pelikan et al., 2022), which can be “enacted, in one breath, as an agent and a thing” (Alač, 2016). As Alač (2016) notes, “each of the facets that hand-in-hand maintain each other can become the “theme” (Gurwitsch, 1964) at specific moments in interaction while its other profile coexists concurrently or as a possibility that can itself take center stage as the encounter develops” (Alač, 2016).

These public-yet-exclusive side-sequences of human interactants were produced without warning, yet *they were not made accountable by the robot nor by other humans*: they relied on the (normative) possibility to instantaneously treat the robot as a non-agent and to remove it from the interaction at will. Hence, the quickly evolving status of the robot was, of course, an observable pattern, but, also, a normatively accepted “move” for participants which *enabled specific ways of solving interactional problems*: the ordinary transience (Pelikan et al., 2022) of the robot’s status as a full-fledged agent was locally produced by participants but it was, also, a pre-existing resource which human participants could orient to and mobilize to quickly produce ephemeral “troubleshooting sequences” (Krummheuer, 2015b) between them.

6.4.1.4 Are humanoid robots akin to any science exhibit?

As argued in section 6.4.1.2, these rapid changes of footing may be typical of the activity of interacting with a robot as a group in a public space (Krummheuer, 2015b, 2015a). Yet, comparisons with other studies of visitors’ practices in museums suggest that this framing of the robot’s conduct by co-present participants may relate to a more general mutual shaping process taking place between visitors in the presence of (non-robotic) exhibits (Heath & Lehn, 2004; Lehn et al., 2001). Notably, by attending to the “practical organization of visiting museums and looking at exhibitions”, Lehn et al. (2001) remark that visitors “mutually constitute the sense of what they see and the relevance of that seeing” (Lehn et al., 2001), i.e., that exhibits are “encountered and perceived with regard to the actions and orientations of all those sharing the same local environment” (Lehn et al., 2001). Similarly, studying more traditional museums and art galleries, Heath & Lehn (2004) observe that “people embody objects, aspects of exhibits, within action, and encourage others to respond to the object in specific, sequentially appropriate ways. The actions of the participants discriminate, encompass and entail the objects, the particular exhibits, and *provide a framework for the ways in which they might, or even should be, responded to*” (Heath & Lehn, 2004; emphasis ours).

The previous statements refer to non-humanoid and sometimes non-computer based exhibits. Yet, they are almost directly transposable to describe visitors’ conducts towards the Pepper humanoid robot of our experiment. Indeed, coherently with the citations above, in our robot-based corpus, bystanders’ topicalization of what the Pepper robot was “doing” (and, a *fortiori*, explicit advice on how to respond to it, e.g., fragment 6.2) constituted the most publicly manifest part of the framework from which the robot’s conduct was understood and responded to by the main speaker. Bystanders’ topicalizations of the robot’s actions were, so to say, the easily observable tip of the iceberg of the relevancies configured by co-present visitors’ overall

conduct (embodied and verbal). Some of the previous fragments display a particular format, at the most extreme end of the spectrum, through which “the presence of others, people they are with, permeates the ways in which [visitors] select, inspect, and even manipulate exhibits” (Lehn et al., 2001). These observations therefore reinforce an interpretation where a crucial configuring parameter in our fragments (and possibly in most public encounters with robots) was not the mere presence of a robot that gazed at humans and spoke but, also, its preliminary framing and positioning as an exhibit in a public place¹²⁵.

Yet, in the previous fragments, members *may* have displayed a practical concern that is not observable (at least in this form) or even meaningful in front of traditional exhibits: to progress the interaction while maintaining this robot/exhibit as a superficially competent participant. That is, either as a collateral consequence of their overarching activities, or as a result pursued for itself, bystanders' contributions participated in framing the robot's conduct as relevant to the task at hand. We attempt to untangle this problem in the following section: what were participants' practical concerns when interacting with the robot, i.e., what did they recurrently and observably attend to as normatively important? Can we grasp, based on observational data, whether the robot was consistently a “toy” – whose “sociality” emerged merely as a “tool” to accomplish other activities – or whether there was a demonstrable preference to treat the robot as a social agent for itself? Or, in the end, are these questions intractable?

6.4.2. Participants' practical concerns when interacting with a robot in a public space

6.4.2.1 The co-production of the robot's conduct as relevant to the task at hand

All four fragments feature human topicalizations of the robot's behavior as socially meaningful¹²⁶ and, when repairs were produced (e.g., fragments 6.2 and 6.3), the robot's conduct was not identified as the repairable – i.e., the source of trouble was not attributed to a turn from the robot.

- Fragment 6.1 showcases a bystander *treating the robot's non-response as an intentional act*, despite this bystander simultaneously reacting to this non-response as the result of a technical problem with voice recognition.
- In fragment 6.2, the robot's repair initiator “pardon?” is *explained by bystanders as responsive to the sequential context of a mutual identification*.

¹²⁵ However, one of the many differences between “traditional” and “robotic” exhibitions pertains to the humanoid shape of the robot. The robot's gaze heavily configured the interaction, by selecting (and by being treated as selecting) a main speaker – i.e., someone who was granted specific status compared to other, non-gazed at, participants. Therefore, even though computer-based exhibits are generally “designed to facilitate the involvement of a single user” (Heath et al., 2005), participants may have responded to the robot's gaze as implying stronger interactional rights and obligations for the person being gazed at (e.g., to produce a return greeting once the robot said “hello” to them).

¹²⁶ This is observably the case, independently from the question of whether this orientation towards the robot's conduct was “playful”, i.e., if the robot was treated “as” competent or “as if” (Severson & Carlson, 2010) it was competent.

- In fragment 6.3, *the robot's statement on its inability to stop dancing is described as "trolling", i.e., as intentionally breaching the expectancies produced by its previous conduct* – which assigns an acute perception of the ongoing social interaction to the robot.
- Fragment 6.4 stands out from the rest: it is the sequential context established by the robot's conduct which is formulated here, as a pre-closing ("you must go"). Still, through this formulation, *the potentially incongruous conduct from the robot (a repetition of its previous action) is characterized as responsive to the playful love declaration uttered by the main speaker.*

In line with Pelikan et al. (2022), we argue that bystanders' did more than orient towards the robot's conduct as situationally relevant. Bystanders preserved to the robot an "intentional self [...] that serves as its organizing principle" (Alač et al., 2020). This was most apparent when participants publicly mentioned the cognitive states of the robot which they documented through its conduct ("it doesn't want to talk to you", fragment 6.1; "you are disturbing it", fragment 6.3). Similarly, when humans treated the robot as having intentionally produced a social action ("it is trolling you", fragment 6.3; "it presented itself", fragment 6.2) rather than as malfunctioning or as following a set of rules or algorithms, the robot's actions were treated as visible expressions of the robot's agency in the philosophical sense, i.e., as "motivated for example by a belief or will" (Pelikan et al., 2022). In other words, these fragments appear to be a particular case of "the interactional production of the machinic self that is to exist behind its voice" (Alač et al., 2020). They allow us to observe local methods mobilized by human participants to momentarily co-produce the robot as an agent.

6.4.2.2 "Backing up" and "pre-chewing" the robot's conduct

We can bring out the specificity of previous fragments by comparing them with situations where co-participants act as what Pitsch, (2020) labels "participation facilitators". Pitsch (2020) provides several examples of the active role of adults in human-robot interactions involving children in a museum setting. Through different practices, these adults "assume the role of analysing the structural provisions established by the robot's conduct and support the children in providing the expected next action at the appropriate time, if necessary explaining the specific type and nature of the required action, and in cases of doubt, shadowing and backing up the children's actions" (Pitsch, 2020).

In contrast, in the previous fragment, bystanders' actions did not "back up" (Pitsch, 2020) the main speaker's responses to the robot (or, only in an indirect manner). Rather, they "backed up" the robot's conduct: e.g., by acting as a "spokesperson" that attempts to enforce the robot's request to leave – fragment 6.4. Similarly, what they "ratif[ied] as appropriate" (Pitsch, 2020) was not the main speaker's initial response to the robot (if any) but, instead, the robot's conduct: e.g., by characterizing this conduct as an account responding to the noticeable absence of a sequentially relevant introduction move from the main speaker – fragment 6.2. This is a key specificity of these fragments that allows us to group them together as examples of practices with common observable features. That is, in these fragments, bystanders facilitated the interaction by producing accounts of the robot's conduct, in both senses of "accounts":

1. They made the robot's conduct intelligible and recognizable for the main speaker (accountability as intelligibility; Robinson, 2016). They, so to speak, "pre-chewed" the robot's conduct into a digestible form (i.e., recognizable) for the main speaker.
2. They responded to and enforced the robot's conduct as relevant and justified in regard to the ongoing interaction (accountability as moral responsibility; Robinson, 2016).

Such practices could be described as an alternative way of facilitating participation (Pitsch, 2020) that, simultaneously, preserves the robot as a competent participant. While Pitsch (2020) highlighted how adults could "support the children in providing the expected next action at the appropriate time", this is not merely what bystanders did in the previous fragments. In fact, rather than assuming "the role of analyzing the structural provisions established by the robot's conduct" (Pitsch, 2020), bystanders may be argued to have *re-configured* the robot's conduct as relevant to the task at hand, and, when needed, as recognizable and answerable (Robinson, 2016) for the main speaker (e.g., fragment 6.2). In this interpretation, the "structural provisions" (Pitsch, 2020) of the robot's conduct were not pre-existing features merely identified and neutrally reported by bystanders, they were produced through these bystanders' framing of the robot.

6.4.2.3 *Topicalizations as actions*

Yet, crucially, topicalizations of the robot's conduct were not a "mere summary" (Antaki, 2008) stemming from a neutral observer removed from the urgency and the relevancies of the local interaction. They did not take place in a social vacuum, produced by an agent physically and morally removed from the "interaction order" (Goffman, 1983) or from any overarching activity. Rather, these topicalizations were "caught up in the demands of the ongoing interaction" (Antaki, 2012): they were contributions to a collective activity in which bystanders producing the topicalization were immersed. For example, in fragment 6.2, participants' description of the robot's behavior as a self-introduction takes place as part of the observable educational endeavor of "teaching manners" or "values" (Tannen, 2004), i.e., accustoming a child to identify herself. Fragment 6.4, instead, displays a playful negotiation of participants' respective (gendered) identities, etc. Through these fragments, bystanders did maintain the robot as a competent participant; but the practical concern of these participants was not necessarily to maintain the "social competence" of the robot for itself – independently from an overarching activity.

In sum, topicalizations of the robot's actions were actions themselves. They did not produce a neutral description, account, judgment, explanation, etc. of the situation, removed from the activity at hand, but were, rather, embedded within conversation (Heritage & Watson, 1980). As Enfield & Sidnell (2017) remark, "what we say (vocally and otherwise) and what we do by saying it are two different, yet related things" (Enfield & Sidnell, 2017). Even when participants produced descriptions of the robot's actions, these descriptions participated in constructing the reality they described, i.e., they were "constituent features of the settings they ma[d]e observable" (Garfinkel, 1967).

6.4.3. The robot as a tool to advance local human activities

6.4.3.1 The “toy status” of the robot

Analyzing participants’ practices as *maintaining the robot as a competent participant... to achieve other activities* suggests that the robot was given a “toy status” (Goffman, 1979b). This concept developed by Goffman (1979b)¹²⁷ is notably used by Roberts (2004) to characterize situations where an animal is “spoken for” in order to deal with interactional issues taking place with co-present human participants. That is, situations where “talk directed at animals is likely in the realm of that which is designed for an overhearing audience” (Roberts, 2004). Through this practice, participants described by Roberts (2004) achieved a variety of actions: “utilizing the pet’s copresence, staff critique client caretaking of the pet, deflate client complaints, palliate client concerns, and maintain their professional stance” (Roberts, 2004).

6.4.3.2 Similarities with “spoken through” animals

Tannen (2004) describes similar situations where participants “speak through” a non-verbal third party like a dog. They do so, for example, by “ventriloquizing” a dog, by producing an account for its growling, by “speaking for another” (Schiffrin, 1993), etc. Again, these practices are subordinated to a variety of “interactive goals” (Tannen, 2004) that family members accomplish with their dog: “occasioning a switch out of an argument frame; rekeying the interaction as humorous; buffering criticism; reinforcing solidarity among family members; delivering praise; teaching values to a child; providing the occasion to talk as a way of enacting affection for pets; re-enforcing a couple’s bond by positioning them as “Mommy” and “Daddy” to their dog; resolving a conflict by conveying and triggering an apology; framing pets as family members; and reinforcing bonds among individuals who live together by exhibiting, reinforcing, and creating their identity as a family” (Tannen, 2004). Some items of this inventory match members’ problems observed in this chapter; problems which were dealt with by producing accounts of the robot’s actions and by speaking “for” the robot. In this sense, the practices of bystanders (described previously) share at least some similarities in their form and in their interactional goals (Tannen, 2004) with those identified in interspecies interactions. Although the extent of these similarities cannot be established without a systematic comparison, their existence begs the question of whether they result from the attribution of common interactional rights and obligations by human participants to robots and pets, and whether they configure the same participation roles.

In sum, in the public setting in which these encounters occurred, the robot often emerged as a competent participant as a means to progress other activities that took place among human participants. Through the topicalizations that participants made of its conduct, the robot was produced as “social” by humans (at least partly) as a resource, or a “toy” (Goffman, 1979b; Roberts, 2004), to accomplish various local “interactive goals” (Tannen, 2004). We argue that this “toy status” of the robot sheds light on the fast-paced changes in the robot’s status mentioned in section 6.4.1.2 of this chapter: i.e., the emergence of side-

¹²⁷ “Next is toy status, namely, the existence of some object, human or not, that is treated as if in frame, an object to address acts to or remarks about, but out of frame (disattendable) in regard to its capacity to hear and talk. Note, this status may be relatively fixed, as with an infant, or momentary, as when a husband comments in passing about his wife as though she were not present even though she is.” (Goffman, 1979b)

sequences between the robot's interlocutor-so-far and other humans, from which the robot was suddenly excluded as a participant. That is, the robot's status as a full-fledged participant was easily dropped by participants when the ongoing interaction was collectively repaired, assessed, judged, laughed about, etc. between human participants (see section 6.4.1.2) or, significantly, at instants where it did not contribute to the progression of a specific activity or goal (see section 6.4.2). We suggest that, on a wider scale, the concept of "toy status" may also appropriately gloss "what is going on" during commonly observed brutal episodes of the transience of the status of a robot: e.g., typical practices in human-robot interactions such as closings "without warning" identified by Licoppe & Rollet (2020), during which "the human participant [...] just stops interacting and moves away from the robot" (Licoppe & Rollet, 2020) even though a response from the participant may be sequentially relevant at this moment (Licoppe & Rollet, 2020). In these cases, as in those observed in our corpus, the "sociality"¹²⁸ of the robot may be, mainly, an expedient tool – built for other means than itself and discarded after use.

6.4.4. Were bystanders doing "not-doing repair"?

6.4.4.1 Doing "not-doing repair"

A problem is implicit in the previous considerations: if topicalizations of the robot accomplished various interactional goals, were they, in particular, doing "not-doing repair" (Pilnick et al., 2021)? That is, through their topicalizations of the robot's conduct as relevant, were bystanders intentionally avoiding repair (Stommel et al., 2022) when the relevance of the robot's contribution was not "immediately discernible" (Pilnick et al., 2021) even to them as bystanders – and not just to the main speaker? This is, at first glance, a reasonable hypothesis for many fragments of our corpus. The practice of doing "not-doing repair" was labeled by Pilnick et al. (2021) in their analysis of the responses of hospital staff to "hard-to-interpret-talk" (Pilnick et al., 2021) produced by people living with dementia. Healthcare professionals observably treated such hard-to-interpret talk as "related to the task at hand" (Pilnick et al., 2021) and, through this, avoided "challenging or drawing attention to the interactional competency" (Pilnick et al., 2021) of the person living with dementia. In spite of radical differences between the overall treatment of a robot and of a human¹²⁹, this strategical move may respond to the same practical concern: how to progress the interaction while publicly maintaining a co-participant as competent in spite of recurring sequentially troublesome behaviors on their part (i.e., when the "relevance of their contributions is not immediately discernible" – Pilnick et al., 2021 – to the participants interacting with them).

However, the practice of "doing not-doing repair" belongs to a category of phenomenon whose presence is difficult to demonstrate within a strip of talk. As Pilnick et al. (2021) note, "this leads to a potential methodological problem for CA researchers: that of analysing something in its absence rather than its presence". Relying on Jefferson (2017), she argues that some occurrences can be rigorously described as "observably relevant errors" (Pilnick et

¹²⁸ To use a different theoretical background, the same considerations apply to the protection of the robot's symbolical "face" (Goffman, 1967).

¹²⁹ On the issue of comparing human-human interactions with human-machine interactions, see section 8.2 of this manuscript.

al., 2021) or “observably relevant breakdowns in shared understanding” (Pilnick et al., 2021): that is, situations where it is observably demonstrable that “some or all participants are aware of the error” (Pilnick et al., 2021).

6.4.4.2 *A non-demonstrable phenomenon in our data*

In our corpus, we find only very rare cases where bystanders can be reasonably argued to be publicly displaying an “awareness of the error” made by the robot; that is, without “postulating ‘invisible’ intentions that are causing observable behavior” (Pelikan et al., 2022). Fragment 6.1 belongs to these rare cases: it provides some level of proof that, even though OLI – the bystander – verbally reacts to the robot’s conduct as displaying an intentional refusal to act, he simultaneously orients towards this non-response of the robot as the result of a (technical) issue of hearing. That is, he pushes the main speaker closer towards the robot, and asks her to lean towards it, changing the location from where is uttered the speaking turn for the sensors of the robot. This dissonance reinforces an interpretation where OLI avoids repairing or exposing as a repairable what he simultaneously treats (in his embodied behavior and in the advice he provides to ISA) as a technical issue on the robot’s end. However, even this argumentation on fragment 6.1 can be subject to criticism, as it relies on what OLI “gives off” (Goffman, 1959) and not on the way OLI’s conduct is treated by other human co-participants as constituting a specific type of action towards the robot. Although this is not fatal, the core of the argumentation unavoidably rests on the analyst’s judgment and not on a next-turn proof procedure.

Moreover, fragment 6.1 is an exceptional case. The same argumentation cannot be held for the three other fragments we provided and for most of the interactions in our corpus. They do not provide observable proof that, in parallel with the topicalizations they address to the main speaker, bystanders oriented towards the robot’s conduct as non-relevant, mistaken or absurd. That is, based on the analysis of our video recordings alone, we can talk about “topicalizations as actions” – using Wooffitt’s (1992) formulation – but we cannot assert, in most cases, that bystanders were doing “not-doing repair”. On the one hand, it was observable that bystanders were (publicly) accomplishing overarching activities by treating the robot’s contribution as relevant to the activity at hand. Similarly, when they occurred, main speakers’ embodied displays of trouble with the robot’s conduct were, by definition, also visually and hearably “displayed” to an analyst on the video recordings. On the other hand, in our corpus, it is generally not demonstrable (i.e., it is an unverifiable assumption about their cognitive state) that bystanders were purposely aiming not to repair the interaction even though the robot’s contributions were, in fact, also devoid of social meaning for them.

In other words, using the argumentative tools of an EMCA approach, we cannot state whether the social meaning bystanders publicly attributed to the robot was a direct reflection of their subjective understanding of the robot’s conduct – or whether what they publicly displayed was completely disconnected from a (hypothetical) cognitive representation of the robot’s conduct as incongruous. In the overwhelming majority of our corpus, that bystanders participated in framing the robot’s conduct as relevant (i.e., the observable impact of their contributions on the course of the talk) does not entail that they were subjectively dedicating their conduct to “hiding” what they, themselves, perceived as a failure from the robot. Hence, “playing the robot’s advocate”, the phenomenon we attempt to specify in this chapter, refers

to the regularly observable result of bystanders' contributions¹³⁰ in maintaining the robot's conduct as situationally relevant. It aims to highlight *how the robot's sociality emerges from a wide variety of actions from bystanders, responding to various practical problems* encountered by these participants in situ. By labeling some bystanders' conducts as "playing the robot's advocate", we do not imply that these practices are *intentionally "saving" the robot from its own interactional missteps while these missteps are perceived as missteps by bystanders themselves*.

6.5. The inner workings of the robot's "sociality"

6.5.1. The interactional "good will" of participants

The previous observations provide empirical material to classic micro-sociological questions: How do participation statuses emerge and evolve? How do participants rely on local features of an interaction to progress it? Do participants make sense of their mutual conducts in an indexical way, i.e., as related to a spatial configuration, a sequential context...? In particular, congruently with the EMCA literature on the topic, these fragments exemplify the robot's status as a moment-to-moment practical accomplishment (Pollner, 1974). This statement, although non-trivial, is both an empirical finding and a methodological axiom of EMCA – it is what we expect to find by using a methodology forged to highlighting moment-to-moment interactional achievements.

However, a broader conclusion can be attempted based on these micro-analytic observations. The participation shifts or "side sequences" highlighted previously – as well as the findings of, e.g., Due (2019), Krummheuer (2015b) and Pelikan et al. (2022) – contribute to specify the inner workings of a recurring phenomenon in human-robot situations: the work required to publicly make sense of what the robot is doing *stems overwhelmingly from human participants, while the robot itself seldom contributes to this task*. That is, the robot rarely clarifies its own behavior in response to visible trouble, it does not produce self-initiated self-repairs, etc. By topicalizing a conduct of the robot which was previously left unspoken, bystanders turned it into an action of a certain type (Sidnell, 2017) and made it accountable as such. These topicalizations of the robot's conduct publicly enforced the "social actions" they appeared to be merely describing, judging, explaining, noticing, etc. That is, the participation shifts we studied are one of many practices through which human participants display a documented tendency (or a recurring "good will" – Pelikan et al., 2022) "to treat the robot's contributions as relevant" (Pelikan et al., 2022).

This tendency (of which the fragments we presented in this work are individual occurrences) can be summarized in a rudimentary manner: humans spoke much more than the robot, produced more accounts and more descriptions of what the robot (or another participant) was doing, observably treated and responded to what was being done as relevant much more often than the robot. There was a striking contrast between the limited interactional contributions of the robot and the alert and fast-paced monitoring of humans to detect

¹³⁰ Contributions which, for the sake of brevity, we artificially limit to one specific type (the "topicalizations of the robot's conduct") in this work.

breakdowns in the interaction and resolve them while maintaining the robot's behavior as situationally relevant. In sum, the "burden" (Brown et al., 2023) of making sense of "what is going on" in a situation, of publicly displaying this interpretation, and of working to progress the interaction overall (Due & Lüchow, 2022) rested essentially with human participants.

The previous data show tentative similarities with observations from Fishman (1978) about the "division of labor" in talk where the "interactional work" (Fishman, 1978) of women was obscured and naturalized, i.e., "not seen as what women do, but as part of what they are" (Fishman, 1978). As West & Zimmerman, (1987) summarize, "women had to ask more questions, fill more silences, and use more attention-getting beginnings". Interestingly, West & Zimmerman, (1987) conclude that "it is precisely such labor that helps to constitute the essential nature of women as women in interactional contexts" (West & Zimmerman, 1987). We argue that, although the concrete practices through which it is realized empirically might be widely different, the existence of an obscured interactional work – unequally distributed between participants – might be common to both settings. To paraphrase West & Zimmerman, (1987), we argue that, in the previous public human-robot interactions, such (obscured and overwhelmingly undertaken by one party) interactional labor "helps to define the essential nature" of human interactants as humans in human-robot interactions. In this interpretation, the previous fragments display a form of typification of "being a human interactant" with a robot where the "work" of making the robot work (monitoring the robot's missteps, describing its conduct as relevant to the activity at hand, accomplishing most repairs, etc.) is displayed and oriented to as natural conduct: it is just what humans do in human-robot interactions.

6.5.2. The burden of interactional work

As a consequence, these fragments constitute local occurrences of the more general "burden" that weighs on humans when interacting with machines in the social world, identified by, e.g., Brown et al. (2023) and Greiffenhagen et al. (2023). Participants' conduct in the previous fragments is part of an overarching work to make technology "work" (Greiffenhagen et al., 2023). Significantly, in a museum setting, this "work" was not only produced by the robot's co-present and officially appointed "caretakers" (experimenters, demonstrators, care workers; Chevallier, 2023)¹³¹. Accounting for the robot's conduct (i.e., the "work" of making the robot's conduct both recognizable and relevant; Robinson, 2016) was partially supported by human co-participants themselves – even by those who were not, in a different vocabulary, the "end-users" of the dyadic interactions the robot was designed for. Human participants, main speaker and bystanders, worked with the behaviors displayed by the robot to progress the talk and to build activities around it.

On this point, these observations are congruent with several studies that hinted at the burden weighing on human co-interactants in human-robot or voice interface interactions (Auer et al., 2020; Due & Lüchow, 2022; J. E. Fischer et al., 2019; Pelikan et al., 2022). As Due & Lüchow (2022) note, it is "the user who does the social interactional work". Whether it is to repair miscommunications (Stommel et al., 2022), to maintain the robot as an apparent autonomous agent (Auer et al., 2020; Pelikan et al., 2022), to manage interactional trouble (Due & Lüchow, 2022; Porcheron et al., 2018), to design turns adapted to the interlocutor's

¹³¹ For example, see Chevallier (2023) for a description of the "pervasive micro-practices of interaction framing in situ" (Chevallier, 2023) produced by care workers or demonstrators to make care robots "sociable".

“perceptive abilities” (Pelikan & Broth, 2016) and, more generally, to “produce actions that ensure progressivity” (Due & Lüchow, 2022), the effort is overwhelmingly on human participants.

6.5.3. Respecifying the “sociality” of a robot

Attending to the moment-to-moment production of the robot’s status as a practical accomplishment (Rollet et al., 2017) – as done by the micro-analytic studies cited over the course of this paper – leads to a respecification (Garfinkel, 1991) of the lay use of the concept of “social robots”. The mostly positive encounters that we studied did not mechanically arise because humans faced a humanoid and intended-to-be-social robot for the first time which, e.g., produced a “wow effect”¹³². Such a summary would suppress the collaborative work produced by human co-participants to enforce the robot’s conduct as relevant. Instead, the robot was observably maintained as a “social robot” *in spite* of the recurring hard-to-interpret talk (Pilnick et al., 2021) it produced for the main speaker.

Ex ante definitions of a robot – or any technology – as “social” (i.e., before any interaction occurred) therefore run the risk of naturalizing, as stable¹³³ (Pollner, 1974) and self-evident, the observable result from micro-sociological, and always fallible, processes. In the case of the humanoid robot used in the fragments above, a definition of the robot’s “sociality”¹³⁴ as the unmediated consequence of its technical abilities (being able to speak, to move its arms, to detect and to respond to humans, etc.) – rather than as a locally emerging property of the interaction (Licoppe & Rollet, 2020; Rasmussen, 2019)¹³⁵ – obscures the micro-interactive work that co-present humans achieved to enforce the robot’s behaviors as “social actions”: that is, as actions responsive to the situated interaction and as “making relevant a set of potential next actions” (Tuncer et al., 2022b). As we attempted to show, used in their lay sense, notions of “social” or “conversational” robots “are merely ‘glosses’ for the processes that constitute them” (Hoeppe, 2023).

¹³²Applied to robots, the “wow effect” has been used as a lay concept to describe the observable tendency of humans to have positive first encounters with a robot (e.g., Etemad-Sajadi et al., 2022; Fuentes-Moraleda et al., 2020; Mubin et al., 2010).

¹³³ “Where others might see ‘things’, ‘givens’ or ‘facts of life’, the ethnomethodologist sees (or attempts to see) process: the process through which the perceivedly stable features of socially organized environments are continually created and sustained” (Pollner, 1974).

¹³⁴ See also Jackson & Williams (2021) for a discussion on the related concept of “social agency”, or Henschel et al. (2021) and Sarrica et al. (2019) for discussions of different definitions of “social robots”.

¹³⁵ The previous analysis provides some empirical content to the emergentist definition of the “sociality” of a robot which we outlined in chapter 3. In spite of its intrinsic properties (a humanoid shape, the ability to gaze at humans, to produce responses to their verbal turns, etc.), its status as a “social agent” only came to be (when it did) because it was treated as such throughout an interaction with other participants. We attempted to specify some of the concrete practices (among many others) which correspond to the local treatment of a robot as “social”.

7. “DISPLAYS OF HEARING” AND SOCIAL AGENCY: THE RELEVANCE OF AUTOMATIC SPEECH RECOGNITION TRANSCRIPTS IN INTERPRETING A ROBOT’S CONDUCT

Someone questions me, and waits for my response for a certain time, which they deem sufficient. If I have not answered after this time, they doubt either of my knowledge or of my sincerity or of my intelligence. – I may: either not have understood, or not have been able to, or not have wanted to. They never consider that I may have been thinking about something else.

—Paul Valéry, *Mauvaises Pensées et autres* (translation ours), 1942

Me when I can’t hear someone: Laugh and hope it wasn’t a question.

—Internet meme

7.1. Introduction

A basic feature of daily human-human interactions is that what is “inside people’s heads” is not a publicly available resource for participants involved in a local situation (Deppermann, 2018; Kristiansen & Rasmussen, 2021). As humans “don’t carry MRI machines with them out in the world” (Kerrison, 2018), how a participant hears, understands, judges, etc. its interlocutors’ actions is not directly accessible information for these interlocutors: there exists no “internalist” window allowing humans involved in a conversation to immediately witness other participants’ cognitive processes in real-time. Most of the time, interactants can only rely on other participants’ responses (verbal and embodied, including facial expressions) to identify if and how their own prior action was understood (as a question about X, as a request, as a greeting, etc. – vom Lehn, 2019), or if their words or gestures were even perceived by their interlocutor. Whether for the involved actor or the researcher studying video data, one must always rely on “inferential procedures” (Deppermann, 2012) to establish relationships between “discourse and cognition” (Deppermann, 2012)¹³⁶.

In particular, as a matter of fact, human recipients do not display on their foreheads the exact words they hear (and potentially mishear) during other humans’ speaking turns. A somewhat obvious consequence of this state of affairs is that, when, from an internalist point of view, an interlocutor completely mishears another participant’s speaking turn, this is not an

¹³⁶ Additionally, this relationship between discourse and cognition is not necessarily a practical concern for participants during their interactions (Albert & Ruiter, 2018).

accountable phenomenon of miscommunication in itself: the only available resource for a co-present participant to detect and repair a potential miscommunication is this interlocutor's embodied or verbal *response* to the previous turn (Mondada, 2011). If, without having heard precisely what was said, this interlocutor produces a second pair part of the type and form normatively expected (Kendrick et al., 2020) after the previous turn produced by a co-participant (for example, by responding with a greeting to a greeting), this response will be treated, momentarily, as adequate to the action produced by the previous participant. By itself (unmediated through gestures, speech, or facial expressions), the cognitive state of a co-participant is not a pragmatically consequential phenomenon.

Yet, this specific informational ecology of human-human interactions appears not to translate entirely to a substantial part of human-agent interactions: those where the agent provides a written trace of what it heard the human say. For example, on the commercial humanoid robot Pepper, the nominal behavior is to display on its tablet (attached to its torso) a transcript of what human interlocutors are saying, as heard by the robot. The top of this robot's belly screen features a "speechbar" where the result of the automatic speech recognition will be written, once the robot hears no more speech during more than 200ms (see Figure 7.1).

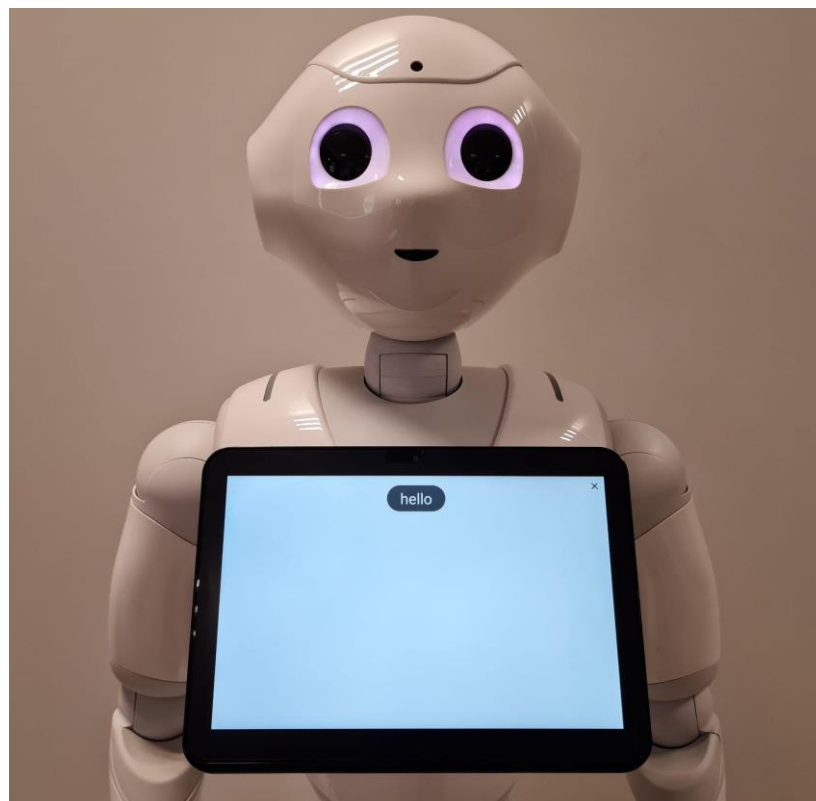


Figure 7.1. Default speechbar setting on a recent Pepper (that is, running its latest operating system, NAOqi 2.9). The robot displays "hello" after a human pronounced this word and stopped speaking for more than 200ms.

This speechbar includes 1) the transcription of the automatic speech recognition, 2) a waveform that moves when the robot hears someone speaking and 3) a blue bar that appears when the robot is listening. Aldebaran's official documentation explicitly states that these features are intended as resources towards which participants can orient to get a better grasp

of the situation when interacting with Pepper: “Providing feedbacks is essential to be sure that Humans really understand what’s going on when talking with Pepper. A bar, called SpeechBar, is displayed on the tablet. It helps to understand if the robot is listening, hearing something and what has been understood”¹³⁷. Similar transcriptions of the “Automatic Speech Recognition” (ASR) can be found, e.g., on smartphone vocal assistants (Siri, Google Assistant, Bixby, etc.). However, these transcriptions are rare among commercial “social” robots.

Significantly, at first glance, publicly displaying the result of the “automatic speech recognition” appears to open a window on what is going on “inside the robot”, before this robot produces any verbal or embodied response (speech, sound effects, gestures, LEDs, etc.) to the previous turn of an interlocutor. For example, a robot that was just greeted with “hello” will display “hello” on the tablet before it starts its return greeting action. In this view, human participants have direct access to the exact receipt of the words they have pronounced¹³⁸ before any return action from the robot can be achieved. When treated as such, i.e., as intended by its designers, an ASR transcript therefore reconfigures the informational ecology of the interaction: it becomes possible for a participant to pinpoint if the upcoming action of the robot stems from a correct receipt of the words of the previous turn.

We argue that the peculiar informational ecology outlined above is consequential on the way miscommunications are detected and dealt with – compared with other commercial robots not relying on a transcription, or even with human-human interactions. That is, if, in human-human interactions, each action “tests the hypothesis a participant has about a co-participant’s response to her/his action” (vom Lehn, 2019), we suggest these “experiments in miniature” (vom Lehn, 2019) take a different form in human-robot interactions where the robot features a transcript of what is said.

To investigate the pragmatic impact of this feature in situated interactions, we base our analysis on our video corpus recorded in July 2022 in Paris, at the Cité des Sciences et de l’Industrie, one of the biggest science museums in Europe. The first half of this corpus features 100 naturally occurring interactions with a Pepper robot placed in the hall of this museum. Chapter 6 extensively analyzed several fragments of this corpus to specify some (a priori different) interactional phenomena. However, the current chapter will also heavily rely on *the second half of this corpus*, composed of 108 additional interactions that occurred with Pepper in a laboratory open to the public at the Cité des Sciences et de l’Industrie, where participants were asked to wear an eye tracking device. In both cases, participants (groups or single individuals) interacted with the same Pepper robot, on which was running a chatbot designed to “converse” on a wide variety of topics. Significantly, the tablet attached to Pepper’s torso was entirely blank, except for the speech recognition transcripts that appeared on it. See section 7.3 for a detailed description of both halves of this corpus, the robot’s conversational design, and the conditions (“Transcript” or “No Transcript”) of this experiment.

We start by prefacing our qualitative findings with an interesting pattern visible in the data obtained from our eye tracking device: as the interaction unfolded, the attention of

¹³⁷ Retrieved January 29, 2024, from https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/conversation/conversation_feedbacks.html#conversation-feedbacks.

¹³⁸ Obviously, these transcripts do not make available to interlocutors how the robot “understood” the actions they performed or, in a less anthropomorphic vocabulary, in which action type it binned their previous action (Enfield & Sidnell, 2017).

humans (materialized by their gaze) slowly focused more and more on the robot's tablet, while its head and gestures were gazed at less and less. Once this quantitative picture of participants' gaze attention has been established, we use an ethnomethodological and conversation analytic approach *to identify the local interactional phenomena aggregated in these eye-tracking data*: that is, how this significant focus of participants' attention on the tablet manifested in concrete, situated, multimodal interactions.

In order to do so, we describe several typical ways in which the tablet was used as a resource by participants in situations of miscommunication, and, more specifically, when the robot misheard the previous turn of a human. These troublesome exchanges appear more likely to reveal participants' orientation towards the ASR transcript – compared to perfectly smooth interactions¹³⁹. First, we focus on the way the display screen may constitute a resource to indicate and manage trouble. In particular, we focus on the specific timing and packaging of repair sequences which are observably indexed to the troublesome speech recognition transcript. We suggest that these repairs are exclusive to human-robot interactions featuring an ASR transcript, i.e., they are a *sui generis* phenomenon that is not observed as such in human-human interactions. Then, we investigate recurring cases where the “erroneous” transcript is treated as a more-or-less autonomous resource: i.e., as disconnected from the embodied robot's vocal and gestural responses, or even as taking priority over them. We focus on the way troublesome displays may become resources to distinctive interactional performances and we try to highlight a typical format through which the incorrect speech transcription is reported to other human co-participants by the principal speaker.

7.2. Previous work

To the best of our knowledge, in the fields of human-computer and human-robot interaction, no study has been yet undertaken about the impact of speech transcripts on the interaction: whether about its pragmatic consequences on an interaction, or about its impact on the perception of a device – robot, phone, computer, etc. Similarly, we could not find existing works applying this line of inquiry to smartphone vocal assistants (Siri, Google Assistant, Bixby, etc.) – which also display transcripts of users' utterances.

Published works whose area of investigation is the closest to this topic can be found about the use of real-time transcripts in human-human interactions (Echenique et al., 2014; Gao et al., 2014; Yao et al., 2011a) or in listening comprehension tasks (Cao et al., 2016a, 2016b; Yao et al., 2011b). In particular, Gao et al. (2014) studied the impact of real-time transcripts in conversations between native speakers and non-native speakers. In one condition, only the non-native speakers had access to the transcripts, while in the other condition both the native and non-native speakers could read them. Based on several indicators, they found that communication was improved when native speakers had access to the transcripts of what they had said: the clarity of native speakers' speech was higher when

¹³⁹ Because “it is hard to determine how something works when we only see it functioning unproblematically” (Stivers et al., 2023), studying “when things go wrong” is a typical strategy to uncover the “seen-but-unnoticed” (Garfinkel, 1967) methods through which humans accomplish their daily activities – ever since Garfinkel's breaching experiments (Garfinkel, 1967). Limited or complete breakdowns in an interaction provide opportunities for participants to publicly display and reconstruct their interpretation of the ongoing situation.

the transcripts were available to them, as well as the ratings of the quality of the conversation made by both the native speakers and the non-native speakers. Gao et al. (2014) suggest that having mutual access to the automated transcripts allowed participants to identify “when and where problems arose in the group communication” (Gao et al., 2014). Chen et al. (2018) and Echenique et al. (2014) compared the impact of “video+audio” condition with a “text transcripts+audio” condition on conversation grounding between native and non-native speakers in English. Significantly for our angle of study, Echenique et al. (2014) found that, compared to the “audio + video” condition, real-time text transcripts were “better at assisting [non-native speakers] in repairing common references”. Likewise, Chen et al. (2018) observed the positive impact of a live-transcript tool that automatically translated and transcribed speech in a communication between Japanese and Chinese speakers. Cao et al. (2016b) compared the problems faced by non-native speakers' in listening tasks within two conditions: “audio only” and “audio+ASR transcripts”. They report that correctly transcribed ASR text helped participants to solve only specific types of problems in their understanding of the audio information, yet, that flawed transcripts worsened the situation by creating additional issues. The authors used an eye-tracking device to establish the frequency and the moments at which the ASR transcripts were gazed at, as well as the exact location of the gaze fixation. They found that some participants constantly gazed at the transcript area, while others only occasionally or never looked at it.

However, those of the previous studies that focus on human-human real-time interactions necessarily rely on a “third party” transcription device: when a participant speaks, the transcript does not display what sentence their interlocutor has heard, but only the transcribed sentence this interlocutor will have access to – while also hearing the original utterance. The configuration of these interactions therefore does not completely overlap with the configuration of human-agent interactions featuring an ASR transcript. For example, in the case of the Pepper robot, the tablet attached to this robot displays the entirety of the speech-to-text information that the robot will use to build an answer. In other words, there is no additional information (prosody, tone, etc.) about the speakers' utterance that the robot would “hear” but not display: what is encoded on the tablet represents all of the audible information available for the robot to respond. As a consequence, we argue that, from the point of view of an observer removed from the local interactions we will study (an etic point of view)¹⁴⁰, the starting configuration of these interactions is *sui generis*: there is currently no equivalent informational distribution in human-human real-time interactions.

¹⁴⁰ That is, from the technical point of view of an external observer with an understanding of the inner workings of the robot. Indeed, as we will see, the ASR transcript can be treated as something else than a mere transcript by participants in situated interactions.

7.3. Method

7.3.1. Natural data collection

7.3.1.1 Setup

Our natural data collection took place at the Cité des Sciences, between the 19th and the 24th of July 2022. The Pepper robot for this collection was placed at the start of an exhibition named “Robots”, which introduced visitors both to the history of robotics and to the workings of state-of-the-art modern robots (either industrial or “social”). Remarkably, visitors had to go through this exhibition to reach the other ones (which were not about robots). This was useful on two levels. First, our Pepper robot was among the first content most visitors were exposed to, which limited the risks of participants being influenced by other exhibitions regarding our robots’ abilities. Second, since this robot was “on the way” to reach other exhibitions, visitors were not yet pre-selected by their intention to interact with robots – even though our participants were inevitably representative of the pool of individuals likely to go to a cultural center during the summer.

During this data collection, our Pepper robot faced the passageway. Two cameras were filming the robot: one on the left, one on the right. At 1.5 meters in front of the robot, a red line was drawn on the ground (see Figure 7.2) to materialize the zone that was being recorded. On both ends of this line, a sign informed visitors that this area was being audio and video recorded for a scientific experiment. This sign explained how, for how long, and by whom the storage of these data was carried out, confirmed that these data would be anonymized, and provided the contact information of the researcher in charge of the experiment if participants wished to exercise their right to access, modify or delete the images which concerned them. Finally, the sign informed participants that by entering this area, delimited by the red stripes on the floor, they consented to the processing of their personal data for the purpose described above and in accordance with the GDPR, as well as to the capture of their image for the purposes of this experiment¹⁴¹.

¹⁴¹ If an unaccompanied child entered the recorded area, data were immediately deleted.



Figure 7.2. Setup for the Pepper robot (loading its chatbot) at the start of the “Robots” exhibition, including an RGPD-compliant sign (left), a red stripe on the ground, and a camera.

7.3.1.2 Scenario

Visitors were not briefed by an experimenter before speaking with the Pepper robot. Only the two signs present on both sides of the setup informed participants that this was an experiment focused on “human-robot interactions”. Participants were given no indication regarding the robots’ abilities (speaking hearing, seeing, etc.), nor about the meaning of the text displayed on the tablet (i.e., the ASR transcript). Coherently, the signs around the robot did not provide participants with any ideas about specific “intended uses” of the robot, in order to maximize the emergence of spontaneous activities with the robot.

7.3.2. Experimental data collection

7.3.2.1 Setup

Our experimental data collection took place at the Cité des Sciences, between the 26th and the 31st of July 2022. The Pepper robot was placed in a closed sound-proof room (see Figure 7.14) which blocked participants from the view of other visitors. Before entering the room, visitors or their legal guardians were briefed, then asked for their written consent to participate in the experiment and, separately, for the use of the video data in which they would appear.

Participants were then asked to wear one pair of eye-tracking glasses (Tobii Pro Glasses 3), which were configured before they started speaking to the robot. If participants entered the room as a group, only one member would wear the eye-tracking device. Inside the room, two video cameras recorded participants' gestural and verbal conduct.

7.3.2.2 Participants

Video data were collected on 108 interactions (groups or single individuals). Among these interactions, 63 participants agreed to wear eye-tracking glasses. Passersby were recruited in front of the room, no matter if they were single individuals or groups. Two selection criteria were applied: being over 18 or being accompanied by a parent or legal guardian and being fluent in French.

7.3.2.3 Instructions

Participants were asked to “speak with the robot in the room for 5 minutes maximum, and 2 minutes minimum”. They were not given explanations regarding the conversational design of the robot, about the meaning of the ASR transcript displayed on the tablet, nor about the robot's abilities – besides its ability to speak, which was presupposed in the instructions. After they had spoken with the robot for 5 minutes, and unless they had already left, participants were interrupted by an experimenter and asked to leave the room.

7.3.3. Robot's Conversational Design

The robot's software and conversational design were the same in the natural and experimental data collections. The robot's behavior (speech and gestures) was handled by a simple rule-based chatbot, locally installed on its tablet. The robot did not generate its answers nor was it provided with the answers by an external API (e.g., from chatGPT). In other words, besides some responses in which the robot could include information about the local situation (e.g., “your name is [name previously given by the participant], today we are [date of today]”, etc.), every sentence that the robot uttered had been directly typed by a human in the chatbot software. However, the robot's responses covered a wide range of domains. Its chatbot could produce a variety of responses on 70 widely defined topics (see Figure 7.3): e.g., geography, sports, animals, personal information about the robot, which movement it could do, which songs it could sing, etc.

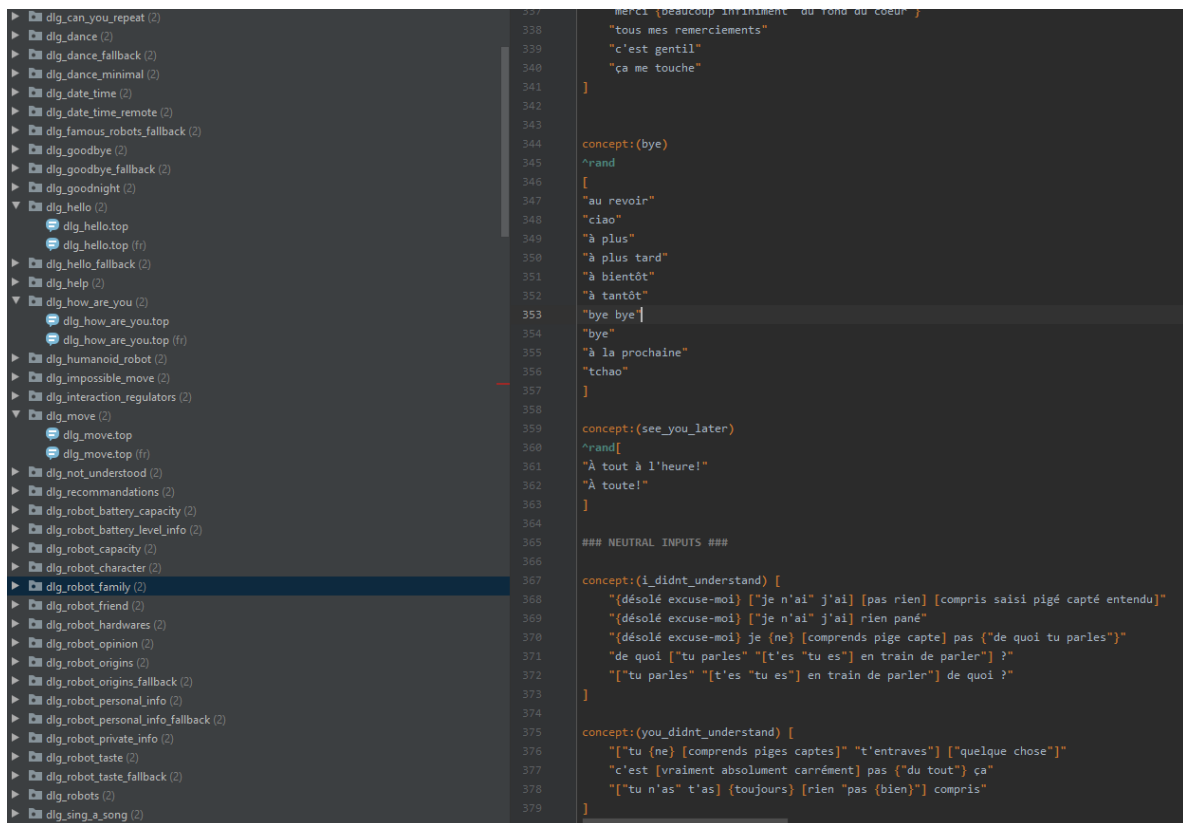


Figure 7.3. Screenshot of the content of the chatbot. On the left, some of the topics about which the robot could produce answers. On the right, some of the concepts it could recognize in humans' speech.

The main characteristics of the robot's conversational design were as follows:

- a. *The robot's speech could be interrupted at any moment.* If the robot was currently producing an utterance, but that it received a new vocal input that it managed to match with an answer, it would interrupt itself and start a new utterance responding to this new vocal input.
- b. *The robot did not produce "listening" sounds.* Earlier versions of Pepper and of the Nao robot featured sounds designed to indicate when the robot started listening (Pelikan & Broth, 2016) and when it stopped listening. This was not the case with the version of the Pepper robot we used: because of its ability to be interrupted, the robot was always listening.
- c. *The shoulder LEDs on the robot were blue when the robot was silent and were turned off when it was producing an utterance.* As the robot was always listening, this color code did not indicate a "listening state", as in previous versions of Pepper or on the Nao robot.
- d. *Except for the ASR transcript, nothing was displayed on the robot's tablet.* Recent Pepper robots normally display a "speechbar" indicating when the robot is listening or hearing sounds. This speechbar is a nominal behavior on these robots but was removed to limit the number of parameters to take in account during the analysis of

the eye-tracking data.

- e. *On the overwhelming majority of topics, the robot did not keep track of the previous turns produced by its interlocutors. As a consequence, our participants faced a conversational design whose limits were similar to many voice user interfaces (thereafter, VUI), for which “every directive needs to be recipient designed for the VUI with reference to immediately prior and/or ongoing VUI actions” (Due & Lüchow, 2022).*

7.3.4. Automatic Speech Recognition Transcript design

Upon hearing a human speak (or a human-like sound), the robot attempted to recognize which words were uttered. To do so, it used both a local Automatic Speech Recognition system and a remote one¹⁴². If any of these services returned a text input that matched one of the possible answers for the chatbot, the robot played this answer: it uttered a verbal response and, if some gestures were scripted in association with this response, it played them. In order to recognize the end of speaking turns, the robot only relied on silences: no grammatical, semantic, or prosodic criteria were used. Any utterance produced by a human followed by a silence of more than 200ms was considered complete, transcribed on the robot’s tablet (for 3 seconds), and responded to by the robot.

In brief, after they produced an utterance, participants encountered one of the following behaviors from the robot:

1. *When the robot was provided with a result from the ASR, and could match a response with this specific ASR result, then the robot displayed the ASR transcript on the tablet (see Figure 7.1) and triggered its verbal and gestural answer at the same time.*
2. *When the robot was provided with a result from the ASR but had no answer to this specific utterance, the robot displayed the ASR transcript but verbally indicated that it did not know what to respond*
3. *When the robot was not provided with a result from the ASR for the sound it received, the robot displayed a question mark “?” on the tablet (see Figure 7.4) and produced an open-class (Drew, 1997) repair initiator (“huh?”, “what?” “pardon?”, “excuse-me?”, etc.)¹⁴³.*

¹⁴² In our case, the remote ASR service was provided by the Automatic Speech Recognition service of Nuance: <https://www.nuance.com/>. Upon being sent the continuous flow of sound received by the robot, Nuance returned one or several possibilities (with associated confidence scores) as to what words composed this audio input.

¹⁴³ Documentation is available here: https://qisdsoftbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/conversation/conversation_feedbacks.html#conversation-feedbacks

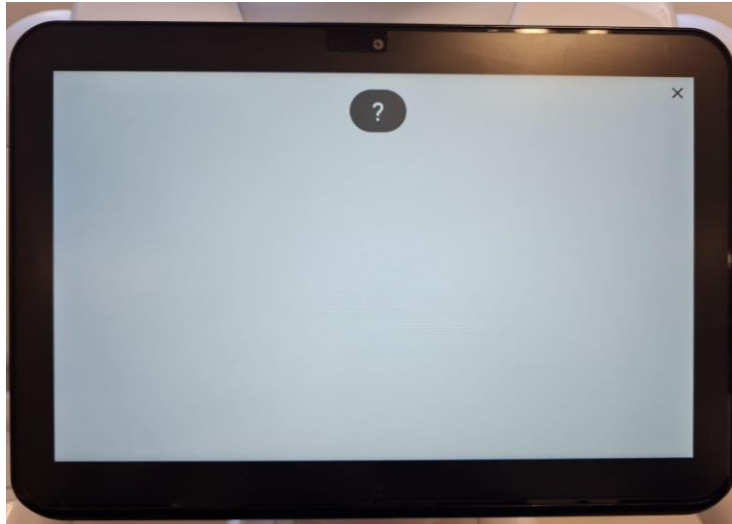


Figure 7.4. The robot received a strip of sound but could not identify any word within it.

In any case, except for the ASR transcripts, *the tablet was blank*. That is, the only information that the tablet displayed were the words returned by the ASR service. Note that these words were also the only information based on which the robot responded to the human. Through this ASR transcript, humans (whether they oriented to the ASR in this manner or not) had exhaustive access to the information from the outside world from which the robot produced its answer: no other information (gestures, tone, appearance of the interlocutor, etc.) was used by the robot.

7.4. ASR transcripts and informational configurations

7.4.1. Similar informational configurations *outside of social robotics*

Ethnomethodological conversation analysis is regularly described as “cognitively agnostic” (Hopper, 2005; Maynard, 2006) or, at least, as uninterested in “the mental processes which go on in the brain when understanding takes place” (Deppermann, 2012). This stance partially results from a basic property of human-human interactions: in typical face-to-face interactions, processes “residing inside the mind of an individual” are not immediately accessible or relevant to co-present participants or to an external researcher. These mental processes are not “a discursive phenomenon, which is publicly displayed and collaboratively oriented to by the parties to a conversation” (Deppermann, 2012)¹⁴⁴. The opposite is true in the intended design behind the ASR transcript on the Pepper robot, i.e., from the designer or engineer’s point of view. In this vision, when a robot writes what it hears (and is treated as doing so by co-present

¹⁴⁴ This is not to say that the presence of an ASR transcript deletes the need for any inferences in an interaction with a robot. However, when participants interpret the ASR transcript as a transcript, they do not need to rely on what is “given off” (Goffman, 1959) by the robot to infer what was heard by the robot – when this is a practical concern for them in an interaction. In other words, in this informational configuration, the phonological reception of participants’ utterances by the robot does not have to be inferred by participants that orient to the ASR transcript as intended by its designers.

interlocutors), the direct phonological receipt of what the robot heard becomes “observable-and-reportable” (Garfinkel, 1967) and, therefore, a potentially interactionally relevant property of the setting.

More or less similar information configurations are extremely rare. There exists no human-human face-to-face configuration based on a *scriptural resource* displaying precisely what words were heard by an interlocutor in real time. Not-too-dissimilar situations would be, for example, speaking to someone connected to a polygraph (i.e., a lie detector) or below an IRM machine. In both these situations, a written support makes available processes going on inside someone’s body or brain, which can be used by co-participants as resources to refine their interpretation of this person’s verbal and gestural conduct. The traces provided by the polygraph or the IRM machine can be considered by co-participants as interactionally relevant, and enrich their interpretation of “what is going on” (Beach & Sigman, 1995). Human-human interactions taking place in configurations like the aforementioned ones are, to this day, extremely rare.

7.4.2. Similar informational configurations *with* conversational agents

Vocal assistants like Siri, Google Assistant, Cortana, Alexa display public ASR transcripts when used on a mobile phone or on a tablet. Even though the way in which they are used might not be conversational (Due & Lüchow, 2022) nor oriented towards the same goals as what mattered situationally for our participants during their interaction with Pepper, all these assistants provide users with a written transcript of the utterance to which the device will be responding to. On the contrary, this informational configuration, to the best of our knowledge, does not exist among commercial “social” or “companion” robots – besides the Pepper robot. Even though screen displays are among the most popular interaction modalities in social robots (Mahdi et al., 2022), the design of these robots shows a clear tendency not to indicate by default what the robot hears.

7.5. Eye-tracking as a complement to an ethnomethodological and conversation analytic approach: some theoretical considerations

Few studies have currently discussed the potential of eye-tracking as an additional tool for EMCA to pursue one of its core goals of “gaining access to participants’ orientations and perspectives” (Kristiansen & Rasmussen, 2021). A major pitfall identified by Stukenbrock & Dao (2019) and discussed by Kristiansen & Rasmussen (2021) is the indeterminate public relevance, for participants involved in situated interactions, of gaze behavior. Accurate and quantified information about eye fixations, saccades, as well as the first-person perspective of video recordings achieved with eye-tracking glasses, are “specifically unavailable to co-participants and cannot for that very reason have any social significance for the participants in the interaction” (Kristiansen & Rasmussen, 2021). Less intricate gaze behaviors can also be perceptually available to co-participants during an interaction (e.g., someone looking in a specific direction, for a long period or not, etc.) without being responded to as socially meaningful, i.e., as publicly available and reportable phenomena. Hence, eye-tracking, when

used as part of a conversational analytic methodology, risks to encourage the analyst to provide *a priori* social relevance (Schegloff, 1993) to all gaze behaviors produced by participants¹⁴⁵. The analyst's description of these gaze behaviors (quantified or not) may not reflect co-present participants' observable orientation to these practices as relevant features of the situation, nor convey what was really perceived by these co-participants while they were immersed in the urgency of a situated interaction.

The previous considerations should not extend to our use of eye-tracking data as a *preliminary step to a qualitative analysis*. Indeed, we compartmentalized, on one hand, the quantitative overview based on the ASR transcript and, on the other hand, the detailed study of interactional phenomena. Eye-tracking was used as a preliminary tool, to get a grasp of whether or not participants focused their attention on the ASR transcript when it appeared. Once it was confirmed that – on average – they did, we focused on the social relevance of this transcript. In other terms, capturing participants' gaze aimed at confirming that the tablet was, indeed, seen by a majority of humans interacting with the Pepper robot. The regular focus of participants' attention on this part of the robot suggested that it had some sort of significance for them, before we applied qualitative methods to understand if the information provided by the ASR transcript was given a *social meaning*. In sum, rather than collecting eye-tracking data to bring “new analytic insights” (Kristiansen & Rasmussen, 2021) to an EMCA analysis, we argue that, on the contrary, an EMCA analysis is fit to bring new insights to purely quantitative eye-tracking results. That is, it is well-suited to “explain” these data by disclosing which interactional phenomena they aggregate.

Occasionally, eye-tracking was used as an additional failsafe against erroneous interpretations of some of the most intricate of our qualitative fragments. Our corpus contained rare fragments where a participant's conduct appeared to be responsive to the ASR transcript but where no deictic gesture nor verbalization allowed the analyst to demonstrate that this conduct was not, instead and exclusively, a reaction to the ongoing verbal turn of the robot. Most of the time, these fragments could not be used as it was impossible to argue rigorously about what was going on. Yet, in specific cases, the eye-tracking data made clear that a participant never gazed at the ASR transcript, which reinforced the interpretation that this participant's actions could hardly be understood (by the analyst) as a reaction to what was written on the tablet.

7.6. Gaze data analysis

7.6.1. Conditions

During the experimental data collection, participants (N=108) were randomly assigned to one of two conditions.

¹⁴⁵ Nevertheless, describing gaze direction, even when it is not used as a public resource, is already a preexisting practice in EMCA. This allows the researcher to describe a modification in the information to which a participant has access to (by looking somewhere), and to which this participant may respond, even if it is not made public yet. In this endeavor, as Stukenbrock & Dao (2019) note, eye-tracking allows the analyst to get a better grasp of someone's gaze direction, rather than inferring it from this person's head angle.

- In the No Transcript condition (see Figure 7.5), participants spoke with a robot that did not display anything on its tablet (it was completely blank at all times).
- In the ASR Transcript condition (see Figure 7.5), participants spoke with a robot that displayed an ASR transcript.

Besides the ASR transcript, the autonomous robot’s behavior was exactly the same in both conditions.

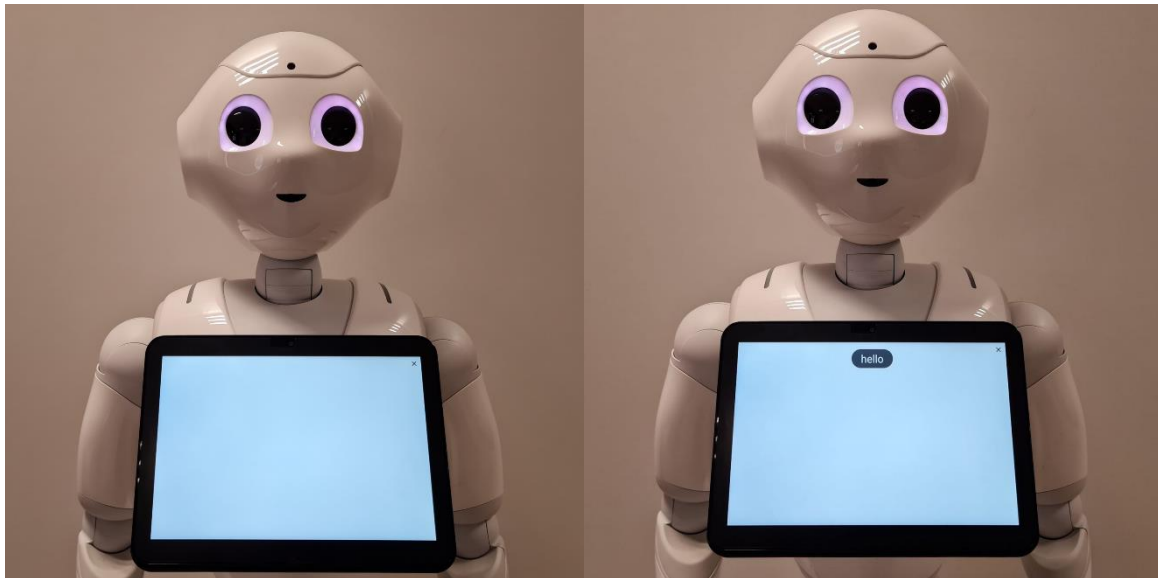


Figure 7.5. Pepper after hearing “hello” in the No Transcript condition (left) and in the ASR Transcript condition (right)

7.6.2. Data Preparation

7.6.2.1 Timecoding of utterances and mapping of fixations

Among participants who agreed to wear an eye-tracking device and whose gaze was captured accurately, we randomly selected 22 participants in each condition (44 overall). For each of these participants’ interactions, we timecoded every moment during which a speaking turn was produced by a human (see Figure 7.6).



Figure 7.6. Speaking turns produced by a participant coded as Times of Interest (TOI) on Tobii Pro Lab.

Meanwhile, a 2D view of the robot was created (see Figure 7.7). It was divided into several zones:

1. Head
2. Shoulders
3. Arms
4. Base
5. Top of the Tablet (ASR Transcript Zone)

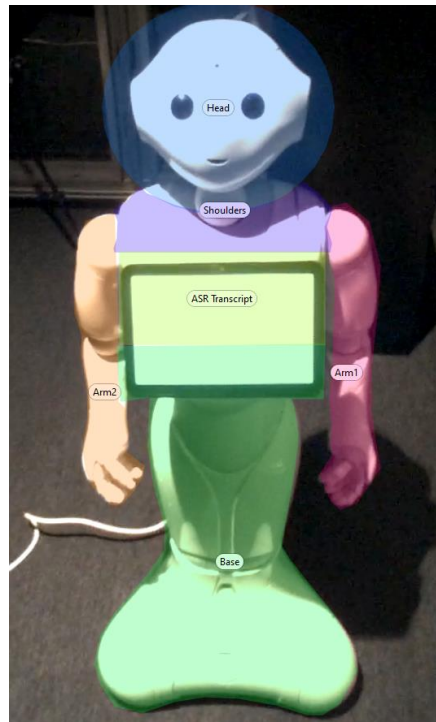


Figure 7.7. Areas of Interest (AOI) set on a 2D picture of the robot on Tobii Pro Lab

We combined zones 1, 2, 3, and 4 under the category “Body and Head”, to compare them with zone 5 “Top of the Tablet” (the zone where the ASR transcript appeared). Each gaze fixation from participants was then mapped onto this 2D view¹⁴⁶ (see Figure 7.6). This provided us with a distribution of participants’ fixations over the robot’s visible features.

7.6.2.2 Isolation of periods during which the ASR transcript was visible

In the ASR Transcript condition, the ASR transcript was displayed exclusively over a period of 3 seconds¹⁴⁷ after the robot heard a speaking turn. The tablet was blank the rest of the time. Hence, in the ASR Transcript condition, to focus on participants’ practices when they could effectively see an ASR Transcript in front of them (and not when the tablet was blank), we extracted participants’ gaze attention during this three-second post-utterance phase. To facilitate a meaningful comparison (see section 7.6.3), we then also extracted this post-

¹⁴⁶ This mapping was achieved manually most of the time. In rare cases, it was produced automatically when Tobii Pro Studio’s mapping tool could match the geometry of the 2D image with the geometry of participants’ view in the video recorded from their glasses. All automatic reportings were thoroughly checked and, when needed, corrected by a human. The Tobii I-VT Attention filter was used for this task.

¹⁴⁷ The tablet was entirely blank except after the robot heard a turn from a human, in which case an ASR transcript would be displayed for a maximum of 3 seconds. In some cases, the transcript would be displayed for 2 seconds only, depending on the length of the utterance recognized by the robot.

utterance phase for the No Transcript condition. In both conditions, this phase of 3 seconds is especially interesting for conversational-analytic concerns: it corresponds, first, to the display of the ASR transcript (in one condition) but, also, to the start of the robot's response to what was said (in both conditions). In other terms, *it is the moment where participants can produce an early interpretation of whether the robot properly heard and understood them, either using the ASR transcript or the robot's verbal and gestural response.*

An additional difficulty was that the number of turns produced by participants varied greatly depending on how long they spoke to the robot: some “conversed” with the robot for a long time, while others for only a few minutes. Hence, for each participant, we considered their gaze behavior during the 3 seconds that followed each of the *20 first speaking turns they addressed to the robot* (see Figure 7.8). In other words, in the ASR Transcript condition, we studied participants’ gaze behavior during the first 60 seconds (3 seconds post-utterance multiplied by their 20 first speaking turns) that they spent in front of a (mis)transcription of what they said. In the No Transcript condition, we studied the same post-utterance period – except that, in this condition, participants were not provided with a transcript.

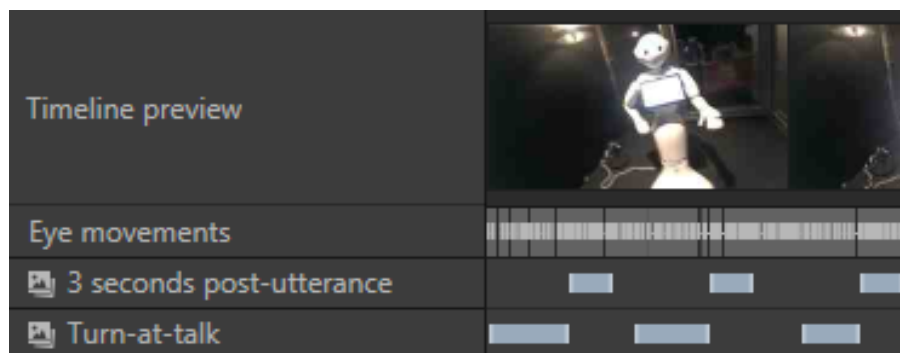


Figure 7.8. Example of timecoded utterances and post-utterance phases for a human participant

7.6.2.3 Binning eye-tracking data in fixed duration time intervals

After extracting the cumulated 60 seconds of “post-utterance” time for each participant, we binned¹⁴⁸ them in 50ms intervals bins (for a total of 1200 bins per participant). Then, for each 50ms bin, we checked how much of this time was spent fixating on each part of the robot. This allowed us to plot the evolution of participants’ average total fixation time (per bin) on the ASR transcript zone compared to the rest of the robot (see Figure 7.11).

¹⁴⁸ The R Studio script that we used for processing these binned data is based on Richard Andersson’s (Tobii’s Chief Product Owner) script. This script is available at: <https://github.com/richardandersson/TobiiProLabScripts/tree/master/plotting%20binned%20metrics>.

7.6.3. Post-utterance gaze behaviors between conditions

7.6.3.1 ASR Transcript condition

During the 3 seconds that followed their utterances, participants' gaze was mostly directed towards the tablet in the ASR Transcript condition (See Figure 7.9). On average, after 20 speaking turns, the average total fixation time per area of interest was:

- 27 seconds for the ASR transcript.
- 18 seconds for the robot's "Body and Head".

That is, 61% of the time participants spent fixating on the robot after they finished speaking was focused on the ASR transcript and not on the robot's gestures or face. The heatmap of participants' fixations similarly illustrates the high number of fixations on the ASR transcript during this condition (see Figure 7.10).

The gap between the average time that participants spent gazing at the robot's "Body and Head" and "ASR transcript" is confirmed by a paired samples t-test. There is a significant difference in mean fixation time between these two parts of the robot, $t(10) = 3.22$, $p = 0.004$. Cohen's d indicates a medium to large effect size (Cohen's $d = 0.70$). Data is normally distributed (Shapiro-Wilk Test $p = 0.5374$).

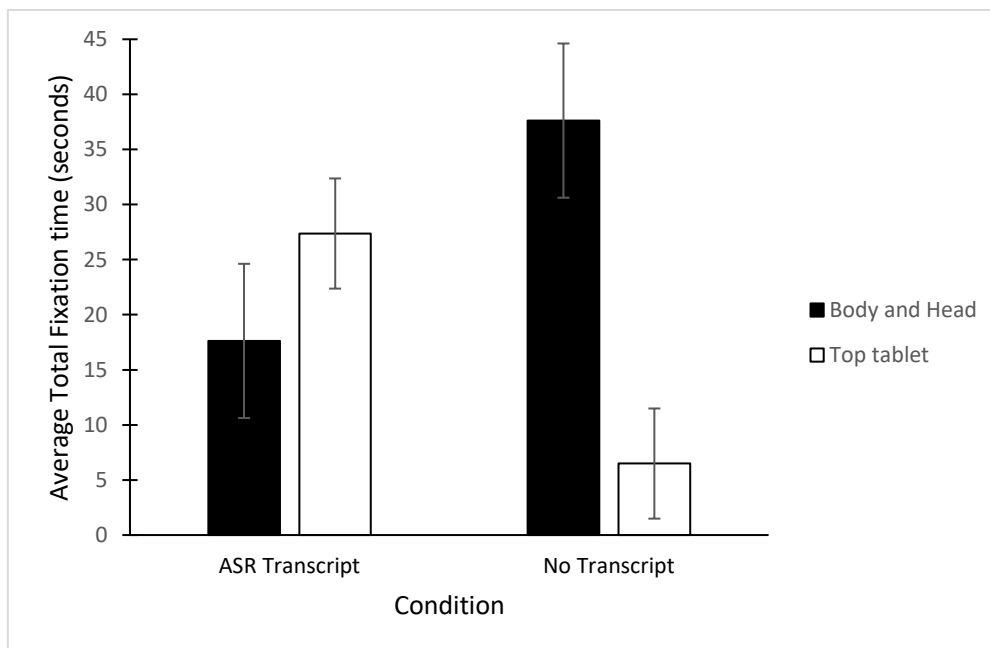


Figure 7.9. Mean total fixation time on areas of interest per condition (post-utterance)

7.6.3.2 No Transcript condition

Unsurprisingly, in the No Transcript condition, participants' gaze was mostly directed towards the robot's "Body and Head" during the 3 seconds that followed their 20 first utterances (see Figure 7.9) and barely towards the top of the tablet. On average, after 20 speaking turns, the total fixation time per area of interest was:

- 39 seconds for the robot's "Head & Body".
- 6 seconds for the top of the tablet (where the ASR transcript would be).

The difference between the average time that participants spent gazing at the robot's "Body and Head" and at the top of the tablet (the "ASR Transcript" zone) is confirmed by a paired samples t-test. There is a significant difference between the mean total time during which participants fixated on these two parts of the robot ($t = 19.0031$, $p < 0.0001$). Cohen's d indicates a large effect size ($d = 3.879$). Shapiro-Wilk Test suggests that the data is normally distributed ($p = 0.4367$).

This relative absence of focus on the tablet is also apparent in the heatmap of participants' fixations after they finished their speaking turn in the No Transcript condition (see Figure 7.10).

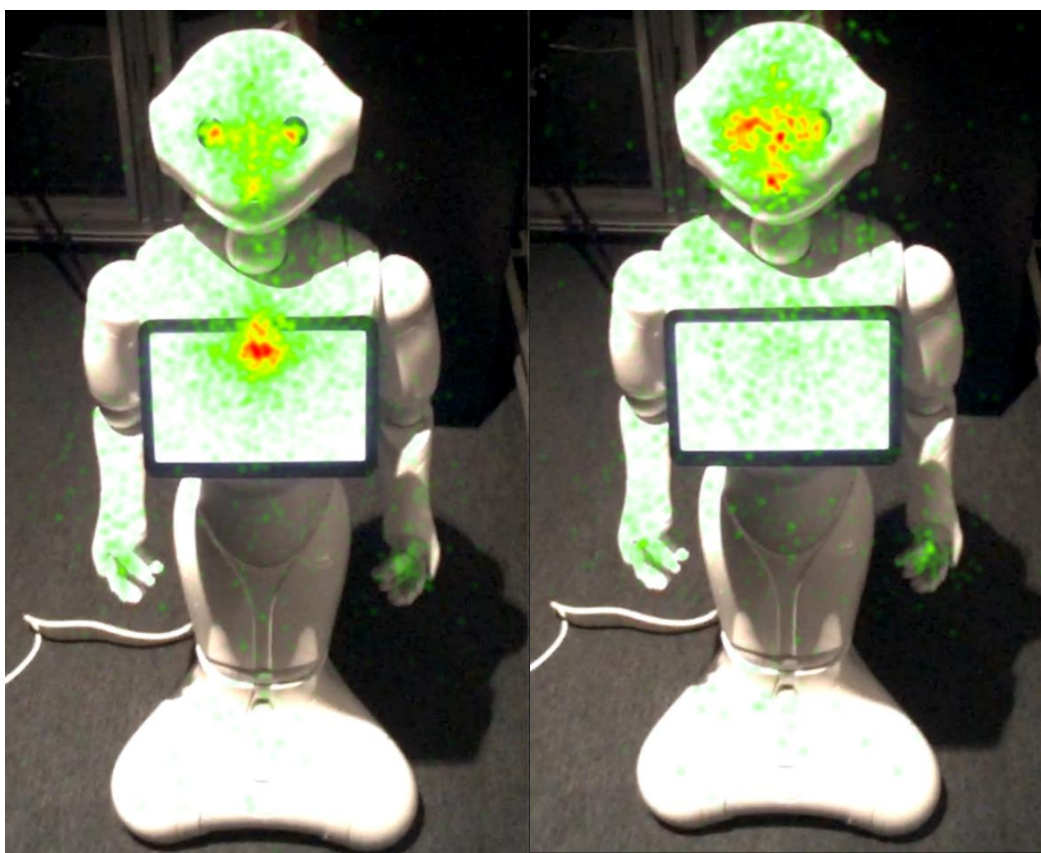


Figure 7.10. Heatmap of participants' fixations after they finished their speaking turn in the ASR Transcript condition (N=22), left, and in the No Transcript condition (N=22), right.

7.6.4. Evolution of post-utterance gaze fixations over the course of the interaction

Because the time participants spent looking at the top of the tablet was minimal in the No Transcript condition, we then focused exclusively on the evolution of participants' gaze attention during the ASR Transcript condition. That is, we sought to determine if participants

changed their gaze behavior over the course of an interaction where they faced a transcript of what they said.

We plotted the evolution of participants' average time spent fixating on the ASR transcript (versus the rest of the robot) over their first 60 seconds of "post-utterance" time (see section 7.6.2.3). The results suggest that, in the ASR transcript condition, participants' gaze attention *gradually focused on the ASR transcript when it was visible* (see Figure 7.11). Conversely, there was a decrease in their gaze attention towards the rest of the robot's body (head, hands, etc.) during this period. In other words, over the course of their first 20 turns-at-talk, participants seemed to increasingly fixate on the ASR transcript each time they finished speaking.

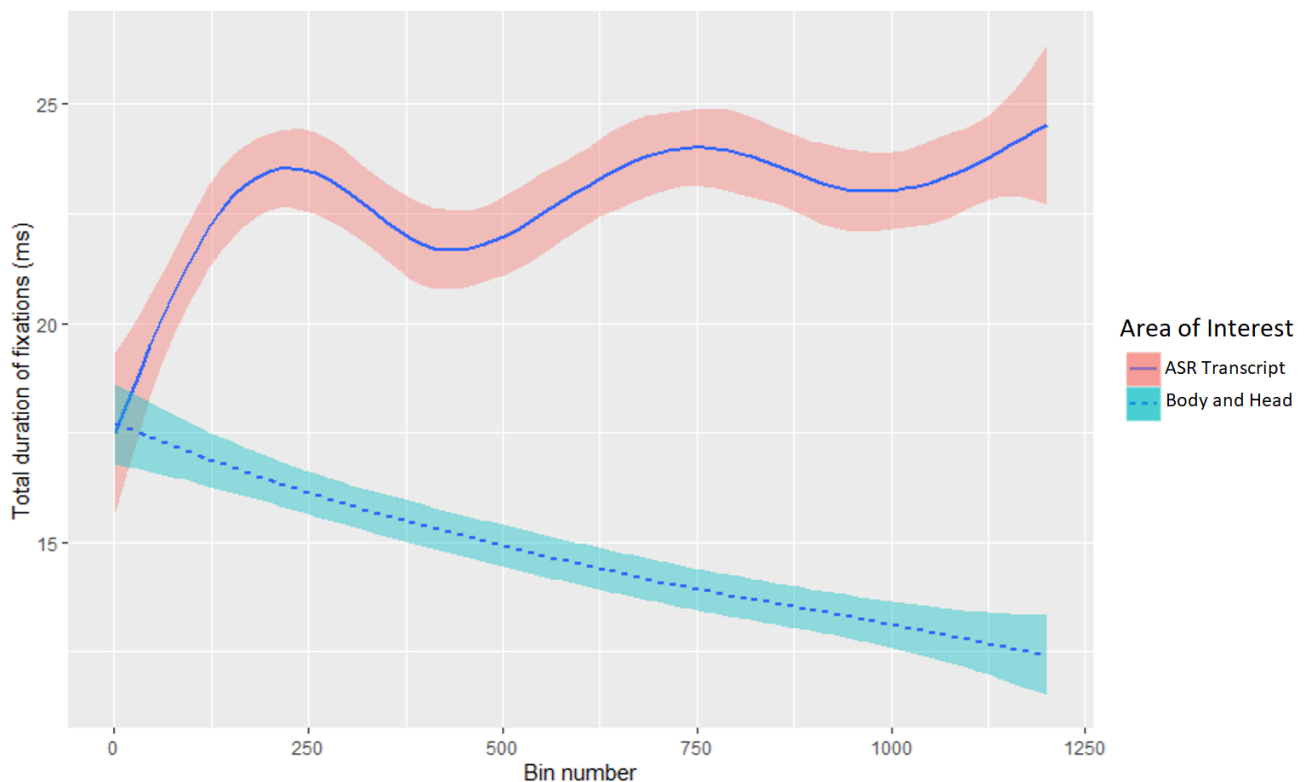


Figure 7.11. Post-utterance fixations over time per areas of interest

To verify this trend, we conducted a linear regression analysis with the bin number (representing consecutive 50 milliseconds intervals) as the predictor and the mean gaze duration as the outcome. The regression model was significant, $F(1,1198) = 18.37$, $p < .001$, explaining approximately 1.5% of the variance in gaze duration (adjusted $R^2 = .014$). For every 50ms increase in time (one bin), participants gazed at the ASR transcript for an additional 0.0023ms on average: $b = 0.0023$, $t(1198) = 4.286$, $p < .001$.

A complementary way of investigating the progressive focus of participants' gaze on the ASR transcript was to analyze the difference between the time they spent fixating on the ASR transcript and the time they spent fixating on the rest of the robot. That is, to study the *relative* focus of participants' gaze attention between both areas of interest (since they could also look elsewhere in the room) over time. Hence, we calculated the difference in mean fixation time between the "ASR Transcript" and the "Body and Head" of the robot for each 50

millisecond interval (i.e., for each bin). We then conducted a linear regression analysis with the bin number as the predictor and this difference as the outcome. The regression model was significant, $F(1,1198) = 48.61$, $p < .001$, explaining approximately 3.9% of the variance in the gaze difference (adjusted $R^2 = .038$). For every 50ms increase in time (one bin), the difference in gaze duration between the “ASR Transcript” and the “Body and Head” increased by 0.0066ms on average: $b = 0.0066$, $t(1198) = 6.972$, $p < .001$. Hence, the post-utterance gradual focus on the ASR transcript is statistically significant but limited: the model's R^2 value suggests that the bin number (i.e., each passing 50 milliseconds) explains only a small portion of the variance in gaze time.

7.6.5. Discussion: Gaze analysis as a preliminary overview

On average, the ASR transcript was heavily gazed at when it appeared, compared to the rest of the robot. This pronounced focus on the transcript begs the question of its local relevance (as a resource, as a remarkable phenomenon, etc.) in situated interactions. In particular, it suggests that, at this critical moment where participants could start to produce an early interpretation of what the robot's response was going to be, the ASR transcript *may* have been a relevant local feature, i.e., one that participants made use of. When it was displayed, it can even be *suspected* that the transcript's relevance for the task at hand was superior to that of the robot's other observable behaviors: as identified in section 7.6.4, the post-utterance focus on the ASR transcript slightly increased during the interaction at the expense of other parts of the robot (its gestures, its head movements, etc.). Yet, a quantitative gaze analysis cannot shed light on these conjectures by itself. That participants gazed at the transcript more than the rest of the robot does not entail that this transcript was, locally, treated as a publicly available and accountable (Kristiansen & Rasmussen, 2021) “conduct” of the robot – the situational relevance of the ASR transcript is not demonstrable by solely describing how long it was looked at. After the previous statistical summary, participants' gaze attention on the ASR transcript is therefore entirely left to be “explained” in an interactional sense by describing how and for which activity this transcript was practically used (if it was) in local contexts.

In his well-known criticism of quantification applied to Conversation Analysis' purposes, Schegloff (1993) suggests that any quantification, if it needs to be done, should be achieved after a careful qualitative analysis of the phenomena studied: so as to avoid aggregating (in the same category) behaviors which were treated as different practices by the members themselves as they were immersed in a local interaction. In the following section, we approach this problem from the other end. Through the detailed analysis of qualitative fragments, we attempt to *highlight the local interactional phenomena aggregated in the quantitative overview we have first produced*. For example, we ask, was the ASR publicly thematized by human participants, and, if so, as part of what local activities? Was the ASR consequential on what actions were “ascribed” (Levinson, 2012) to the robot when it started to respond (or even before it responded)? Was the ASR transcript indexed by participants during the repair sequences they initiated? An EMCA approach appears fit to clarify the typical interactional practices in which took place the “attention economy” objectified by the previous eye-tracking analysis.

7.7. Qualitative fragments

The following fragments were transcribed following Mondada's (2016) multimodal transcription conventions for embodied conduct and Jefferson's (2004) conventions for verbal conduct. All transcription conventions can be found at the beginning of this work, page xiv. Occasionally, when the fragments were especially intricate, a “description” section was added below them to help the reader make sense of the transcripts.

7.7.1. Responding to the ASR transcript as a noticeable source of trouble before the robot's verbal and gestural answer

7.7.1.1 Fragment 7.1: Responding to the ASR transcript as a source of trouble that has priority over the robot's verbal and gestural answer

1. @ (0.3)
@leans towards ROB-->
 2. PA1 **quel est votre travail?**
what is your job
 3. (0.5)@ (1.1) % (0.2)
-->@
- rob %displays "quel est votre mel"-->
"what is your mel"
4. ROB **j'ai entenduf# quel est votre £**
I heard what is your
- pal £throws head backward and right£
img #img.1.1



Figure 7.12. Image 1.1 – PA1 turns away from ROB after reading the ASR transcript

5. ROB **m[el ^ (.) mais] je ne sais pas quoi répondre**
mel but I don't know what to respond
- pal ^walks away from ROB-->
6. PA1 [() mel]
mel

```

7.      (0.3)%(0.7)@(0.9)
rob     -->%
pa2     -->@leans towards ROB-->
8.  PA2  quel: est (.) votre (.) travail:?
       what is your job
9.      (0.3)%(0.3)
rob     %displays "quel est votre copain"-->
       "what is your friend"
10. ROB  j'ai@ deux s[up*ers]# @ amis+%# (.)^(.) [nao et rom]eo
       (.)
       I've got two super friends          nao and romeo
pa2     -->@-straightens up--@             ^steps towards PA1-->
pa2     *points tablet-->
pa2     +gazes towards PA1-->
img     #img.1.2   #img.1.3

```



Figure 7.13. Image 1.2 & 1.3 – PA2 points at the robot's tablet

```

11. PA2      [thh]                [(          )]
12. ROB  +(.)[tous les^ humains]% et robots sont^mes amis aussi
       all humans and robots are my friends too
pa2     ->+gazes at ROB>>
pa2     -->^-----steps towards ROB-----^

```

```

rob [ ( (laughs) ) ] -->%
13. PA2

```

7.7.1.2 Analysis

This fragment displays two treatments of the robot's conduct as inadequate, both of which are initiated before the robot's vocal and gestural response. While ROB's answer (L.4) is still underway, PA1 manifests trouble with the robot's conduct in an embodied way, by throwing his head backward and by walking away from ROB (L.5). He then publicly singles out the troublesome word ("mel") by repeating it out loud (L.6). PA2 attempts to repair this interaction by formulating the same inquiry but, this time, with longer pauses between each word (L.8). Unlike what it did after PA1's inquiry, ROB responds to PA2 without verbally indicating what it heard (L.10), but still displays an ASR transcript (L.9). As PA1 just did before, PA2 starts to react to ROB's conduct as troublesome and noticeable when no meaningful content has yet been made hearable in ROB's ongoing verbal answer: ROB has only uttered "I've got" ("j'ai") (L.10) when PA2 starts to straighten up. She then points towards the belly screen, laughs, and produces a participation shift by gazing and stepping towards PA1 while speaking with him.

Significantly, when PA2 is already responding to the appearance of the ASR, she disregards the vocal response that ROB keeps producing: during ROB's turn, she shifts her attention towards PA1 and speaks or laughs in overlap with ROB's hearable response. In other words, the transcribed information appears to have priority over the verbal response of the robot as an "answerable" or, at least, as a "noticeable" conduct: the delivery of a verbal response by the robot does not interrupt the reactions that follow the appearance of the ASR transcript, and, through this, does not reconfigure the manner in which participants visibly orient to the situation. This is unlike the symmetrically opposed situation visible in, e.g., fragment 7.8, where new information about the ASR transcript reconfigured the participant's interpretation of the previous verbal conduct from the robot.

7.7.1.3 Fragment 7.2: Using the ASR transcript in repairing a mishearing before the robot's verbal and gestural response

```

1. PA1 %comment* te déplaces-tuh?
    how do you move around?
    pal *holds his hand flat and stretches it towards rob->
    rob >>%-displays "?"-->
2. (0.2) * (0.7)
    pal -->*
3. ROB comment?
    what?
4. (3.1)
5. PA1 +tu* sais avancer et reculer un pti %peu?
    do you know how to move forward and backward a little?
    rob >>+gazes at PA1>>
    pal *extends and retracts extended hand towards ROB->
    rob -->%
6. (0.1) * (0.2) ^ (0.6) % (0.1)

```

```

pal      -->*      ^steps back-->
rob      %displays "?"-->
7. ROB   pardon?
         pardon?
8.       (0.1)^(0.4)*(0.4)
pal      -->^      *extends hand towards ROB-->
9. PA1   est ce que tu peux avanc*er un peu?
         can you move forward a little?
pal      -->*retracts hand towards his torso-->
10.      (0.1)*(1.2)*£(0.1)
pal      -->*      *extends and retracts hand towards his torso-->
pal      £gazes at belly screen-->
11. PA1  [a%vanc:e]
         move forward
rob      -->%displays "est-ce que tu peux danser"-->
         "can you dance"
12. ROB  [oh oui](0.4)encore*[une fo]is
         oh yes      one more time
pal      -->*
13. PA1  [av-]*>avan#:cer<£
         mo      move forward
pal      *raises a finger-->
pal      -->£gazes ROB's head-->
img      #img.2.1

```



Figure 7.14. Image 2.1 – PA1 raises a finger

```

14.      (0.2)
15. PA1  pas dan:£ser      (0.1)      £(0.1)$<avan*cer>.
         not dance      move forward
pal      -->£gazes belly screen-->£gazes at ROB's head-->
pal      -->*
rob      $dances>>

```

16. £ (0.1) *# * (0.1)
 pal -->fgazes belly screen-->
 pal *extends hand held vertically towards ROB*
 img #img.2.2

17. * (0.8)%# * (0.1)
 pal *retracts hand towards his torso*
 img #img.2.3
 rob -->%



Figure 7.15. Image 2.2 & 2.3 – PA1 extends and retracts his hand

18. (6.0) £ (3.2)
 pal -->fgazes at ROB's head>>

19. PA1 **est ce que tu peux me donner la météo.**
can you give me the weather forecast

7.7.1.4 Description

L.1, the participant produces a question about the mean or the manner in which the robot moves in space. This question is emphasized gesturally by the participant, as he mimics a movement in space with his hand going forward and backward. The robot then produces the repair initiator “comment ?” (L.3) and keeps displaying a question mark on the tablet. After a long pause of 3 seconds (L.4), the participant utters a different request (L.5) which now questions the robot about its mere ability to go forward and backward. Doing so, he manifests treating the generality of his first question as problematic. His new request is coordinated with the same hand gesture as the previous one. Towards the end of this turn, the robot stops displaying a question mark on the tablet. The participant then starts to move backward (L.6), supporting an interpretation of his previous request as a request to move forward for the robot (instead of a literal question about the robot’s ability to move). The robot then displays a new question mark (L.6) and another repair initiator (L.7) “pardon”. The participant produces a question about the robot’s ability to move forward. This request immediately follows the end of the backward movement achieved by the participant (L.8), which left some empty space in front of the robot. This tends to characterize this utterance as a modalized request to move forward (i.e., an indirect speech act; Searle, 1969), rather than a purely factual interrogation about the robot’s abilities. This sentence takes place as the participant moves his hand forward

(L.8), then backward, mimicking a movement in space as in L.1 and L.5. ROB does not answer for the next 1.4 seconds and, immediately after his gaze focuses on the tablet (L.10), PA1 relies on this silence to self-select and produces a request through the imperative form “avance” (L.11). While the participant just started this turn, the robot displays on the tablet “est-ce que tu peux danser”, i.e., “can you dance” (L.11). From a technical point of view, this ASR transcript corresponds to the (erroneous) transcription of the previous turn produced by the participant (L.9). Then, before the end of the participant request, the robot produces a turn in overlap (L.12). The first part of this turn displays the acceptance of a request (“oh oui”), while, after a short pause of 0.4 seconds, the second part characterizes the activity being agreed upon as having already been achieved in the past by the robot (“encore une fois”).

In overlap with the last part of the robot's turn, the participant reacts with a false start “av-” (L.13) followed by a repetition of the verb used in his prior request “avancer” and a rejection of the verb displayed by the robot on the ASR transcript “pas danser”. He then repeats “avancer” once more. These two repetitions of “avancer” are produced with a slower (L.13) then faster (L.15) tempo – compared to this participant's previous utterances – and by stressing the second (L.13) and then first (L.15) syllable. They are co-occurring with the participant placing his extended index finger in front of his lips (L.13 to L.15). These combined gestural, grammatical, and prosodic features support an interpretation of this turn as pedagogical. The overlap and the “shush” gesture (finger on the lips) indicate a possible attempt at interrupting the robot's ongoing course of action. Remarkably, the participant's gaze is directed at the tablet when the robot displays the ASR transcript (L.10 and L.11) and stays fixed on the tablet as he produces the first part of his response (“avancer”, L.13). He then looks at the robot's head as he says “pas danser”, then alternates his gaze between the tablet and the robot's face during the rest of his turn (L.13 and L.15), and finally maintains his gaze on the tablet for several seconds after the end of his turn (L.16 to L.18). This suggests that PA1 is monitoring ROB's transcriptions of his own repair.

As the participant finishes his turn, the robot starts to move its arms, head and torso in (what is intended by its designers as) a “dance animation”. Immediately after the end of his turn, the participant produces once more an iconic gesture by advancing and retracting his hand in space (L.16), after which the current ASR transcript disappears from the robot's tablet. The participant then stays silent and static for 9 seconds while watching ROB dance (L.18). After 9 seconds, he changes topic and asks if the robot can provide a weather forecast (L.19).

7.7.1.5 Analysis

Until L.12, PA1 produces different requests, focused on ROB's ability to move in space, to which ROB either responds by producing the open class repair initiators (Drew, 1997) “what” (L.3) and “pardon” (L.7), or to which it does not produce any response (L.10). Moreover, during most of this period, ROB is displaying a question mark on its tablet. Through these cues, ROB does “not locate a specific repairable element of the utterance” (Griffiths et al., 2015), nor does it manifest an understanding of any part from the previous talk of PA1. This situation changes after L.12, where ROB positively responds “oh yes, one more time” (“oh oui, encore une fois”) to PA1's previous request and displays the ASR transcript “can you dance” (“est-ce que tu peux danser”). At this point, unlike what he did previously, PA1 does not start a new autonomous request: he reiterates the verb that ROB did not display (“avancer”, L.13 and L.15) and, by using the construction “not dance” (“pas danser”), disallows (Lerner & Kitzinger,

2019) the term transcribed by the robot (“danser”). Because PA1’s turn begins with a false start in overlap with ROB’s utterance (“[av]-avancer”), this turn is designed as purposely interruptive and competing for the turn with ROB (Deppermann et al., 2021; Oloff, 2009): PA1 is attempting to stop ROB’s ongoing action in its tracks. Yet, it is only after this repair that, by starting to dance, ROB displays what action it agreed to by saying “oh yes”. In other words, PA1 characterized ROB as having misheard (or misunderstood) what action was requested *before ROB enacted the action in question* (L.15) – and while ROB never thematized this action verbally. Since it was the first time PA1 heard this response in his interaction with ROB, the timing of this repair is understandable in connection with the display of the ASR transcript L.11.

This repair is remarkable on two points. First, it arises in overlap with ROB (L.12), when ROB’s ongoing turn-so-far is displaying alignment with what came previously (“oh yes”) and does not verbally index the specific content it aligns with. Since, at this moment, ROB is still standing still, no embodied conduct from ROB (gazing, pointing, starting to dance, etc.) can be used as a resource by PA1 to produce an early and evolving interpretation (Deppermann & Schmidt, 2021; Mondada, 2014, 2021) of what action from ROB is ongoing, imminent or “projected”. Secondly, this repair instantly identifies the misheard word (“avancer”, L.13 and L.14) and allows PA1 to stress the misheard syllable (“avancer”, L.15). As such, it constitutes a specific case of self-initiated third turn repair (Schegloff, 1992c) where:

1. A1’s turn (L9. or L.11) is treated as the trouble source.
2. ROB’s transcript (L.11) reveals a misheard term.
3. PA1’s third turn (L.13) initiates and completes a repair by repeating the trouble source¹⁴⁹.

In other words, ROB’s conduct “reveal[s] to the co-participants understandings of the previous talk which the latter find problematic” (Sidnell, 2005), yet without producing a verbal response to PA1. In this sense, ROB’s conduct is treated akin to an embodied conduct revealing a “problematic sequential implicativeness (or what action a speaker has meant to be doing with the turn)” (Schegloff, 1992c), as when, during a driving lesson, the student driver displays a misunderstanding of the instructor’s ongoing instruction through the way in which she turns the steering wheel, and is corrected by said instructor (Deppermann, 2018)¹⁵⁰. Or, when an assistant surgeon’s gesture is aimed in the wrong direction, to which the surgeon responds by initiating a repair before the completion of this assistant’s gesture (Mondada, 2014). Yet, on the other hand, the ASR transcript allows PA1 to pinpoint exactly which term was misunderstood in his previous turn (rather than indirectly inferring it from ROB’s embodied response). In this sense, the ASR transcript is used similarly to a “display of hearing” (Svennevig, 2004) or an “information receipt” (Svennevig, 2004) produced by ROB. The ASR transcription of PA1’s turn provides no demonstration of understanding: it does not formulate the upshot (Enfield & Sidnell, 2017) of what PA1 was saying, and provides no new material about PA1’s course of action which could be “amenable to correction” (Heritage, 2007). This transcript is therefore treated by PA1 as demonstrating, a minima, the “phonological form”

¹⁴⁹ More precisely, PA1 publicly manifests accomplishing “the practice of ‘doing clearer repeat’” (Schegloff, 1992c). That is, not only does he repeat what ROB said, but, by producing his embodied turn in a specific way, he makes publicly accountable that this is the (typical) repair practice currently being done.

¹⁵⁰ “[i]mmediately after the instructor begins to formulate this request, the student starts to turn right, which may be occasioned by a misinterpretation of the upcoming request. The instructor brakes and makes the car stop, while requesting the student to stop” (Deppermann, 2018).

(Svennevig, 2004)¹⁵¹ grasped by the robot from the previous turn – on which PA1 relies to identify the troublesome term.

7.7.2. Responding to the ASR transcript as a noticeable source of trouble after the robot's verbal and gestural answer

7.7.2.1 Fragment 7.3: Using the ASR transcript in accounting for an inadequate verbal answer from the robot

1. ROB %ça va très bien (.) merci (.) et toi?
I'm very good thanks and you?

rob >>+gazes at PA1>>

rob >>%displays "comment tu vas aujourd'hui"->
"how are you today"

2 (0.4)

3. PA1 t'as pas% de la fièvre? (.) le covid?#
you don't have fever? the covid?

rob -->%

img

#img.3.1



Figure 7.16. Image 3.1 – PA1 asks a question to ROB

4. (0.3) £(1.8)

pal £gazes at belly screen->

5. ROB ou% plein.
or a lot

rob %displays "t'as parlé la fièvre le corps vide"->

¹⁵¹ It is possible that this repetition is simultaneously perceived by PA1 as a "claim" from ROB to have understood what is displayed on the ASR transcript, but this hypothesis is not demonstrable from this fragment alone. However, even in this hypothesis, a "claim of understanding" by repetition is, by default, also a display of hearing. Claiming to understand by repeating the previous turn produces at least one element that is "amenable to correction" (Heritage, 2007), and which is therefore "demonstrated" rather than "claimed": i.e., it produces a demonstration of the correct hearing of the previous turn.

"you spoke the fever the empty body"

6. (0.3)£(0.3)%(0.4)
pal -->£gazes at "le corps vide" on the belly screen->
rob ->%
7. PA1 h [°fle @cor:ps vi]de°# he^hehe £.hh (.) y comprends@ pas
the empty body it doesn't understand
pal -->£-----gaze at PA2-----£gaze at ROB's head>>
pal @-----turns towards PA2-----@
pal ^steps towards PA2-->
img #img.3.2



Figure 7.17. Image 3.2 – PA1 steps and turns towards PA2

8. PA2 [le corps vide]
the empty body
9. (0.1)
10. PA1 .h^h@hh (0.3) ^tu t'appelles comment
what's your name
pal -->^takes a step towards ROB^
pal @leans towards ROB>>

7.7.2.2 Description

This fragment starts (L.1) with the robot answering a “how are you” question previously asked by PA1. As it utters this turn, ROB displays the transcript of this question as it received it (L.1). Even though ROB returns the question “and you?”, PA1 does not respond and, instead, maintains the current topic on the robot’s well being by asking if it feels feverish (L.3). As PA1 delivers this turn, ROB stops displaying the ASR transcript for the previous sentence. Yet, 0.3 seconds after the end of her turn, PA1 starts looking at the (then blank) zone where ASR transcripts are displayed on the tablet (L.4). After 2.1 seconds of pause (L.4), ROB answers “ou plein” and displays the ASR transcript “t’as parlé la fièvre le corps vide” (L.5). PA1 is still looking at transcripts zone (L.6) and, less than a second later, repeats a part of the ASR transcript “le corps vide” (L.7). She utters this repetition in quasi-perfect overlap with a similar repeat produced by PA2 (L.8): the display of a troublesome transcript from the robot is initially responded to in a choral way (Lerner, 2002) by human participants.

However, in the same turn, PA1 produces a laugh and an account for the robot's troublesome ASR transcript by stating that it "doesn't understand" (L.7). As she delivers the beginning of this turn, she looks at PA2, steps forward and turns towards PA2 (L.7). This embodied delivery produces a participation shift, characterizing PA2 as the recipient, and momentarily forming a shared inner space (Kendon, 1990) between PA1 and PA2. Moreover, even though PA1 looks back at the robot when she produces the evaluation at the end of her turn, she refers to it using the third person ("it doesn't understand"). After her account of the robot's troublesome transcript, PA1 changes topic (L.10) by asking ROB his name, while, this time, addressing it using the second person. As she utters this turn, she looks back at ROB, leans, and takes a step towards it (L.10). As a whole, this turn and its embodied delivery constitute the end of the shared inner space between PA1 and PA2, and enact a new participation shift which includes the robot again.

7.7.2.3 Analysis

While the troublesome part of the ASR transcript becomes interactionally relevant, ROB's verbal answer ("ou plein", L.5) is never commented on nor responded to, as participants change topic in the end by asking the robot its name (L.10). The robot's answer is not treated as conditionally relevant (i.e., as producing a pressure to produce an answer of a certain type). Moreover, the thematization of the ASR transcript occurs after an unusually long silence of one second following ROB's turn (L.6). This suggests an interpretation of ROB's turn as *an incongruous or unexpected second pair part* after PA1's previous question.

PA1 and PA2's verbalizations (L.7 and L.8) of the robot's ASR transcript do not accomplish an action akin to some repetitions of a previous turn which, in human-human interactions, "problematise what is being repeated and typically solicit a response" (Rossi, 2020). Indeed, the absence of response from the robot to PA1's reading of the transcript (L.7) is not made accountable by the participants. In other words, PA1's reading of the ASR transcript is closer to a "repetition as quotation" which "indicate[s] an individual's intention to comment upon the other's talk, by taking it up as the theme and object of reference of his own discourse" (Perrin et al., 2003).

Moreover, because PA1 and PA2's verbalizations of "le corps vide" are choral (Lerner, 2002), they demonstrate the co-alignment of PA1 and PA2 in perceiving this specific part of the ASR transcript as a noticeable phenomenon. Yet, this choral noticing and what follows exclude the robot as a co-present interlocutor. This is displayed by the linguistic and embodied participation shift produced by PA1: she looks at PA2, torques her body, and takes a step towards him (L.7), while laughing and referring to ROB in the third person (L.8). Doing so, she creates a shared inner space with PA2. The treatment of ROB as a "non-interactant" is also displayed by the account and the laugh that PA1 produces about ROB. It evaluates ROB as unable to grasp the meaning of the previous sentence ("it doesn't understand") and, by producing this third person account in front of ROB, treats it as a (socially) absent third party whose face would not require to be protected from negative evaluations or mockery. However, this exclusion of the robot is only momentary, since PA2 reverts the participation shift on every point: she turns, leans, and looks towards ROB, while addressing it in the second person ("tu", L.10). This is a striking example of the severe changes in interactional status that the robot can undergo during an interaction (K. Fischer, 2021) where, as a "transient phenomenon" (Pelikan et al., 2022), it can be "enacted, in one breath, as an agent and a thing" (Alač, 2016).

On a final note, eye-tracking data reveal that PA1's original inquiry L.3 is followed by an anticipatory monitoring of the robot's ASR transcript. Indeed, 0.3 seconds after her turn (L.4), she starts to gaze at the zone on which the ASR will appear (and where it appeared so far since the beginning of the interaction), then keeps her gaze mainly on this zone for the next 1.8 seconds of silence, through ROB's answer, and until the start of her own turn (L.7). This visible anticipation of the appearance of an ASR transcript constitutes one of the individual cases aggregated in Figure 7.11, which displays the (minor) tendency of participants to focus more and more on the ASR over the course of an interaction.

7.7.3. Reading the transcript as a full-fledged speaking turn

7.7.3.1 Fragment 7.4: Exploiting what is written on the tablet as a resource for a public performance

```

1. ROB  +%plait-il?
        excuse me
    rob  >>+gazes at pal>>
    rob  >>%displays "?"-->
2.      (0.1)@          (0.3)          @ (0.5)
    pal  @straightens up slightly@
3. PA1  je @suis Jeanne
        i am Jeanne
    pal  @leans forward-->
4.      (0.1)@(0.2)@          (0.8)          @ (0.7)%(0.1)
    pal  -->@          @straightens up completely@
5.      (0.7)%(0.1)
    rob  -->%displays "je suis chiant"-->
        "you are annoying"
6. ROB  okai(0.1)Δ(0.1)@(0.1)tu* es #chi@ant.
        okay          you are annoying
    pal  Δopens mouth wide-->
    pal  @,,,,,,,,,,,,,,,,@turns towards group-->
    pal  *points tablet-->
    img  #img.4.1

```



Figure 7.18. Image 4.1 – PA1 points at ROB's tablet

```

7.      (0.3)@(0.1)
  pal      ->@
8. PA1    Δy% ((vous dit)) j'étais chiant!*
          it ((told you)) i was annoying
  pal -->Δ
  pal      -->*
  rob     -->%
9.      (0.1)@(0.1)@
  rob      @,,,,,@turns towards ROB->
10.PA1    ai* chui @ pas chiant #@j'v[ais% t'pu]nir!
          hey i'm      not annoying i'm gonna punish you!
  pal      *waves finger in front of ROB's tablet>>
  pal      ->@leans towards ROB@
  img      #img.4.2
  rob      %displays "?">>

```



Figure 7.19. Image 4.2 – PA1 waves a finger in front of ROB's tablet

7.7.3.2 Description

ROB initiates a repair L.1 while a question mark is displayed on the tablet. After a short pause (L.2), PA1 self identifies (L.3) by uttering “I’m Jeanne” (“Je suis Jeanne”) while leaning towards ROB. After a long silence of 1.7 seconds (L.4) during which PA1 stands up straight, ROB displays “I’m boring” (“je suis chiant”) on its tablet. Immediately after, L.5, ROB produces a claim of understanding (Heritage, 2007) by repeating and aligning with the previous sentence of PA1 (as heard by ROB): “Ok, you are boring” (“Ok, tu es chiant”), L.5. As ROB delivers this turn, PA1 produces a participation shift on several levels: in quick succession, she opens her mouth wide, turns towards the group behind her and points to ROB’s tablet. Once PA1 is fully oriented towards the group (L.6.), she produces a report (L.7) of ROB’s previous utterance, using indirect speech, which characterizes the group behind her as the intended recipient of the robot (“he told you I was annoying!”). Meanwhile, the ASR transcript disappears from ROB’s tablet. Nearing the end of her turn, PA1 initiates a new participation switch back to ROB: she fully turns back towards it (L.7 to L.8) and stops pointing the tablet (L.8). In this stance, she then disaffiliates with the robot (“I’m not annoying”, L.9) and ostentatiously threatens to reprimand the robot (“I’m going to punish you!”) as she leans towards it and wags her finger in front of the tablet (L.9).

7.7.3.3 Analysis

Similarly to fragment 7.1, the embodied reaction of PA1 to ROB’s erroneous transcript (L.5) is initiated before ROB has uttered the object of its turn. When PA1 starts to display the robot’s conduct as noticeable (i.e., when she starts to turn towards the group, with an open mouth while pointing to the tablet), ROB’s verbal turn-so-far is limited to a claim of understanding (“ok”) and has not yet named what the robot aligns with (“you are annoying”). This supports an interpretation where, even though PA1’s indirect speech L.7 may be referring to ROB’s entire hearable turn, her original theatrical noticing exclusively indexes and responds to the ASR transcript.

PA1 does not treat the ASR as a display of misunderstanding or as a direct transcript of her previous turn, but, rather, as *an evaluation of herself (by the robot) as “boring”*. Through her conduct, she makes both the robot’s ASR transcript and the robot’s verbal answer publicly relevant (to her audience) as a contribution. She produces a blame of the type “A tells B that C (generally absent) acted badly” (Laforest, 2009). However, while she does this through the typical participation shift format identified in previous fragments, and while referring to the robot as the third person, her overall behavior is less exclusionary towards ROB than what is observable in these previous fragments. Through her behavior, PA1 uses the robot’s conduct as a resource to configure co-participants as an audience for ROB. Indeed, once PA1 has made relevant ROB’s conduct as an insult, she switches back to ROB to deploy a theatrical remonstrance based on this understanding of the situation (as ROB being insulting).

This fragment illustrates that a troublesome ASR transcript can be treated as an autonomous contribution to the interaction (i.e., what the robot “says”) rather than as a mere “information receipt” (Svennevig, 2004). ROB’s behavior is responded to as remarkable and socially meaningful – and not as the result of a technical problem where the robot’s microphones did not properly receive PA1’s voice – in order to engage co-present interlocutors. In other words, the ASR transcript (and possibly the robot’s verbal answer) are turned into a “resource for a playful performance”. This may constitute a case of what Pelikan

et al. (2022) name “ascribed agency”: PA1 works to topicalize ROB’s conduct as introducing a meaningful and remarkable contribution to the ongoing activity, preserving ROB’s agency and competence as an interactant in the same move.

7.7.3.4 Fragment 7.5: Reading the ASR transcript as “what the robot says”

```
1. PA1    allez *on se checke? (0.3) [on se %checke.]#
           come on can we fist bump    can we fist bump
rob      -->%displays "?"-->
pal      *extends closed fist in front of rob->
img      #img.5.1
```



Figure 7.20. Image 5.1 – PA1 extends his closed fist in front of ROB

```
2. ROB      [comm+ent?]
           what
rob      +gazes fist pal-->
3.      (0.4)
4. PA1    on se check?
           can we fist bump
5.      (0.4)%(0.1)%(0.3)
rob      -->% -->%displays "hello c'est vrai"-->
           "hello it's true"
6. ROB    bonjour:
           hello
7.      (0.1)
8. PA1    hello@* c'est ^vrai?#
           hello its true
pal      -->*opens hand palm facing the ceiling-->
pal      @torques slightly towards PA2-->
pal      ^steps back-->
pal      #img.5.2
```

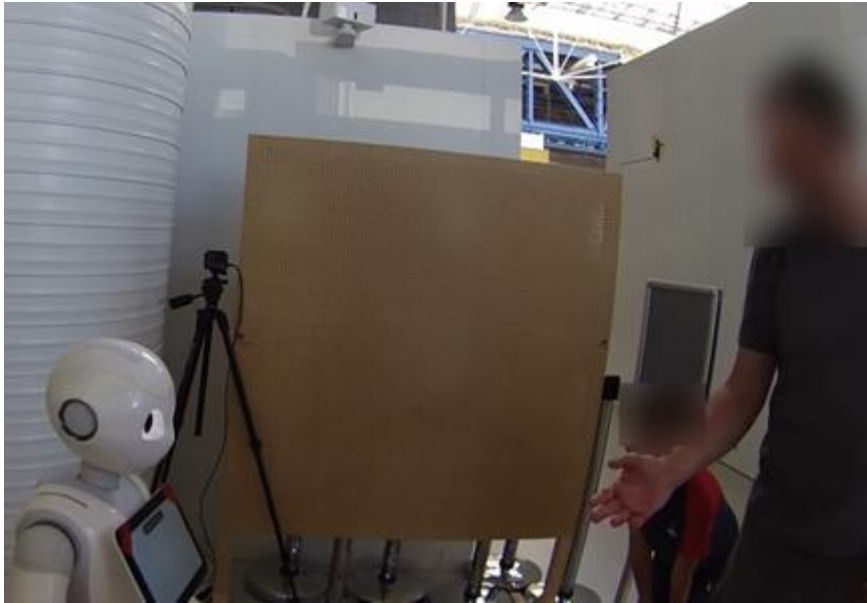


Figure 7.21. Image 5.2 – PA1 opens his hand, palm facing the ceiling

```

9.          (0.4)
10.PA1     qu'est ce* @qui^ racon*te
           what is it/he talking about
           pal          -->*retracts arm*
           pal          -->@
           pal          -->^
11.          (0.3)%(0.2)
rob        -->%
12.PA2     hello
           hello
13.          (0.2)
14.PA1     eh$ ça mar[che]$ (.) ça marche [trop] mal
           it works           it's working so poorly
           rob        $. . . . . $-waves----->>
15.          (0.2)*(0.2)*
           pal        *. . . . . *-waves>>
16.PA2     [eu]          [hello?]
           hello
17.PA1     sa[lut!] (.) allez [salut]
           bye           alright bye
18.PA2     [hello!]    [hello!]
           hello           hello
19.          (0.2)@
           pal        @turns away from ROB>>

```

7.7.3.5 Description

PA1 initiates a request to produce a mutual “check” gesture (L.1). The continued relevance of the request is enforced by keeping his hand extended in front of ROB for a long period (L.1 to L.8), as a form of “frozen gesture” (Schegloff, 1984). After PA1 marks a short pause (L.1),

ROB displays a question mark on its tablet and produces a repair initiator (“what?”) in overlap with the end of PA1’s turn (L.2), as it gazes at PA1’s hand. PA1 complies with ROB’s repair initiation by repeating his previous request (L.4). After a short silence (L.5), ROB stops gazing at PA1’s hand, displays “hello it’s true” (“*hello c’est vrai*”) on its tablet (L.5) and responds with a greeting (“*bonjour*”, L.6). PA1 immediately reads out loud the ASR transcript, ending his turn with a rising intonation (L.8). As he does so, he opens his hand palm up in a perplexed gesture (L.8) and takes a step back, both of which constitute embodied markers of trouble. As a whole, this interplay between visual and linguistic resources (Mondada, 2011) publicly reveals the information provided by the ASR transcript as inadequate with the ongoing sequence.

PA1 then explicitly questions the current relevance of the previous conduct from ROB (“what is he talking about”, L.10). As he delivers this turn, PA1 lowers his arm, closing the possibility for ROB to satisfy PA1’s original request by hitting his fist with its own: it becomes impossible to co-produce a “check”. One of the children accompanying PA1 (thereafter PA2) then greets the robot (L.12). After a pause of 0.2 seconds, PA1 produces a negative evaluation of the robot’s functioning (“it works so badly”, L.14), still referring to it in the third person and using the neutral form “it”. This formulation – combined with PA1 stepping away from ROB – tends to characterize the robot as a (socially) absent third party: in spite of its physical and perceptual proximity with the human members, the robot is not addressed as a full-fledged interlocutor in the current interaction by PA1. However, after a short pause, PA1 utters the closing token “*salut*” (L.17) with a waving gesture extended towards ROB – which initiated its own waving gesture at the beginning of PA1’s turn (L.14). In French, the term “*salut*”, uttered another time by PA1 at the end of his turn (L.14), can constitute either a greeting or a goodbye, depending on the context. Here, having previously canceled his “check” gesture (L.10), PA1 starts to turn away from ROB after finishing his waving gesture (L.19): this supports an interpretation of PA1’s “*salut*” as an interaction closing “goodbye”.

7.7.3.6 Analysis

One of the most significant phenomena of this transcript is that, when it comes to construing the ongoing course of action of the robot, only the ASR transcript is demonstrably treated as relevant for “the purposes at hand” (Garfinkel, 1967) by PA1. Even though ROB produces a greeting token (“*bonjour*”) L.6, PA1 does not deal with it. He does not respond to it (e.g., with a return greeting), nor does he evaluate it or specifically topicalize this sudden greeting as inadequate. What he demonstrably treats as relevant for the ongoing activity is the ASR transcript produced (L.5) before the robot’s greeting.

By reading the transcript out loud (L.8) with a rising final intonation and pointing the tablet with his extended hand, PA1 makes public that something is not working. Because it is quickly followed by an evaluative turn from PA1 (L.10), this repetition is not a repair initiator addressed to ROB to re-establish intersubjectivity. Rather, by reading it, PA1 topicalizes the troublesome character of the ASR transcript, and, similarly to human-human “repetitions as quotations” (Perrin et al., 2003), signals what he reads as the content of his upcoming discourse. Indeed, PA1 turns L.8 and L.10 are not constructed as self-talk. Through his hand-pointing gesture (L.8) and his turn towards one of the children around him (L.8), PA1’s contribution is visibly recipient designed for co-present participants, themselves configured in space as an audience (Keevallik, 2018) around PA1 and ROB. In sum, PA1 produces a “comment as a sequence initiation” (Keevallik, 2018): in an activity taking place among co-

present participants, PA1 presents ROB's tablet as an element of their surroundings which can be commented upon.

Remarkably, once the ASR transcript has been made publicly relevant, PA1 treats it as a contribution stemming from ROB (L.10) – and not from, e.g., an autonomous tablet disconnected from what ROB does or says. Indeed, the turn L.10 both characterizes the ASR as something that ROB “says” (“what is he/it talking about”) – rather than what ROB hears or understands (e.g., fragments 7.2 and 7.5) – and attributes the authorship of what the transcript “says” to ROB. PA1 makes ROB, as a unified and coherent entity, accountable for what is written on the tablet: PA1's embodied and verbal conduct orients to (and publicly “enforce”) the tablet and ROB as a single and cohesive self, rather than as two independent entities.

However, PA1's turn L.10 also excludes the robot as a recipient of his evaluation. Even though the format of the participation shift he produces L.8 is less marked than in fragments 7.1 and 7.3, it shares their most significant traits. PA1 does not turn or torque towards co-participants but bodily disengages from the focused interaction with ROB by taking a step back (L.8), while he suddenly starts to refer to the robot in the third person. As in fragments 7.1 and 7.3, this participation shift arises in the turn immediately following the appearance of the troublesome ASR. It lasts throughout the negative evaluation sequence (“ça marche trop mal”), which is produced in the third person and without any visible precautions to preserve a hypothetical “face” (Goffman, 1967) of the robot, or to maintain it as a competent interactant. Moreover, similarly to fragments 7.1 and 7.3, PA1's evaluation sequence is followed by another participation shift, where PA1 reengages the robot (end of L.14), this time by addressing a verbal goodbye and a waving gesture to ROB.

7.7.3.7 Fragment 7.6: Reading the ASR transcript as a full-fledged speaking turn

1. ROB +%plait il?
pardon
rob +alternates gaze between participants>>
rob >>%displays “?”-->
2. (0.2)
3. PA1 fais nous une blague/
tell us a joke
4. PA2 =hhhh%h.
rob ->%displays “fais nous une blague”->
“tell us a joke”
5. (0.3)
6. ROB ok (0.5) comment fair tenir une giraffe dans% un frigo
en trois étapes
ok how to fit a giraffe in a fridge
in three steps
rob -->%
ROB [tu ouvres la porte (.) tu fais rentrer la girafe (.) tu]

```

fermes la porte
you open the door      you put the giraffe in      you
close the door

7. PA2 [t'ouvres la porte tu rentres la gira:fe et tu fermes la porte]
you open the door you put the giraffe and you close the door
8. (0.1)%(0.4)
rob %displays "ferme la"-->
"shut up"
9. ROB min:ce(0.5)désolé
darn sorry
10. PA1 =°fεfrme la°#
shut up

pa1 εgaze towards PA2-->
img #img.6.1

```

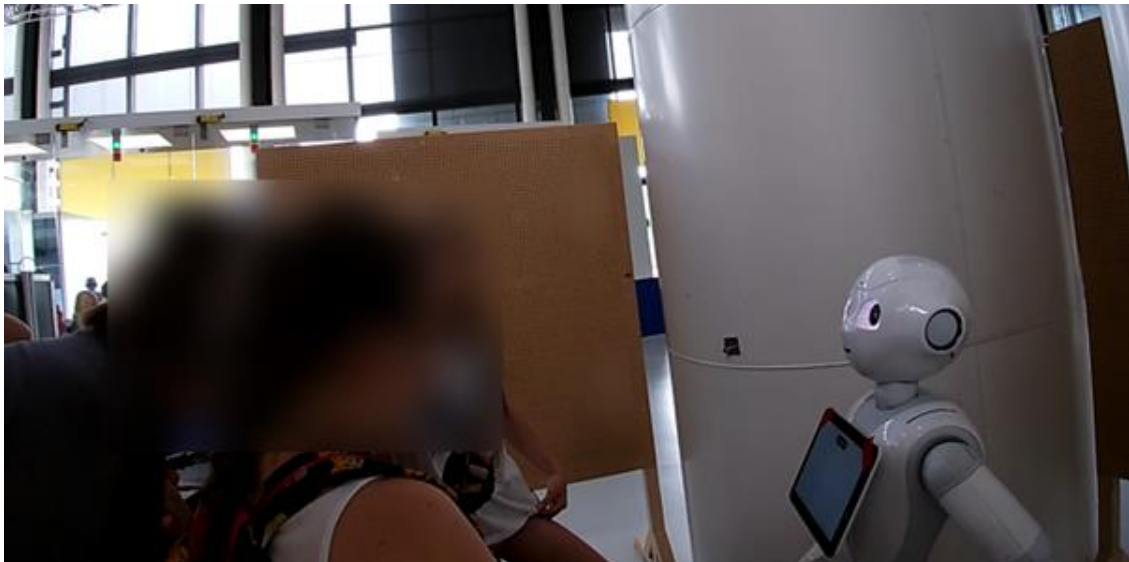


Figure 7.22. Image 6.1 – Group of visitors looks at the ASR transcript

```

11. (0.2)ε(0.1)
pa1 -->ε
12. ((participants laugh chorally))%
pa1 -->%
13. PA3 * @hé:/#(0.5)εon parle * pas εcomme ça hein/
we don't talk like that huh

pa3 *waves finger in front of ROB's head*
pa3 @leans towards ROB->
pa3 εgazes at tabletεgazes at ROB's head>>

```



Figure 7.23. Image 6.2 – PA3 waves a finger in front of ROB's head

```

14 .          % ( 0 . 1 ) @ ( 0 . 1 )
    rob       %displays "?">>
    pa3       -->@
15.  ROB    hein?
        huh?

```

7.7.3.8 Analysis

As ROB is telling a joke (L.6), PA2 demonstrates her knowledge of the joke being told by reciting it in overlap with ROB (L.7). After this choral telling of the joke, ROB displays “ferme la” (L.8), which can either mean “shut up” or “close it” in French, depending on the context¹⁵². It then produces an apology with “darn, sorry” (L.9). PA1 reads out loud the ASR transcript “ferme la”, observably and hearably addressing PA2 by looking at her and by speaking in a lower voice than before (L.10). The entire group of co-present participants then laughs for several seconds (L.12). At the end of this collective laughter, and similarly to fragment 7.4, PA3 produces a theatrical remonstrance by ostensibly waving her finger in front of ROB’s head and by stating “we don’t talk like that” (L.13). By publicly addressing the robot’s head as part of a blame that indexes what appeared on the tablet, PA3’s remonstrance explicitly topicalizes the utterance written on the tablet as an utterance over which the (humanoid) robot has control, rather than as a mere transcript. The ASR transcript is not treated as the result of

¹⁵² From a technical point of view, ROB detected only the words “ferme la” in the previous utterance from the participant, in which they meant “close it (the door)”. Having not grasped the rest of the sentence, ROB’s chatbot matched these two words with pre-written responses to the words “shut up”. The utterance “darn, sorry” was among these responses.

a process that automatically transcribed some elements of PA2's speech, but as original content over which ROB is held normatively accountable.

7.7.4. The ASR transcript as a resource for third parties in contesting the sequential relevance of (embodied or verbal) responses from the robot

7.7.4.1 Fragment 7.7: Using the ASR transcript in accounting for a breakdown in interaction after it arose

1. PA1 £(.) .hhh@je *peux t'serrer * la main?
can I shake your hand?
pal -->fgazes at ROB's head-->
pal @leans forward-->
pal *opens her hand*
2. (1.0)@(0.4)%(0.2)£(0.1)
pal -->@
rob -->%displays "tu peux fermer la main"-->
"can you close your hand"
pal -->fgazes at belly screen-->
3. ROB quelle \$ main?
which hand?
rob \$lowers and raises arms-->
4. (0.1)£(0.1)%(0.1)*(0.3)
pal -->fgazes at ROB's right arm-->
pal *extends arm towards ROB's right arm-->
rob -->%
5. PA1 eu::: (.) la gau\$che/
the left one
rob -->\$
6. (0.1)£ (1.4)%(0.1) £(0.2)
pal -->fgazes at ROB's headfgazes at belly screen-->
rob %displays "la gauche"-->
"the left one"
7. ROB d'a\$ccord
all right
rob \$raises both forearms-->
8. (0.1)£(0.1)*(0.3)#\$(0.4)
pal fgazes at ROB's right hand-->
pal *extends right arm towards ROB's right hand-->
rob \$retracts right arm-->

img

#img.7.1



Figure 7.24. Image 7.1 – PA1 extends her hand towards ROB's right hand

9. (0.4)*(0.1)£(0.1)
pal *retracts right arm->
pal £gazes at belly screen-->
10. (0.2)\$ (0.2)£*(0.1)+(0.4)
rob -->\$extends left arm->
pal -->£gazes at ROB's left hand->
pal -->*switches eye-tracking recorder to right hand-->
rob +gazes at its left arm-->
11. PA1 ah.*\$okai
oh okai
pal -->*extends left arm towards ROB's left hand-->
rob -->\$closes its left hand-->
12. (0.3)%(0.1)£(0.1)
rob -->%
pal -->£gazes bellyscreen
13. PA2 non* mais# il a compris *\$fermer la main
no but he/it understood close the hand
pal -->*touches top of ROB's left hand*
rob -->\$lowers left arm-->
img #img.7.2 and 7.3

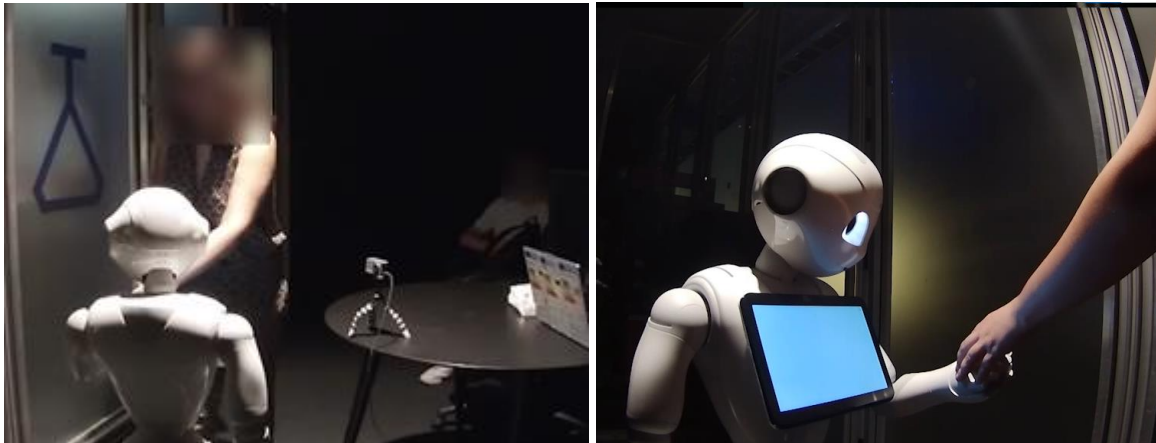


Figure 7.25. Images 7.2 & 7.3 – PA1 touches ROB’s left hand while PA2 states that ROB understood “close the hand”

```

14.          (0.2)$(0.3)+(0.5)
rob          -->$
rob          -->+gazes PA1>>
15. PA1      nan::* mais@£ c'est s- @          £oui
           no    but    it's          yes
pal          *leans head towards PAR2>>
pal          @---shrugs--@shrugs-->
pal          -->£gazes ROB's right arm£gazes belly screen-->
16. PA1      d'accord@ se£rrer la main(.)ok/*          £(.) ok/ ca va
           all right shake the hand    ok          ok it's all right
pal          -->@
pal          -->£gazes ROB's right hand-->£gazes ROB's head>>
pal          *waves towards ROB->
17.          (0.1)£@(0.1)*
pal          ->£
pal          @rotates towards PAR2>>
pal          ->*

```

7.7.4.2 Description

PA1 produces a request to shake ROB’s hand (“can I shake your hand”, L.1), while leaning towards ROB and opening her hand (L.1). After a silence of 1.4 seconds (L.2), ROB displays instead “can you close your hand” on its tablet. As PA1 focuses her gaze on ROB’s tablet, ROB asks “which hand” (L.3) and moves its arms up and down. PA1 then extends her right arm towards ROB’s right arm (L.4) and responds “the left one” (L.5). After a silence of 1.6 seconds, ROB displays “the left” (“*la gauche*”) on its tablet (L.6) and produces an agreement token (L.7) while it raises both forearms. PA1 observably responds to the robot’s verbal alignment and embodied conduct by extending her right hand towards ROB’s right hand (L.8).

Yet, shortly after PA1 started her movement, ROB starts to retract its right arm, in reaction to which PA1 also withdraws her own arm (L.9). ROB then starts to extend its left arm (which, from a technical point of view, corresponds to the “closing hand animation” which was scripted on the robot) and looks at this arm (L.10). After 0.2 seconds, PA1 gazes at ROB’s left arm, puts the object she was holding in her right hand (L.10), and extends her left hand towards ROB’s left arm (L.11). As she extends her arm, she verbally formulates her understanding of the robot’s embodied conduct with the change-of-state token “ah” (Heritage, 1985). However – in line with its understanding of PA1’s initial turn as a request to “close its hand” (L.2) – ROB starts to close its left hand (L.11), making it impossible for PA1 to hold ROB’s hand as a handshake would require. After a short pause (L.12), PA1 adapts to this physical obstacle to the completion of a mutual handshake by placing her hand on top of ROB’s closed hand. (L.13). PA2, who was watching the scene unfold silently until this moment, produces an account of the robot’s understanding of the situation (“he/it understood close the hand”, L.13) prefaced by a “no but” (L.13). Doing so, PA2 demonstrably contests PA1’s embodied interpretation of the interaction (as an ongoing mutual handshake) and attributes a different cognitive state of understanding to the robot. As PA2 produces this turn, ROB starts to lower its arm, which PA1 gazes at (L.13). After a 1-second pause, PA1 answers a similar “no but” prefaced turn (L.15 and L.16).

7.7.4.3 Analysis

This fragment illustrates a nominal use of the ASR transcript as a resource to identify, account for, or repair trouble in interaction. PA2’s account (L.13) comes after a long sequence of visible struggle to coordinate between PA1 and ROB: PA1 originally interprets ROB’s left arm gesture as the initiation of a handshake (L.8), reconsiders this course of action when ROB lowers its left arm to raise the right one, frees her right hand from the object she was holding (L.10), and is finally confronted to ROB’s closed hand when she approaches her right hand (L.11 and L.13).

Significantly, even though PA1 already displayed her treatment of the ongoing action as a coordinated handshake since the beginning of the fragment (L.1), it is only after PA1 meets no adequate embodied response from ROB that PA2 chooses to index the previously visible ASR feedback. This transcript, which was never indexed by PA1 or PA2 before this moment, is therefore (publicly) used as a resource to re-establish a form of intersubjectivity *only after* an interaction breakdown is manifested by the mismatch between PA1 and the robot’s embodied response. When the ASR transcript (erroneously) displayed a receipt of PA1’s request as “can you close your hand” (L.2), PA2 did not produce an account of this mismatch. It is only when trouble arose with the robot’s gestural response that the ASR transcript was observably indexed to account for this breakdown in the interaction. This differs from, e.g., fragments 7.1 and 7.2, where the ASR transcript was used *before any trouble had arisen* with the robot’s verbal or gestural conduct.

As a consequence, on the interactional surface (independently of the question of whether PA1 is *pretending* the robot is collaborating to achieve a mutual handshake), this fragment displays two different treatments of the same gesture produced by the robot. The first one, enacted by PA2, uses an informational configuration that includes the tablet to contest the sequential relevance of the robot’s conduct. The second one, carried out by PA1, orients to the “what-is-going-on” of the situation by (publicly and demonstrably) relying

exclusively on the robot's verbal and gestural conduct in each sequential context. However, PA2 treats these two interpretations as unable to coexist. By prefacing his turn with a “no” (Lerner & Kitzinger, 2019), L. 13, PA2 does more than accounting for PA1's trouble to properly complete her handshake gesture with the robot. He negates¹⁵³ that a mutually understood handshake was ever present – i.e., that there existed an “intersubjectively shared reality” on this point between the robot and PA1. In other words, in PA2's interpretation, *the action that ROB achieved by raising its hand partly depended on what was written on its tablet – and not only on the relevance of the robot's gesture as a potentially adequate response to PA1's request.*

Indeed, more than a simple resource to infer a breakdown in intersubjectivity, this fragment constitutes a case of the ASR feedback oriented to as an externalization of “what is inside the head” of the robot. Using the ASR transcript, PA2 inputs a mental state to ROB by asserting that it “understood ‘fermer la main’” (L.13). In other words, PA2's delivery of his turn does not formulate a hypothesis based on external cues about what ROB “understood”: he is not applying a “documentary method” (Garfinkel, 1967) to produce a reasonable depiction of an underlying phenomenon. On the contrary, it displays direct access to information about the external world on which ROB relies to trigger its subsequent embodied conduct, and which overrides other possible interpretations (by PA1) of ROB's course of action. PA2's statement on the action being achieved by ROB through its arm gesture is produced by indexing the ASR transcript as a *publicly accessible cognitive state* of the robot.

Finally, and crucially for our analysis of this fragment, to the best of our knowledge, PA2 could not pinpoint what ROB understood without the help of the ASR transcript: PA1 or PA2 had never observed nor interacted with a Pepper robot before, and none of them had triggered the “hand closing animation” until this point of the interaction. The presence of the ASR transcript is therefore consequential for the interaction, as it makes it possible for PA1 to produce the specific format of account mentioned above – i.e., a direct peek into the robot's head. The ASR transcript allows PA2 to clarify as inadequate the second pair part produced by the robot (here, extending and closing its left hand) that could, otherwise, be treated as sequentially relevant (as PA1 was doing until then¹⁵⁴). In other terms, the presence of the ASR was central in the reevaluation of the situation as a misunderstanding rather than as a (non-smooth) mutually coordinated handshake.

¹⁵³ Lerner & Kitzinger (2019) note that “repairs can be accomplished with a no-prefaced rejection of the trouble source followed by its replacement. The no—by explicitly disallowing the trouble source as having been mistaken—also prepares recipients to understand what is said next as being a correction” (Lerner & Kitzinger, 2019). In our case, the PA2's no-prefaced turn casts PA1's displayed interpretation of the robot's course of action as mistaken.

¹⁵⁴ It is not possible to reasonably assert that PA1 and PA2 represent the two extreme poles of “preference for progressivity” versus “preference for intersubjectivity” (Heritage, 2007). The publicly available information available to the analyst in this fragment does not allow to demonstrably show that PA1 has, indeed, read the ASR transcript but decided to “go on” with the breakdown in intersubjectivity – or, on the contrary, that PA1 simply did not realize that the robot has misheard her. Therefore, even though we are faced with two different treatments of the robot, only PA2's behavior can be reasonably described as *responsive* to the ASR transcript. PA1's conduct may very well constitute a case of doing “not doing repair” (Pilnick et al., 2021) or, conversely, result from a perception of the robot as involved in a properly coordinated activity. These internalist hypotheses are beyond the scope of an EMCA approach.

7.7.4.4 Fragment 7.8: Using the ASR transcript in contesting the definition of the situation as one of intersubjectivity

1. PA1 est-ce que tu peux lever la tête en:: l'air?
can you look up?
2. (2.9)%(0.3)
rob %displays "est ce que tu peux laver la tête en l'air"-->
"can you clean up the head upward"
3. ROB désolé (.) je ne dispose pas de cette fonctionnalité.
sorry I do not have this feature
4. (0.3)%(1.9)
rob -->%
5. PA1 °bah je sais pas°@ -il fait peur un peu* hh
well I don't know it scares me a little bit
pa1 @turns towards PA2-->
6. (1.6)
7. PA2 tu @parles pas assez fort () par laver (y voulait dire) lever
you're not speaking loud enough by "clean up" it meant "look up"
pa1 -->@turns towards ROB-->
8. (2.0)
9. PA1 il @a marqué laver?#
it wrote "clean up"?
pa1 -->@turns towards PA2-->
img #img.8.1



Figure 7.26. Image 8.1 – PA1 points at ROB's tablet while asking PA2 for confirmation about what it previously displayed

10. (1.0) £ (0.5) @£ (0.5)
 pa2 £ nods £
 pa1 @turns and leans towards ROB>>
11. PA1 est ce que tu peux >leuver< la tête?
 can you >look< up?

7.7.4.5 Description

The fragment starts with the rejection of PA1's request (L.1) by ROB. Even though ROB's utterance denies what PA1 was asking for, it is not made relevant as an inadequate second pair part. In other words, even though ROB's utterance halts the activity at hand by rejecting (L.3) PA1's request, *it is a sequentially relevant response, of the type and form made relevant by PA1's request*. PA1 then produces a complaint to PA2 about being scared of ROB. Doing so, she creates an answer slot for PA2. However, PA2 does not use this available slot to answer PA1's complaint but, instead, to produce an alternative account of the interaction that just unfolded. Before explaining the trouble with ROB hearing "laver" ("cleaning") instead of "lever" ("looking up"), he starts with an account of why this trouble happened ("you don't speak loud enough").

7.7.4.6 Analysis

Once information about what was written on the ASR transcript is provided to PA1 by PA2, this information completely reconfigures, a posteriori, the action attributed to ROB by PA1. By asking a confirmation question "it wrote clean?", PA1 manifests she is dealing with new information, on which she did not previously rely to interpret ROB's conduct. This is confirmed when, after this turn, she immediately casts the robot's previous answer as inadequate by repeating her own previous question – while insisting on the term which was stated, by PA2, to be mistranscribed. This re-evaluation of ROB's conduct illustrates how much the social meaning of verbal and embodied responses from the robot can be contingent upon what the robot transcribes on its tablet. Here, the ASR transcript is used as a resource to reevaluate the relevance of the second pair part produced by the robot.

Crucially, this re-evaluation is not about the *type* of response that ROB does. In this fragment, ROB's conduct *grants* a request: at this level of description, on the surface of the interaction, ROB produces an adequate second pair part. However, this action (of granting a request) is reevaluated as *indexed to the wrong speaking turn*. Therefore, rather than speak about reevaluating an "action" in a generic sense, we should say that the *type of social action produced* is not problematized: what is reevaluated is this action as *activity-relevant*, i.e., its *status as a response to what was really requested by human participants*. In this case, it is not problematized as a response to a request but as a *response to a specific request*.

This fragment also displays how two alternative understandings of the same scene can differ, based on the (non-)use of the ASR text. PA2 characterizes ROB's answer as responding to a turn that was not produced by PA1: by informing PA1 that ROB transcribed "laver" instead of "lever", and by stating that she did not "speak loud enough", he treats PA2's

original question as a source of trouble connected to ROB's mismatched transcript. In the meantime, without the information about the ASR transcript provided by PA2, PA1 treated ROB's answer ("I do not have this functionality") as a relevant second pair part to her question: she did not attempt to repair the interaction, nor characterized ROB's response as a breach of the conditional relevance established by her question, nor as a contribution inadequate with the activity currently at hand. From the external point of view of the analyst, who has access to what the robot heard, the beginning of this fragment therefore displays an example of a misunderstood speaker who "do not grasp that they have been misunderstood" (Schegloff, 1992c).

7.8. A wide range of situated uses for the ASR transcript

7.8.1. Troublesome ASR transcripts could override the verbal and gestural response of the robot

The previous fragments can be organized on a continuum on which the ASR is treated, on one extreme, as a mere receipt of information (Svennevig, 2004) from which the robot will generate its response and, on the other extreme, as a full-fledged speaking turn with a meaning of its own. On the first side of this continuum (fragments 7.1, 7.2, 7.3, 7.7, and 7.8), the transcript displayed on the tablet is an additional resource that "gives off" (Goffman, 1959) what the robot heard or, depending on the participants, understood from the human's previous turn. On the other side (fragments 7.4, 7.5, and 7.6), the ASR is treated as "saying something", i.e., as purposely communicating information that is not a simple transcription of what the robot heard.

A recurring feature of the previous fragments is that the content displayed in the ASR transcript was often responded to or used for a public performance, while the speaking turn produced by the robot was not dealt with. In sum, in case of trouble, the ASR transcript – more than drawing attention to the robot's tablet in purely quantitative terms – was often connected with a *treatment of other modalities of expression of the robot (gestures, speech, etc.) as less interactionally relevant*. This lower relevance is manifest in fragments 7.1, 7.2, 7.3, and 7.5 where the robot's gestural or hearable conduct is not responded to (fragments 7.3 and 7.5) or is overlapped (fragments 7.1 and 7.2) by participants' reaction to the ASR transcript. Such conducts manifest an orientation to the ASR transcript as a *sufficient* contribution from the robot for participants to produce a response – which the verbal and gestural reaction from the robot will not re-configure a posteriori. In situ, the text displayed on the tablet "overrides" the robot's embodied and verbal response: it is given priority over other behaviors from the robot as an "answerable" or, at least, as a "noticeable" conduct. The production of a verbal answer from the robot does not interrupt the reactions that follow the appearance of the ASR transcript, and, through this, does not reconfigure the manner in which participants demonstrably orient to the situation. This is unlike the symmetrically opposite situation, visible in, e.g., fragment 7.8, where new information about the ASR transcript reconfigured the main speaker's interpretation of the previous verbal conduct from the robot. In most of the previous fragments, the ASR transcript gives participants access to information which are exceptionally

relevant to the practical problems (the “purposes at hand”; Garfinkel, 1967) they encounter in their situated interactions with the robot.

These observations connect with data obtained from our eye-tracking device. The limited tendency to look at the ASR more and more as the interaction went on (see section 7.6.4) may correspond, in many cases, to a transition from a *human-robot interaction*, in the sense of a treatment of the robot’s vocal and embodied responses as the most relevant features of the interaction, to a *human-to-written-transcript* interaction, in the sense of a treatment of the text displayed on the tablet as the main parameter from which to build an answer for participants – often overriding anything the robot might say or do afterward.

7.8.2. The specific situated functions of the ASR transcript

Because of the informational configuration of human-robot interactions featuring a transcript, we argue that the diverse situated functions of the ASR transcript were frequently *sui generis*. It was often impossible to rigorously overlap the interactional relevance of Pepper’s tablet transcript with the relevance given to resources typically featured in face-to-face human-human interactions. As mentioned previously, the ASR transcript was regularly treated by participants as a non-verbal “display of hearing” (Svennevig, 2004) or of understanding. Consequently, since humans do not display what they hear on their forehead, participants interacting with a Pepper robot comparatively had access to an additional element that was specific to human-robot interaction – and which they often treated as crucial information.

For example, fragment 7.2 is a typical case of repair produced when the robot barely started to respond, i.e., before “an understanding of at least the general thrust of the utterance can have been achieved” (Jefferson, 1982). Yet, the human participant in fragment 7.2 is able to pinpoint precisely which word has been misheard by the robot. As we attempted to show in section 7.7.1.5, the repair produced by this participant in fragment 7.2 (L.13) differs from phenomena identified in human-human interactions. It does not respond to the ASR transcript as a mere embodied “demonstration of understanding” (Heritage, 2007) from the robot: that is, the participant does not *infer* the existence of a misheard word from external cues but can immediately *notice* which word was misheard. Simultaneously, the timing of this repair does not match a reaction to a “verbal display of hearing”: because of its scriptural nature, the ASR transcript is displayed simultaneously with the turn-at-talk produced by the robot. This configuration was consequential on the timing of the repair initiation and on its packaging. It impacted which syllables were stressed, what terms were identified as troublesome in the participant’s previous turn, and how fast they were repaired. In other words, it changed the “precise moment when the point of a turn or action becomes recognized in the course of its very production” (Deppermann et al., 2021). As a result, the repair analyzed in fragment 7.2 – and virtually any repair in human-agent interactions that relies on information provided by an ASR transcript – stems from a configuration from which one of the most consequential traits is absent during daily face-to-face human-human interactions.

Hence, the tablet transcript was often mobilized by human participants in ways that did not perfectly overlap with how typical human-human informational resources are mobilized in face-to-face interactions. This is mainly linked to two parameters mentioned above: first, the scriptural nature of the ASR transcript (which can appear at the same time as the gestural and

verbal answer of the robot); second, the use of the ASR transcript by some participants as an “internalist window” on what the robot heard or understood.

7.8.3. The ASR transcript as a publicly noticeable phenomenon

Among the previous fragments which featured several participants, a common feature is the work deployed by the main human speaker to make publicly relevant the ASR transcript as a noticeable feature of the interaction. That is, either as an interactional contribution “given” (Goffman, 1959) by the robot, or as a relevant cue that the robot “gives off” (Goffman, 1959). When main speakers were the ones originally identifying the ASR as troublesome (here, fragments 7.1 to 7.5¹⁵⁵), they typically communicated it as relevant and remarkable through the following format of participation shift, which can be divided into 4 steps:

1. Reading the ASR out loud (fragments 7.1, 7.3, 7.5)
2. Doing a body torque towards co-present participants (fragments 7.1, 7.3, 7.4, 7.5) and/or taking a step backward (fragments 7.1, 7.3, 7.4, 7.5)
3. Hearably or gesturally orienting to the ASR as remarkable: by laughing (fragments 7.1, 7.3), pointing towards the tablet (fragments 7.1, 7.3, 7.4, 7.5), and/or producing a report of the ASR (fragment 7.4)
4. Commenting on the robot’s functioning or on its understanding of the situation (fragments 7.1, 7.3, 7.5), often while referring to it in the third person (fragments 7.3, 7.5)

The exact embodied and linguistic format of this participation shift is extremely prevalent in our corpus when the main speaker publicly thematizes the ASR transcript (in front of all co-present third parties). However, these side-sequences are not specific to our corpus: similarly brutal shifts in participation (excluding the robot) towards third parties have been noted by, e.g., Due (2019) about interactions with a Pepper robot in a corridor, or by Krummheuer (2015) based on interactions with an embodied conversational agent in a mall¹⁵⁶. When produced while orienting towards an ASR transcript as troublesome, this format served, among other endeavors, to maintain intersubjectivity in the group about what action the robot was doing or what it “understood” or “heard” – especially when the robot’s verbal or gestural behavior might have been complicated to make sense of for all participants. Through it, the main speaker made public the relevance of the content of the tablet for the ongoing interaction and, simultaneously, enforced their interpretation of the robot’s conduct.

7.8.4. The crucial relevance of the ASR transcript for the robot’s own perception of the situation

The impact of the ASR transcript on participants’ conduct highlights that, from the point of view of the robot itself, considering the ASR transcript is crucial to make sense of the interactions in which it is involved. In fragments analyzed in this work, human participants’ reactions are

¹⁵⁵ Fragment 7.2 was not a multiparty interaction.

¹⁵⁶ See section 6.4.1.2 of this manuscript for an extensive analysis of the regular side-sequences produced by human participants

only fully understandable if the presence of an ASR transcript in front of them is taken into account. For example, in fragment 7.3, it is not possible to understand participants' choral utterance of "le corps vide" as a repetition (projecting an "intention to comment upon the other's talk"; Perrin et al., 2003) without taking in account what is written on the robot's tablet. If a robot were to attempt to interpret a participant's behavior without this crucial contextual element of "what it displayed on its tablet", it would often miss one of the most consequential features of these situated interactions. In other terms, it is critical for the robot to be able to identify when the conduct of human participants (repairs, accounts, repetitions, etc.) is not produced in reference to *something it said* (i.e., it is not an answer to a previous speaking turn) but, rather, in reference to *something it is currently displaying*. This applies even in situations when the ASR transcript is not oriented to as a transcript by participants, but, instead, as "what the robot intended to say" (e.g., fragment 7.4) – and more generally, any situation where the ASR transcript is used as a resource to advance the interaction.

These observations therefore add to the list of key parameters that embodied agents should perceive (Vinciarelli et al., 2009) and consider to better understand what human participants are indexing in their speaking turns. As Tuncer et al. (2023) state, "a robot should not produce actions that make relevant next actions it will not be able to respond to", which, among other prerequisites, implies that the robot only "produce[s] actions which it is itself capable of understanding" (Tuncer et al., 2023). In the situations we studied and with the conversational software we used, the ASR transcript contravened Tuncer et al.'s (2023) guideline: the display of a transcription of what the human said was an action from the robot that the robot was unable to make sense of.

Yet, the relevance of the ASR transcript to participants also relates to the typical design question: what should a robot display of what it grasps from the world? Should the robot function as a complete "black box", or should it make public some of the information it uses to generate its actions? For example, should the robot display how many humans it currently sees, the confidence score attributed to the presence of each human, or even the confidence score attributed to each of the ASR results it shows? A specific answer to this question can be found with autonomous cars (Brown et al., 2023), which usually provide passengers with feedback about what they "see" of the outside world: cars, pedestrians, motorcycles, lanes, etc. – that is, one possible discretization and representation of these data. Significantly, there can be no "neutral" or "hands-off" response to this inquiry. Any form of perceptual data made available by a robot during an interaction (raw data from its sensors or processed data) is a choice of design with potential interactional consequences: what the robot makes publicly available is always one way (among many others) to select, process, and represent information it uses in local interactions. Rather, the answer regarding "what to display" may need to be sought *in the pragmatic consequences of information made available to humans on the interaction itself*.

7.9. Some theoretical implications for mutual understanding

7.9.1. Sense-making and sequential adjacency

The previous discussion on the pragmatic impact of an ASR transcript raises questions about the inner workings of human-human and human-agent understanding. To what degree are face-to-face interactions based on a form of “wobble room” about what the other interactant heard or understood from our own behavior? How important is a form of interpretive fuzziness in smoothing human-human interactions and in preserving participants from being faced too often with ostensible breakdowns in intersubjectivity – which participants then have to repair or, on the contrary, in spite of which they have to “go on”? Would a hypothetical *internalist window* be damaging overall for what is typically treated as a satisfying face-to-face conversation between humans? In sum, *what would happen if humans automatically displayed on their forehead what they heard from other interactants?* Specifically, does having access to the phonological reconstruction of the utterance that a turn responds to modify the relevance of this turn as an adequate situational response (if the phonological reconstruction is erroneous)? By disrupting taken-for-granted ways of communicating with another speaker (among which the fact that, in human-human interactions, participants rarely produce “displays of hearing”; Svennevig, 2004), interactions with a Pepper robot may shed light on some of the inner workings of face-to-face communication.

Garfinkel’s early experiments (Garfinkel, 1967; Ivarsson & Lindwall, 2023) highlighted the demonstrable tendency of participants to interpret immediately adjacent talk as responsive to their own and as displaying an understanding of their own conduct – even when, from a technical point of view, the agent cannot be responsive to what the human is doing. In a classic study, Garfinkel (1967) studied how subjects made sense of an interaction where their questions were responded to by “yes” or “no” but that, unbeknown to these subjects, these answers were triggered randomly by a researcher. In spite of the disconnection between the provided response and the subjects’ verbal conduct, these answers were treated by subjects as “motivated by the intended sense of the question” which had been asked (Garfinkel, 1967). Similar observations have been made about interactions with even the earliest “conversational lures” (Relieu et al., 2020) like the chatbot ELIZA. In the same study, Relieu et al. (2020) documented how the chatbot LENNY produces behaviors that “mould in different sequential trajectories”, despite the total absence of real-time adjustments from this chatbot to what the caller is saying. The design of LENNY’s speaking turns (even though these “turns” are, from a technical point of view, pre-recorded and pre-ordered) facilitates their treatment as relevant contributions to the ongoing interaction by human callers.

This tendency observable in human-agent interactions – to treat some behaviors as responsive contributions to the activity at hand – might be a product of the inner workings of conversational mechanics. More precisely, the previous observations are congruent with the stance following which, in typical human-human interactions, “understanding is not treated as a mental process but is related to the next action achieved by the co-participant and demonstrating her understanding” (Mondada, 2011). Because it “represents one of the most elementary bases of intersubjective understanding” (Relieu et al., 2020), sequential adjacency is heavily at play in successful interactions with conversational agents, even when, from a technical point of view, the conversational agent does not react to what its human interactants

are saying (Relieu et al., 2020). As Tuncer et al. (2023) remark about interactions with a Pepper robot, “[p]articipants pre-suppose that their actions are organised sequentially, and so they expect the robot to also organise its conduct likewise” (Tuncer et al., 2023).

7.9.2. What is the minimum “interpretative latitude” required in human-robot interactions?

In his aforementioned experiments, Garfinkel (1967) notes that, when participants in his experiment faced (random) “yes” or “no” answers to their questions, “[t]here was vagueness (a) in the status of the utterance as an answer, (b) in its status as an answer-to-the-question [...]” (Garfinkel, 1967). By comparison, as we observed in fragments 7.1, 7.2, 7.3, 7.7, or 7.8, the presence of an ASR transcript was a parameter that contributed to removing the vagueness of the robot’s responses “as an answer-to-the-question”: the exact phonological form of what the robot was responding to was apparent (to the main human speaker or, at least, to a bystander) and, when this phonological form differed from what the speaker actually said, a repair or an account was generally produced. In this sense, the ASR transcript had the (intended) effect of reducing the interpretative latitude (Heritage, 1989) of participants as to what the robot was answering to. Yet, doing so, it may have constituted an unexpectedly radical change to the inner mechanics of conversation. Systematically producing conduct akin to a “proof of hearing” disrupted the usual working of social encounters, where it is expected that “a recipient of one’s remarks will, ‘without requirement of a checkout’ understand what was intended by the remarks” (Maynard, 2011). The ASR transcript offered such a “checkout”.

Some consequences can be drawn from these observations for conversational design. Provided that ASR transcripts are treated as displays of hearing by human participants, many typical conversational practices are likely to be difficult to perfectly replicate for a robot that displays transcripts. Independently from a conversational agent’s competence or technological advancement, systematically displaying an ASR transcript unavoidably complicates the accomplishment of “conversational moves” that rely on the absence of an internalist window on “what was heard”. For example, a conversational move like what Liberman (1980) calls “gratuitous concurrence” may, a priori, be hindered by the informational configuration associated with ASR transcripts. In Liberman’s (1980) definition, gratuitous concurrence is the action of providing an interlocutor with a “confirmation of comprehension” (Liberman, 1980) without having comprehended what this interlocutor has said: e.g., when a recipient produces the change-of-state token (Heritage, 1985) “oh” in response to an utterance that they did not understand (Liberman, 1980). Between humans, a useful resource for gratuitous concurrence is therefore the inaccessibility, for the other parties, of “what is being concurred with” in the speaker’s head. When a speaker concurs with prior talk, other parties have no access to obvious mismatches between what they said and what the speaker heard (the option relevant for us in this work) or understood from what was heard. In the case of a mishearing, the conversational “move” of gratuitous concurrence is therefore not always replicable for a robot displaying a transcript (e.g., by responding “yes” to a sentence that it did not hear entirely) because part of the internal state of the robot is publicly visible. For such a robot, being able to pass over an ambiguity about the phonological identification of what the human said (by producing an alignment token as a form of gratuitous concurrence) requires that its interlocutor either disregards the information of the ASR transcript or orients to it as something else than a transcript. Similarly, laughing to camouflage having completely

misheard what an interlocutor said (a strategy often depicted in the internet online culture¹⁵⁷) is unlikely to have the intended effect in this informational configuration: the blank ASR transcript provides readily accessible proof that the conversational agent did not, in fact, hear anything. Even in the heuristic hypothesis of a conversational AI sufficiently advanced to be, strictly speaking, purportedly “ambiguous” about the actions it is producing – unlike the robot we used, which can merely be said to produce actions that are treated as ambiguous by human participants –, the aforementioned “moves” would not produce the same effect when the agent displays an ASR transcript that is treated as such by its interlocutors. That is, any ambiguity regarding the strip of words indexed by the conduct of this agent would not exist and, therefore, not be available as a resource for the robot.

Hence, we argue that, in the presence of a systematic ASR transcript, what “interacting with a robot” consists of practically is pulled even further away from the concrete processes that constitute human-human interactions: the transcript impacts the methods that are (or can be) used, the pragmatically relevant properties of the interaction, etc. A design dilemma (as well as a moral dilemma) stems from this state of affairs: should conversational agents prevent misunderstandings at all costs? Or should they act similarly to typical human interlocutors? In sum, do we want conversational agents to be able to use the same conversational moves as humans? Beyond questions regarding the ideal “interpretative latitude” between robots and humans to facilitate the progress of the talk, it is worth considering if, for ethical reasons, the robot should be provided the same “tools” (e.g., indeterminacy about what it really heard) than humans to maintain the sequential relevance of its turns – and, through this, its perceived conversational competence.

7.9.3. Impact on the “sequential plasticity” of the robot's turns

In human-robot interactions, many behaviors from a robot treated by co-present members as a successful second pair part are, from the point of view of the engineers who programmed this robot, the result of an unforeseen sequence of events in which the robot (Pelikan et al., 2020; Rudaz et al., 2023) did not, in fact, adapt to the human. In other words, temporarily putting aside the perspective of members directly interacting with the robot, this robot is rarely able (in a technical and internalist sense) to accomplish the action it is treated to have accomplished by participants. Tuncer et al. (2023) similarly note that humans may orient towards an action from the robot “as responding to their own previous conduct and meaningful, notwithstanding the possibility that the action be fortuitous, and its appropriateness as a response a coincidence” (Tuncer et al., 2023).

Hence, when considering the regularly “fortuitous” character of the robot’s successful responses to humans’ actions, the ASR transcript limited the “sequential plasticity” (Relieu et al., 2020) of the robot’s conduct by clearly indicating – for the numerous participants who treated it that way – what sequence of words the robot was responding to. The resulting informational configuration narrowed down the range of behaviors from the robot which could be directly treated as *social actions* – that is, as responsive to the situated interaction and as “making relevant a set of potential next actions” (Tuncer et al., 2022a). We argue that these

¹⁵⁷ E.g., <https://www.instagram.com/p/CwxZSNorv-Y/> (Accessed November 2023). Meta-conversational content like this one provides a list of “moves” that are available to participants in an interaction after mishearing their interlocutor.

behaviors became less simple to treat interactionally as stemming from an agent (Pelikan et al., 2022) imbued with intentionality. That is, more (incongruent) parameters had to be taken into account during the interactional work accomplished by some participants to preserve the appearance of “an intentional self” (Alač et al., 2020) to the robot. In sum, because the ASR transcript limited the range of meaningful intentional patterns that could be connected with the robot’s observable behavior, such a context offered fewer resources for human participants to “safeguard the robot’s status as an agent” (Pelikan and al. 2022). If our interpretation is correct, some of the typical methods through which a robot is maintained as a competent participant in a public setting (detailed in chapter 6 of this work) should require more work to be carried out when this robot displays an inaccurate transcript of the previous utterance: for example, when the robot clearly misheard an instruction, it may become more intricate for humans to attribute a mental state to it to account for a “missing action” on its part (Pelikan and al. 2022).

In this sense, the “speechbar” feature of the Pepper robot (which displays an ASR transcript in its default settings) contrasts sharply with the ideas behind the Cozmo robot’s sound design analyzed by Pelikan et al. (2020): the Cozmo robot’s sound was intentionally designed to “blur the boundaries between failure to act and intentional refusal to act” (Pelikan et al., 2020). As we saw, through the ASR transcript, instead of hiding the absence of a cognitive “intersubjectivity” (or at least the absence of a correct receipt of human participants’ speech) behind behaviors able to “mould in different sequential trajectories” (Relieu et al., 2020), the Pepper robot provides additional resources to make these breakdowns ostentatious and interactionally relevant. Doing so runs the risk of submerging human interactants with (more or less important) “repairables” which, otherwise, might have gone unnoticed.

7.9.4. Does the ASR transcript enforce a cognitivist definition of mutual understanding?

Based on the previous discussion, the orientation of many of our participants towards the ASR transcript can be described in two ways. A safe manner of verbalizing our participants’ practices is to say that it became a members’ problem that the ASR transcript displayed an adequate receipt of their previous turn. The performance that was regularly expected from the robot was not only a gestural and a verbal one. It was also an “auditive” one. During their interactions, it generally mattered to participants when the robot displayed a too-dissimilar transcript of what they said. Yet, in the EMCA endeavor to represent as precisely as possible participants’ own emerging categories during their situated activity, another level of description might be more faithful to the orientation displayed by many of our participants towards the ASR transcript: they were led to enact a cognitivist definition of mutual understanding. In other words, participants cared about what was inside the robot’s “head” or “algorithm”.

More precisely, as a technological artifact, the ASR transcript reifies a specific representation – possibly shared by the designers of the robot – of what human-human understanding is (in a conversation), which stands at the other end of the spectrum compared to the way it is conceptualized in ethnomethodology and conversation analysis. If, as Suchman (1987) mentions, “[e]very human tool relies upon, and reifies in material form, some underlying conception of the activity that it is designed to support”, the speechbar enforces a conception

of “understanding”¹⁵⁸ as sharing a “similar mental representation about the world” (Albert & Ruiter, 2018) – rather than, as it is generally described in conversation analytic works, “related to the next action achieved by the co-participant” (Mondada, 2011; Mondémé, 2022)¹⁵⁹.

We argue that the ASR transcript – which stood out in front of participants each time they stopped speaking – encouraged an orientation to “what was inside the robot’s head” as interactionally relevant. For the majority of observed participants, the ASR transcript facilitated a cognitivist definition of mutual understanding “as a mental process” (Wittgenstein, 1953), i.e., as “grasping what is in the other’s mind” (Shotter, 1996). It did not only matter for participants that the robot produced verbal and gestural answers that fit what was just said, *it also mattered that these verbal and gestural answers stemmed from an entity that had properly heard what they said*. In order to ascribe actions to that entity, in order to grasp what it was “doing”, an element of the cognitive state of that entity mattered (what phonological reconstruction it had produced of the previous talk), possibly more than its verbal and gestural response. For those participants, it favored an orientation to progressivity in conversation as based on a perfectly shared mental reality¹⁶⁰, rather than as “being able to ‘go on’ with each other” (Sterponi & Fasulo, 2010) or “to progress with whatever [participants] are doing together” (Albert & Ruiter, 2018)¹⁶¹.

Verbeek (2006) notes that “[m]edical imaging technologies, such as magnetic resonance imaging and ultrasound [...] make visible parts of the human body, or of a living fetus in the womb, that cannot be seen without them. But the specific way in which these technologies represent what they ‘see’ helps to shape how the body or a fetus is perceived and interpreted and what decisions are made.” Similarly, in the previous fragments, the ASR transcript made visible a specific aspect of the robot’s perception of the world – however, unlike the ultrasound, without hiding the other dimensions from view. Through some participants’ verbal and embodied behaviors (pointing at the tablet, repeating what was written, etc.), this display from the robot of its perception of the word became a public and a consequential phenomenon: when they faced an erroneous ASR transcript, many participants attempted to repair this transcript before the robot produced a verbal and gestural answer

¹⁵⁸ Ivarsson & Lindwall (2023) remark that “[t]here is no point in imparting the analytic registers of HCI research with conceptual ambiguities by building technical terminology out of vernacular expressions”. Speaking about “understanding” or even “hearing” falls within the scope of this remark on vernacular expressions. Yet, we argue that, here, what the robot “understands” or “hears” is an adequate description of what many participants, emically and locally, were concerned about. Since participants used these terms (when stating that “the robot understood” or “heard” something), we consider them as valid categories.

¹⁵⁹ This definition of coordinated action is also used by researchers who focused on interspecies interactions where “no matter if members share similar mental states, what matters is that the orderly phenomena which emerge from their actions are identifiable and intelligible, i.e. that they are supports for practical action” (Mondémé, 2019; my translation). This non-cognitivist definition is especially efficient at explaining how we can coordinate with entities whose mental categories or *umwelt* are potentially completely alien to ours (Mondémé, 2016b).

¹⁶⁰ If our analysis is correct, this intention is as intentionally “inscribed” (Akrich, 1989) as can be. As mentioned previously, the official documentation of Aldebaran on the “SpeechBar” (a feature that includes the ASR transcript) justifies its existence by clearly stating the relevance of this set of information “to understand what’s going on”: “Providing feedbacks is essential to be sure that Humans really understand what’s going on when talking with Pepper. A bar, called SpeechBar, is displayed on the tablet. It helps to understand if the robot is listening, hearing something and what has been understood”. This documentation is available here: https://qisdsoftbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/conversation/conversation_feedbacks.html#conversation-feedbacks

(fragment 7.2) or after the robot’s answer – even if this answer had been treated as relevant so far (fragment 7.8). In the interpretation that we tried to sketch out, these participants appeared to display an orientation towards the existence of a form of “shared reality” (at least regarding the phonological identification of what they said) as required to progress the interaction.

Meanwhile, in the second condition of our experiment, where no ASR transcript was available, participants did not produce conduct in which “that the robot accurately heard or understood what was said” was treated as a condition to continue with the activity. This is strikingly different from what we saw in previous fragments, where the presence of an ASR transcript – in addition to being associated with a higher focus of participants’ gaze attention on the tablet – induced a treatment of other modalities of expression of the robot (gesture and speech) as less relevant for the task at hand. The ASR transcript was treated as a central phenomenon to coordinate the ongoing activity, while the head, body, and voice of the robot were sometimes disregarded. In other words, displaying the *phonological reconstruction of the utterance that a turn responded to* impacted *the relevance of this turn as an adequate response*. Because of this increased focus on the tablet and its interactional relevance at specific moments of interactions, the ASR transcript fits the category of what Verbeek (2006) names “mediating technologies”, that is, a technological artifact that “facilitates people’s involvement with reality, and in doing so, [...] coshapes how humans can be present in their world and their world for them.” (Verbeek, 2006)¹⁶².

7.10. “Interactional work” and the ASR transcript

7.10.1. Does the ASR transcript increase the “interactional work” required from human participants?

We argue that the informational configuration connected to the presence of the ASR transcript complexified the task of engaging in a smooth interaction for humans, as it clearly manifested any small deviation to a cognitive conception of mutual understanding. Each time “what was heard by the robot” differed from what was publicly said by the human, this information was accessible by said human, independently from what the robot’s verbal and gestural response would be. Through the robot’s ASR transcript, participants had constant access to deviations to (cognitive) intersubjectivity. We suggest that these participants had to produce additional work during the ASR condition, because, in this condition, (perceived) misunderstandings could less easily be left tacit as non-consequential “for current practical purposes” (Albert & Ruiter, 2018; Linell & Lindström, 2016; Schutz, 1972). In many cases, the ASR transcript could be said to have troubled participants’ practical sense of what was a disregardable non-significant mismatch to reciprocal alignment (Shotter, 1996; Sterponi & Fasulo, 2010) or, conversely, of what was a significant misunderstanding which should be addressed – even at

¹⁶² This design is extremely rare in social robotics. As we remarked in section 7.4, unlike the Pepper robot, most “social” robots do not provide users with a public ASR transcript (e.g., the Cozmo robot analyzed by Pelikan et al., 2020). The previous analysis suggests that this difference in design may reveal a deeper conflict between two conceptions of coordinated action among designers themselves: as having similar representations of the word or, conversely, as “being able to ‘go on’” (Sterponi & Fasulo, 2010) together.

the cost of a “time out” in the current course of action. In other words, the ASR transcript modified “the limits of what [the parties in a conversation] will seek to bring to determinacy” (Lieberman, 1980). When the ASR transcript was treated as a display of hearing, “what has been heard by the robot” was not information that could be inferred (Deppermann, 2018) or documented (Garfinkel, 1967) from the robot’s verbal and gestural conduct: it became an unambiguous, immediately accessible fact. Even in the case of a response from the robot treated (so far) as adequate, the ASR transcript made apparent each situation where this response was produced based on a misheard input (e.g., fragments 7.7 & 7.8). The ASR was, therefore, a crucial constituent of “action recognition (Schegloff, 2007) or “action ascription” (Levinson, 2012). Displaying an ASR transcript, by clarifying what a turn responded to, clarified (not in principle, but in our empirical observations) what conduct was produced through this turn: the ASR transcripts added to the many “features of context and design that interactants use in the action ascription process” (Stivers et al., 2023). It is possible that, faced with such unambiguous mishearings, participants were confronted with the additional choice of dealing with these mishearings in any given manner or to pass over (Lieberman, 1980) them, i.e., to “let it pass” (Garfinkel, 1967).

7.10.2. A new parameter to “work with” during tensions between progressivity and intersubjectivity

Hence, the previous fragments may display different practical solutions to the tension, well established in human-human interactions, between progressivity (Schegloff, 1979b; Stivers & Robinson, 2006) and intersubjectivity (Heritage, 2007). In human-human interactions members’ actions generally hint at a preference for “advancing in-progress activities through sequences” (Stivers & Robinson, 2006). When trouble arises in “speaking, hearing and/or understanding the talk”, members act in a way that prevents the interaction from “freez[ing] in place” (J. E. Fischer et al., 2019; Schegloff, 2007) so that the “turn and sequence and activity can progress to possible completion” (J. E. Fischer et al., 2019; Schegloff, 2007). Significantly, this preference for progressivity was observed in human-agent interactions by J. E. Fischer et al. (2019), who partially transferred Stivers & Robinson’s (2006) framework onto interactions with the vocal user interface Alexa. They observe that many of the manifestations of progressivity identified by Stivers & Robinson (2006) can be found within interactions with this vocal user interface: for example, participants “work” to receive and provide answers to Alexa, participants offer accounts when Alexa does not provide an answer, etc. (J. E. Fischer et al., 2019). Yet, this preference for progressivity can conflict with another documented preference: the maintenance of intersubjectivity and common ground (Heritage, 2007; Schegloff, 1992c), which, in conversation analysis, is conceptualized as the “communicative achievement of mutual understanding” (Sterponi & Fasulo, 2010) rather than as “a condition for communication” (Sterponi & Fasulo, 2010). In this acceptation, intersubjectivity and mutual understanding correspond to “being able to “go on” with each other” (Sterponi & Fasulo, 2010) rather than being able to grasp “what is in the other’s mind” (Sterponi & Fasulo, 2010).

In most formats, repairs, as resources to maintain intersubjectivity (Schegloff, 1979b, 1992c), bring a “time out” (Heritage, 2007) from the ongoing conversational activity: when initiating a repair, participants temporarily interrupt the progressivity of the current turn or sequence (Heritage, 2007), since initiating a repair “suspends the ongoing course of action” (Bolden, 2012; Schegloff et al., 1977). Therefore, consistently with a preference for

progressivity, unnecessary repairs tend to be avoided: “as long as the verbal interchange evolves relatively smoothly, participants do not engage in metalinguistic commentaries or repair” (Linell & Lindström, 2016). That is, *participants do not initiate repair as soon as a pre-existing and spotless state of intersubjectivity starts to deteriorate even in the slightest*; they initiate repair when their shared understanding is not sufficient for their “current practical purposes” anymore (Linell & Lindström, 2016; Schutz, 1972). For example, when participants refer to persons or places, Heritage (2007) notes that “recognition is assumed unless some form of trouble is indicated. In other words, recognition is treated as the 'default' condition”. In their everyday activities, members are not concerned “with attaining absolute terminological precision as in certain scientific genres” (Linell & Lindström, 2016). The practical problem participants generally face is not to keep, at all costs, “sufficiently similar mental representations about the world” (Albert & Ruiter, 2018) but “to progress with whatever they are doing together” (Albert & Ruiter, 2018). In other words, while they are immersed in the urgency of a social interaction, participants are primarily concerned with how to advance this interaction: unless it is specifically made relevant in regard to the task at hand, “what is inside the head of their interlocutor” is usually not a members’ problem when they cope (H. L. Dreyfus, 2001) with daily social encounters.

7.10.3. ASR transcripts as a systematic threat to (assumed) common understanding

In the specific informational configuration of human-robot interactions that featured transcripts, this tension between progressivity and repair was put to the test through the regular mishearings detected on these ASR transcripts. In particular, as we saw in fragments 7.1, 7.2, 7.3, 7.7, and 7.8, the robots’ ASR feedback was regularly treated as a form of “display of hearing” (Svennevig, 2004) or – depending on the participant’s interpretation – as a “claim of understanding” (Heritage, 2007; Sacks, 1968) of what was previously said. By displaying (or failing to display properly) the previous speaking turn of a participant, the robot indicated at the very least¹⁶³ its (in)adequate identification of the phonological properties of the participants’ utterance (Svennevig, 2004).

By constituting a recurring indicator (often overriding other verbal and gestural cues) about the existence of a breakdown in “intersubjectivity”, the ASR was both a resource and a new set of opportunities for participants to repair the interaction at each erroneous transcript. Following Schegloff’s (2007) canonical definition of progressivity as “moving from one element to a hearably-next-one with nothing intervening” (Schegloff, 2007), the ASR feedback was frequently oriented to as something which intervened “between some element and what is hearable as a/the next one due” (Schegloff, 2007). After each utterance produced by a human, the robot displayed information susceptible to revealing the previous turn of the human as “repairable” or, at least, as inadequately heard or understood – before the robot itself produced

¹⁶³ Indeed, even if, among humans, repeating the previous turn is often a “claim of understanding” (Heritage, 2007; Sacks, 1968) rather than a demonstration of what was understood, the act of properly repeating a turn is, at the minimum, a demonstration of having properly heard the previous speaking turn. Since a repeat is a common practice for claiming understanding (Bertrand & Goujon, 2017), some of our participants may have treated the ASR as a “claim” of understanding rather than as a mere display. However, even in such cases, the packaging of this “claim” relied on a repetition of the immediately preceding speech (and was treated as such by our participants when they produced a repair): it always constituted simultaneously a display of hearing.

a verbal or gestural response. In other terms, while demonstrations of understanding are not the “default condition” (Heritage, 2007) in human-human interactions, and displays of hearing are also rare (Svennevig, 2004), the ASR transcripts – systematically – provided by our Pepper robot were regularly treated by participants as such displays of hearing.

Consequently, the ASR transcript may constitute an exception to a “maxim of minimization” (Levinson, 1987), as argued by Heritage (2007) about demonstrations of understanding. Fragments 7.1, 7.2, 7.3, 7.7, and 7.8 constitute local examples that attempting “to produce unambiguosness” (Meyer, 2019) can “unnecessarily complicate the practical situation” (Meyer, 2019). Of course, displays of hearing are occasionally produced by vocal interfaces other than the Pepper robot – but not in a systematic way. J. E. Fischer et al. (2019) provide an example where Alexa “makes the speech-to-text transcription of [a speaker]’s prior request hearably available to [this speaker]”. However, in J. E. Fischer et al.’s (2019) example, rather than a default behavior triggered independently from the content of the speech that Alexa heard, this specific hearable ASR transcript responded to a turn characterized (by the vocal interface) as a request to “hear a station”. Hence, in the sequential context described by J. E. Fischer et al. (2019), the (erroneous) transcript constituted a relevant resource for participants to resolve the ongoing breakdown in communication, i.e., it could be used by participants to produce an account of “what went wrong” (J. E. Fischer et al., 2019) in the previous exchange. Conversely, based on the previous analysis, we suggest that, in a conversational context, systematic and non-context-sensitive ASR transcripts hinder intersubjectivity enacted as a “situated, temporarily sustained and only partially shared experience” (Linell & Lindström, 2016). Indeed, as Linell & Lindström (2016) suggest, experiencing intersubjectivity in this manner is dependent on the fact that, in human-human interactions, “a single contribution within a reasonably coherent sequence presupposes an understanding of the prior contribution(s)” (Linell & Lindström, 2016). The adequate understanding of a previous contribution is usually displayed “indirectly in the design of the next relevant action” (Svennevig, 2004). Yet, in many of our observations, the display of ASR transcripts removed this essential presupposition of one’s previous contributions as sufficiently understood: when the robot responded verbally, the ASR could override the relevance of this response *as a response to a specific turn*. In this new configuration, the “weave of interactional moves” (Linell & Lindström, 2016) produced between humans and robots was less likely to offer the (superficial) coherence which is the bedrock of an experience of the ongoing interaction as intersubjectively shared.

7.11. Design recommendations

7.11.1. Preventing from using ASR transcripts in leisurely conversational interactions

Based on the previous observations, we recommend not activating ASR transcripts on a humanoid robot primarily involved in purely leisurely conversational interactions or in small talk. In these conditions, a conversation with a multimodal humanoid robot risks not being treated as fully multimodal by many participants. The analysis of quantitative eye-tracking data indicates that the narrow zone of the tablet where the ASR transcripts appear occupies a

significant part of participants' attention. Congruently, the analysis of qualitative fragments of typical uses of the ASR transcript highlights the overwhelming relevance of the transcribed ASR for the participants in construing their interpretation and their response to the robot's conduct. More than a parameter that participants merely use, among other multimodal cues, to clarify and repair breakdowns in interaction, the ASR transcript often *overrides* the subsequent verbal and gestural conduct of the robot – which are often disregarded or overlapped by participants as they respond to the content displayed on the tablet. In sum, whether we study participants' attention or what they treat as socially relevant in situ, the ASR transcript diverts their focus from the robot's verbal and embodied behavior towards the scriptural content displayed on the tablet. As we mentioned in section 7.8.1, the presence of the ASR induces a transition from a *human-robot* interaction towards a *human-to-written-transcript* interaction.

7.11.2. Displaying a non-systematic written transcript

An alternative solution to the complete absence of an ASR transcript would be a transcript displayed only after specific utterances, i.e., not mechanically after each speaking turn from a human. For example, VUIs like Alexa (J. E. Fischer et al., 2019) sometimes provide “accounts of what went wrong” (J. E. Fischer et al., 2019) by repeating exactly which terms they recognized in a prior utterance – although such verbal accounts differ on several points from a (written) transcript that appears simultaneously with the robot's verbal answer (see section 7.8.2). Of course, the optimal – “easier said than done” – solution would be an ASR transcript displayed *only when it is interactionally relevant for co-present members themselves*. Indeed, one of the core issues with the ASR transcripts that we analyzed here is their invariable appearance no matter the situation at hand. They were displayed by our Pepper robot after each utterance from a human, independently from the situational relevance of showing them.

Leveraging the similarity between the interactional function of a written ASR transcript and the function of a verbal “display of hearing” (Svennevig, 2004), a non-systematic and situationally relevant ASR transcript would be coherent with what is usually done in human-human conversations. Indeed, as mentioned in section 7.8.2, humans do not systematically produce displays of hearing by repeating their interlocutor's previous turn. Rather, these displays of hearing are produced in response to specific local contingencies where they occupy a meaningful function for participants, who only “sometimes need to display explicitly that they have registered a piece of information” (Svennevig, 2004). For example, Svennevig (2004) indicates that, among other situations, displays of hearing become relevant when the previous contribution “does not project any further talk to come” and that, as a consequence, the adequate registering of the prior turn cannot be displayed “indirectly in the design of the next relevant action” (Svennevig, 2004) – as is usually the case. Alternatively, in different configurations, these information receipts can constitute a way for recipients to display their construal (Svennevig, 2004) of the main speaker's contribution while immediately giving the floor back to the said main speaker, etc.

As a consequence, fully preventing the risk of producing ASR transcripts that unnecessarily slow down the progressivity of the interaction – and, possibly, which transgress

Levinson's (1987) maxim of minimization¹⁶⁴ – would require that embodied conversational agents are able to recognize situations where producing a scriptural display of hearing is interactionally relevant. This would, of course, imply an extraordinarily precise understanding of complex interactional dynamics¹⁶⁵ whose feasibility is beyond the scope of this work. However, one could imagine a rougher solution, where an ASR transcript only appears when the robot understands information whose understanding is generally not possible to display “in the design of the next relevant action” (Svennevig, 2004): for example, *when being provided with someone’s name, with a date, or with someone’s schedule* (Svennevig, 2004). Providing an ASR transcript after such information would provide better odds that these transcripts will be locally relevant for human participants – although these ASR transcripts would be triggered by anticipation and not as the result of a detailed understanding of the local context by the robot.

7.11.3. The specificity of leisurely interactions with a robot in a public space

The previous analysis of the interactional consequences of the ASR transcript was achieved based on encounters where no task was pre-defined besides the conversation itself. Our experiment featured a natural and an experimental setting (see section 7.3). In the natural setting, participants freely stopped and spoke to the robot without any verbal or written instruction: they were not required to collaborate with the robot in any way. In the experimental setting, participants were asked to speak with the robot, but no specific result was expected from them or from the robot. By contrast, vocal assistants like Siri, Google Assistant, or Bixby (which also provide ASR transcripts) are regularly tasked with solving precise problems, involving references to persons or places: i.e., to get a list of restaurants in a given city, to get the name of the writer of a given song, etc. In these endeavors, the flawless identification of the content of the human speaker’s utterance becomes critical for the task at hand. A mistake in the reception of an instruction (e.g., misrecognizing the name of a city or a song) is likely to result in failure and/or be repaired by the human interactant. Moreover, our data are overwhelmingly composed of situations of “first encounters” with a humanoid robot, involving participants unfamiliar with this technology. Therefore, even the packaging of speakers’ turns may differ when they interact (repeatedly and over the long run) with VUIs: as Due & Lüchow (2022) note, over time “people attune the way they speak to suit the computational system”. In this general context, our findings about the ASR transcript – the orientations of participants to it, its use as a resource for different tasks by these participants, its consequences on the progressivity of the interaction – are not generalizable to non-leisurely interactions with a VUI. The same goes for the recommendations we produced. The “purposes at hand” (Garfinkel, 1967) in these different configurations might differ widely and, accordingly, so do the relevance and consequences of the systematic display of an ASR transcript.

¹⁶⁴ This may be the case of ASR transcripts interpreted as demonstrations of understanding (Heritage, 2007) by co-participants. As mentioned in section 7.10.3, following Heritage (2007), some demonstrations of understanding may constitute exceptions to the minimization maxim of Levinson (1987).

¹⁶⁵ Another, drastically different solution would be for the robot to display the action it is going to do in response to what was said, rather than a transcript of this previous utterance. This behavior is closer to a typically ethnomethodological definition of understanding as displayed by the next action of an interlocutor. It is, for example, what Amazon's robot Astro does.

7.12. Methodological takeaways

7.12.1. What is the human responding to? The tablet or the robot?

These observations open up a methodological question for the transcription norms in conversation analysis: to which entity do human participants respond to, and how should it be conveyed in the transcripts that conversation analysts produce from recordings of human-robot interactions? When a human moves his or her arm, this movement will typically be translated as an “action” from the human: the arm will generally not be oriented to by co-present participants as a full-fledged entity displaying intentions of its own. It will, most of the time, not be treated as the “author” of its movement. Similarly, when writing about human-robot experiments, any moving and speaking bundle of plastic, sensors, screens, LEDs, etc. is traditionally referred to as a single “robot” by the researcher. This “raw physical artifact” (H. Clark & Fischer, 2022) is categorized as a homogeneous entity in the transcription or description of the experiment. Yet, one could imagine transcribing “it” as a combination of several autonomous entities.

As Mondada (2002) discusses, typical transcription practices suppose attributing each utterance to a speaker, usually identified on the left of the transcript, at the beginning of each new turn. Yet, Mondada (2002) remarks that this convention imposes a discretization of streams of sound from the raw data¹⁶⁶ as well as an “autonomization and a specification of the uttered speech” (Mondada, 2002; my translation). Consequently, and relevantly to the analysis of human-robot interactions, it produces an effect of authorship (Mondada, 2002). In the case of robots wearing tablets, two alternative transcription choices each capture a different fraction of the “emerging categories” (Mondada, 2002; my translation) which become relevant for participants over the course of the interaction. On the one hand, the content displayed on the tablet can be transcribed (as was done in this work) as a conduct *stemming from* the robot – in the same way that, in human-human interactions, an arm gesture *is produced by* a participant without the arm being annotated as an autonomous entity separate from this participant’s self. For example, in the transcripts used in this study, the ASR feedback displayed on the tablet was referred to as one of the multimodal actions (voice, gesture, tablet content, flashing blue LEDs, etc.) produced by the preexisting participant “ROB” (the robot), identified on the left of the transcript¹⁶⁷. Following this choice, two lines from fragment 7.6 were transcribed in this manner:

```
8.          (0.1) % (0.4)
   rob      %displays "shut up"-->
9.  ROB     darn (0.5) sorry
```

¹⁶⁶ For example, when an utterance made during a choral response is transcribed as produced by a specific participant, even though it was indistinguishable (as a standalone part of this choral response) by co-present members themselves when it was uttered.

¹⁶⁷ Participants themselves sometimes actively worked to publicly characterize the ASR transcript as an intentional contribution from the robot (instead of an automatic behavior over which the robot would have no control), as in fragment 7.4.

Yet, on the other hand, the tablet could be transcribed as a full-fledged participant. That is, the content written on the tablet could be noted as an action produced by a new participant “TAB”, separate from “the rest of the humanoid robot” ROB. Then, fragment 7.6 should be transcribed as follows:

8. **TAB** **shut up**
9. **ROB** **darn (0.5) sorry**

One effect of this second way of transcribing is to convey that, when participants are responding to something appearing on the tablet, they are responding to a full-fledged participant “TAB”, whose actions are not authored by the humanoid “ROB” – which, as a consequence, is not accountable for any source of interactional trouble (e.g., a wrongly transcribed sentence, an insult, etc.) or any sequentially implicative turn (e.g., a greeting) produced on the tablet¹⁶⁸.

This dilemma is even more significant in other uses of robots featuring a screen where, as often with Pepper robots, is running an application that can be interacted with through touch (e.g., a menu, a virtual keyboard to type something, a game) or even a telepresent interlocutor’s face. In these extreme cases, transcribing as an “action from the robot” each new content displayed on the tablet (e.g., when the tablet suddenly displays a question, a picture, a different menu, a change in the UI, etc.) is likely to artificially produce a unified “robot+tablet” entity which did not exist in participants’ orientation to the local interaction which is being transcribed. Hence, we suggest that transcription choices are at risk of imposing etic categories on human-robot interactions even more than on human-human interactions. It is relatively safe to postulate that, when a human moves their arm, other participants will not orient towards this arm as a rational and autonomous actor accountable for its actions (e.g., an arm is rarely blamed for having pushed a glass off a table by itself). The situation is not so simple with robots. As potentially new ontological categories, it is much more intricate to determine what is locally treated as a “body part” and what is treated as a full-fledged actor among the many components of what we predefine as a unique robot. Human-robot data heavily limit the number of postulates that the analyst can reasonably produce before a detailed analysis. This type of data requires the analyst to attend, more than usual, to the locally emerging categories of participants even before a formal transcription is attempted.

7.12.2. The limits of eye-tracking to capture participants’ involved experience

Eye-tracking data can be subjected to similar scrutiny as to how well it captures participants’ orientation to their situated encounter with a robot – although the difficulties are elsewhere in

¹⁶⁸ Mondada’s (2016) multimodal transcription conventions, on which we base our transcript, allow for any of the two transcription choices we described. We made the choice to keep transcribing the ASR feedback appearing on the tablet as “an action from the robot”. Indeed, since the tablet was treated as such *most of the time* by participants, we decided to keep this transcription practice across all transcripts for the sake of readability. However, if the accuracy of transcriptions were to be the only criterion, the most rigorous choice could be to adapt the transcription to the way the robot is being treated: i.e., to transcribe the tablet as an autonomous actor only when it can be observably demonstrated that it was being treated as such in a specific interaction.

this case. Kristiansen & Rasmussen (2021) note that one of the pitfalls of using eye-tracking in EMCA research is “the special status which is assigned to gaze *a priori* by recording it separately by means of eye-tracking equipment” (Kristiansen & Rasmussen, 2021). Yet, a similar remark can be made regarding the division of the visual flux of participants between consistent categories (e.g., “the robot’s head”, “the robot’s arms”, etc.) across all interactions, which is required to aggregate and communicate quantitative eye tracking data.

Indeed, our statistical results suggest that the presence of a tablet diverted the attention from the embodied responses produced by the robot, since a significant fraction of our participants’ gaze fixations was focused on the top of the tablet in the “ASR transcription” condition. To understand to which degree participants focused their attention on the tablet, the preparation phase of our data implied to draw two “areas of interest” on a 2D image of the Pepper robot (“tablet” *versus* “head and body”) on which participants’ fixations were subsequently mapped (see section 7.6.2). These areas are congruent with the two main clusters of gaze fixation from participants (see Figure 7.10): even though they were predefined by the researcher, these zones of fixation match the two areas of the robot where participants focused their gaze. However, such divisions (even generated from clusters of attention) of the flux of visual information experienced by participants in situated interactions are likely to misrepresent the idiosyncrasy of these participants’ overall involved experience (H. L. Dreyfus & Kelly, 2007). That is, by isolating and by naming two components of the robot (“tablet” and “head and body”), we do not want to suggest that these zones exist as such “in the head” of participants (Garfinkel, 2019): i.e., that these participants mentally represented the robot as an object of theoretical inquiry, possessing a set of individualized and clearly designated properties (eyes, a tablet, arms, etc.) that they mentally singled out – visually, from the rest of the robot, but also in temporality, from the urgency of the ongoing sequence of action – when gazing at them. Instead, it is possible (and, we claim, likely) that the robot was experienced as a *gestalt*, i.e., as a coherent whole without identifiable independent parts¹⁶⁹.

¹⁶⁹ As Holthaus & Wachsmuth (2021) have demonstrated, human-robot interaction models can sometimes profit from treating the interaction as holistic.

8. CONCLUSION AND OPEN QUESTIONS

[...] one has to reverse the movement that is exalted by the myth of the cave, the professional ideology of the professional thinker, and return to the world of everyday existence, but armed with a scientific thought that is sufficiently aware of itself and its limits to be capable of thinking practice without destroying its object. Put less negatively, it is a question of understanding, first, the primary understanding of the world that is linked to experience of inclusion in this world, then the – almost invariably mistaken and distorted – understanding that scholastic thought has of this practical understanding, and finally the – essential – difference between practical knowledge – reasonable reason – and the scientific knowledge – scholastic, theoretical, reasoning reason – that is generated in autonomous fields. (Bourdieu, 2000, p. 50)

8.1. The practices glossed by “social agency”

This work has examined *the members’ methods in and through which a robot emerges as a social agent in different settings*. Rather than attempting to formalize an ahistorical and transsituational model of how a robot comes to be treated as an agent “in general”, we have described several locally relevant methods, practices, processes, etc. that participants displayed when organizing their activities with a robot – while maintaining this robot as a full-fledged and competent social participant. That is, as announced at the beginning of this work (see section 1.2), our goal has been to enrich the repertoire of already documented interactional processes occurring during human-robot encounters, indexed to specific settings, sequential contexts, spatial configurations, etc. Among the potentially infinite number of properties of a setting that can be perceived and attended to, we have strived to understand what is *used or accomplished* by human participants (and, to a certain extent, by the robot) to progress the activity at hand. This analytical lens has been maintained throughout the study of different phases of human-robot interactions, taking place in different settings (natural and experimental), and involving robots programmed in various ways.

In other words, relying on the growing body of EMCA literature on human-agent interactions, we have attempted to clarify (both theoretically and empirically) what the analysts themselves generally gloss as “social agency”. Namely, in the definition employed throughout this work, the robot’s sociality is not contemplated as a property of the robot, nor as a feature that floats above the interaction: rather, the robot’s social agency *is* the describable local organization in which specific rights or obligations are observably attributed to the robot – and in which all participants are *embedded* no matter if they are “robots” or “humans”. Hence, in a “deflationary move” (Lynch, 2022 – quoted in Mlynář & Arminen, 2023) typical of EMCA, social agency has been argued to be identical with the observable practices (produced by members of a setting) *by which* an entity is treated as an accountable participant whose conduct is contingent on the ongoing interaction. In this understanding, the notion of social agency indexes an organization (or a system, a structure, a publicly observable distribution of rights

and responsibilities, etc.) “that stands apart from any particular actor” (Button & Sharrock, 2016). Social agency is about the “doing” rather than the “doer” (Button & Sharrock, 2016).

8.1.1. Empirical results

In short, we have examined the emergence of a robot as a greetable social agent (chapter 5), its continuous accomplishment as a social agent in a public setting (chapter 6), and, finally, the crucial relevance of the “automatic speech recognition transcript” as a configuring parameter for a robot’s social agency (chapter 7). To elaborate further, the empirical findings presented throughout this work indicate that:

1. Evidently, a robot pre-labeled as “social” does not necessarily emerge as a social agent *in situ*. However, even the very first instants of an encounter with a robot can be consequential on its momentary emergence as an accountable conversational partner, or, instead, on its persistent treatment as an object merely producing pre-recorded turns. In other words, the first moments of physical co-presence between a robot and a human are not an interactional vacuum: the intertwining between participants’ actions and the very first behaviors (or, often, the motionlessness) displayed by the robot produces a priori unpredictable sequential trajectories, which are susceptible to configuring the timing and the manner in which the robot emerges as a social agent.
2. Yet, even when a robot may be glossed as being treated like a “social agent” in public interactions with groups of humans, a closer look reveals the recurring work stemming from bystanders to re-configure the robot’s conduct as relevant to the task at hand. Not only can we observe a *scaffolding* of the robot’s ability to produce relevant contributions – i.e., as a significant body of literature has demonstrated, the robot and its environment are configured *before the examined interactions occur* to facilitate the robot’s success at accomplishing its tasks or actions (Chevallier, 2023; Kamino & Sabanovic, 2023) –, but the robot’s conduct can also be framed *a posteriori* by the audience to be meaningful for the person directly interacting with this robot. This “pre-chewing” of a robot’s conduct by a third party (after this robot has spoken, gestured, moved, etc.) – *to re-configure it as a relevant contribution for the main speaker* – is one facet of the work to make technology “work” (Chevallier, 2023; Due & Lüchow, 2022; Greiffenhagen et al., 2023; Lipp, 2022; Pelikan et al., 2020, 2024; Stommel et al., 2022; Tuncer et al., 2023).
3. Finally, an especially configuring parameter of the treatment of the Pepper robot as a (competent) social agent is the “automatic speech recognition transcript”: that is, the public display by this robot, on a screen placed on its torso, of the words it “hears” from its interlocutor’s speaking turn. Indeed, among other pragmatic consequences, having access to what the robot (mis-)heard regularly led participants to *the a posteriori re-evaluation* of the situational relevance of the robot’s actions – even when these actions were so far treated as adequate responses. Through this re-evaluation, participants overrode their interpretation-so-far of the robot’s conduct as activity-relevant: they displayed a change in their footing regarding the status of the robot’s conduct *as a response to what they really requested, asked, remarked, etc.* In sum, we argue that, in an informational configuration where participants have direct access to a robot’s

reconstruction of the external world, the “weave of interactional moves” (Linell & Lindström, 2016) produced between humans and robots is less likely to offer the (superficial) coherence which is the bedrock of an experience of the ongoing interaction as intersubjectively shared.

8.1.2. Theoretical and practical frictions

This research has examined human-robot interactions with analytical tools and concepts originally designed for and from the analysis of human-human interaction (Mondada & Meguerditchian, 2022; Mondémé, 2022) – and has done so in a context still uncommon for EMCA practitioners: the inside of a technology company. As much as possible, we have attempted to describe the theoretical and concrete tensions that arose between, on the one hand, these analytical practices and concepts typical of EMCA and, on the other hand, the empirical study of human-robot interactions – within an industrial context.

- a. First, section 2.2 investigated the “conceptual loosening” implied by the analytical treatment of the robot as a full-fledged participant. In particular, it discussed the extent to which a central tool for the EMCA practitioner (the “next-turn proof procedure”) remained relevant during the examination of human-robot interactions.
- b. Section 2.3 then outlined the challenge of deriving interaction designs for rule-based robots from the examination of human experts' practical activity (such as that of ordinary expert conversationalists) understood as non-rule governed (Bourdieu, 1977; H. L. Dreyfus, 2014; S. E. Dreyfus, 2014). After discussing some of the fundamental postulates of doing “applied CA” (Antaki, 2011a; Ten Have, 2007) to improve rule-based robots, we claimed that a humble yet realistic goal for an EMCA-inspired design approach is to narrow the gap (Coulon, 1993) between the robot's “plan” and the emergent local practices of its human interactants (Suchman, 1987): that is, to help the designer to consider more of the common-sense resources and procedures that conversationalists ordinarily use and orient to. With rule-based robots, the point of an applied EMCA approach is not to collapse the distinction between the map and the territory, but merely to draw better maps.
- c. Finally, chapter 4 described some of the disjunctions that arose between our EMCA approach and the ordinary practices, processes, and tools of the employees of a technology company. After focusing at length on the “agency” (Battentier & Kuipers, 2020; Verbeek, 2006) of some design and programming tools available for the Pepper robot, we argued that these tools reify specific representations of what a human-robot interaction should be (and hinder the design or the programming of different types of actions or conducts from the robot). As such, they convey an “unnatural use of the natural language” (Moore & Arar, 2018).

8.1.3. A (fleeting?) specification of the emergence of robots as social agents?

8.1.3.1 *Beyond human-robot interactions as “black boxes”*

With these focal points and this analytical lens, we attempted to remedy a common orientation to human-robot interactions as “black boxes” (Heritage, 2001). Instead, because they are part of human activities that display “order at all points” (Sacks, 1995) – or, at least, which are not “fundamentally disorderly” (Heritage, 2001) – it was possible to highlight micro-level phenomena that are accomplished or relevant for humans (and, sometimes, robots), as they coordinate their respective conduct. Doing so led us to respecify some lay concepts used to gloss human-robot interactions. Descriptions of strips of interactions as “encountering”, “getting acquainted” or “having a conversation” with a robot may aggregate widely different local organizations under a common label and/or obscure the work to make technology “work” (Greiffenhagen et al., 2023) produced by co-present humans.

Even more, we started this research in the hope of identifying core features of situations commonly glossed as “having a social interaction” – features that make these situations recognizable and typifiable as a certain type of activity. Ironically, the details of the “work” and constant monitoring that human interactants accomplish to make the technology “work” (Greiffenhagen et al., 2023) may be a significant feature of these activities that make them recognizable as “social interactions with a robot”. In other words, the interactional “burden” (Brown et al., 2023) weighing on human interactants is managed by these humans through the use of specific methods (likely to be “seen but unnoticed” – Garfinkel, 1967) that constitute distinctive features of human-robot “social” interactions. In this sense, we argued that labels such as “social robots” or “conversational” robots are, currently, wishful mnemonic (McDermott, 1976) for social interaction.

However, descriptions of micro-order phenomena glossed by lay terms (“social robots”, “encountering a robot”, “having a conversation with a robot”, etc.) do not have to be the end goal of an EMCA study – unless one explicitly follows a quietist approach to the study of social facts. Pointing to the relevant features of human-robot encounters (as made visible by human and robot interactants) is to highlight the properties of local situations through which a robot emerges as social. Understood as such, the micro-order phenomena identified by EMCA (which can be mobilized to do “applied CA” (Ten Have, 2007) and produce EMCA-inspired designs) allow the designer or the engineer to identify some of the variables, parameters, features, etc. on which they should act – among the potentially infinite number of properties of a situation that can be discretized from an etic perspective.

8.1.3.2 *The embeddedness of stable “human-robot” practices in a setting*

As mentioned at the outset, the format of this research – based on several corpora – stems from the idea that different settings (etically defined as “natural or experimental” and “professional or recreational”) may offer a more complete view of the methods, practices, categorizations, etc. that human participants display in their interactions with robots. Now that this empirical work has been carried out, our results indeed reflect the relevance of these settings for human participants: they were connected with different observable expectations towards the robot, with different levels of interactional “good will” (Pelikan et al., 2022) to treat the robot’s contributions as relevant (e.g., when interacting with a robot in a museum versus

in the hall of an office building), and with more or less exploratory practices from humans (e.g., to grasp the extent of the robot's abilities – Tuncer et al., 2023).

Significantly, because of their embeddedness in a physical and social setting, some of the practices described in this research are likely to be ephemeral (Mlynář & Arminen, 2023) – traces of a distinct period in the history of human interactions with artificial agents. In another vocabulary and theoretical perspective, these practices may one day be said to reflect and constitute a momentary degree of “acquaintance” (or “comfort”, or “familiarity”) of “users” with a given state of robotic technology and AI. As Mlynář & Arminen (2023) note, “EM/CA research also inevitably and unavoidably—though mostly inadvertently—provides accounts of practices that are reflexively entrenched in the exogenous time of social processes”. Like the telephone etiquette of the early days of landline telephony (Mlynář & Arminen, 2023), even relatively stable and recurring practices currently observed in human-robot interactions are meaningful only in relation to a lifeworld that may, sooner or later, fade away (Mlynář & Arminen, 2023).

Still, we argue, the variability of what constitutes “normal” human-robot practices (in both the moral and statistical senses of “normal”) should also be pictured as a blurry background against which *long-lasting* micro-order phenomena can be contrasted: methods, processes, practices that are fundamental machinery of the local organization of human-robot or human-human activities. As suggested in the introduction section of this work, the current shifting fresco of human practices involving artificial agents is an opportunity to reveal more fundamental or stable features of what constitutes “sociality” (Mondémé, 2022).

8.2. Do human-robot interactions shed light on the inner-workings of “human-human” interactions?

8.2.1. Human-robot encounters as breaching experiments

In Förster et al. (2023), Albert argues that “moments of trouble and failure can provide researchers with ideal empirical material for observing the structure of the participation frameworks we use to get things done in everyday life”. Given the many human-robot encounters that can be analyzed as displaying occurrences of “trouble and failure” (including a consequent part of those presented in this work), it is therefore tempting to instrumentalize these encounters as breaching experiments (Garfinkel, 1967). Indeed, although they do not necessarily exhibit complete breakdowns of the interaction order, the potential strangeness of these settings could allow such pre-existing “background expectancies to come into view” (Garfinkel, 1967). In this perspective, human-robot first-time encounters would constitute unintentional breaching experiments (Garfinkel, 1967) that produce “circumstances [...] where sense-making activities are more prominent” (Have, 2005): they would function as “natural experimental situations” (Mondada, 2002; translation ours) where participants can be expected, more often than not, to “thematize generally invisible ordinary phenomena” (Mondada, 2002; translation ours). The “estranging” (Have, 2005) task would therefore be mainly entrusted to the robot's interactional (in)competence and to its non-human appearance.

While psychologists have long used robots as tools to shed light on human cognition (Broadbent, 2017), using robots to document humans in regard to sociological and

praxeological questions is, to the best of our knowledge, a relatively recent idea. One of the earliest intentional use of robots as estrangement tools in an EMCA perspective can be found in Weiss et al. (2010), who explored how the presence of a robot navigating city streets breached the everyday order, as observable in the reactions of passers-by. Even more recently, Pitsch (2020) develops the idea of “HRI as a tool to investigate situated action”. Although she does not take position on the potential estrangement provoked by the presence of a robot, she suggests that

“[...] HRI can serve as a methodological and conceptual tool for investigating situated (inter-)action on at least five levels: “(1) exploring basic building blocks of situated (inter-)action involving the orchestration of communicational resources as complex ‘multimodal gestalts’ under the condition of interactional contingency and the co-participants’ interpretation; (2) theories and concepts of interaction; (3) socio-technical constellations; (4) exploring human sociality and moral orders as they transpire in novel technologically rich situations; (5) exploring the integration of novel technologies in the ecology of our daily lives and resulting societal questions.” (Pitsch, 2020)

Congruently with this line of inquiry, we argue that a fruitful exercise for EMCA analysts is to study what is “breached” in human-robot encounters: can we observe breakdowns or disruptions in preexisting “background expectations” (Garfinkel, 1967) or “natural facts of life” (Garfinkel, 1967)? That is, are these settings observably oriented to by participants as inconsistent with a “jurisprudence” (Meyer, 2019, based on Schneider, 2013), where “past judicial decisions are taken as precedents for new decisions” (Meyer, 2019)? Does the robot’s conduct produce a “breach”, in the sense that “practices that have proven successful in my interaction [...] in situations in the past, that I judge similar to the one present” (Meyer, 2019) do not work anymore when interacting with this robot? Or, instead, is the robot encountered as a new ontological category with which none of these preexisting expectations will be visibly disrupted? In the latter case, participants may be seen working to categorize or test the robot, but these activities will not allow the analyst to shed light on mechanisms at play in interactions taking place exclusively between humans. In fact, in the most extreme scenario, the robot would merely be “life as usual” (Garfinkel, 1967), a technological device like any other, whose proper use has to be figured out: no more “anthropologically strange” (Garfinkel, 1967) than a new phone whose user interface has to be mastered¹⁷⁰. Yet, in the opposite scenario, the robot’s intermediary status between an agent and a non-agent may constitute a powerful “estrangement device” (Lynch & Eisenmann, 2022) likely to shed light onto core features of everyday human sociality – up to the “interaction engine” (Levinson, 2020b) underpinning human interactions and, possibly, non-human interactions (Heesen & Fröhlich, 2022; Mondémé, 2022).

8.2.2. Should we use “placeholder” analytical categories?

As noted at the very beginning of this work, the 1960s sociologist or linguist studying a strip of talk would have found “few systematic resources” (Heritage, 2001) with which to analyze a strip of talk “and none which could offer any significant clues as to the details of the actions the participants are engaged in” (Heritage, 2001). By comparison, how well-equipped is the

¹⁷⁰ Which, arguably, is already a situation of trouble likely to reveal some of users’ expectancies.

current-day researcher attempting to make sense of human-robot encounters? Undeniably, we now have incomparably more “systematic resources” at our disposal to account for the local organization of human-human settings. A researcher interested in this topic can currently profit from the multiple incremental discoveries achieved regarding sequential organization (turn-taking, conditional relevance, etc.) and from numerous empirical corpora organized around heavily documented practices (repairs, requests, etc.) or typically ritualized sequences (Stivers & Rossano, 2010b): interaction openings (Harjunpää et al., 2018; Mondada, 2015; Mortensen & Hazel, 2014; Robinson, 1998), pre-closings (Schegloff & Sacks, 1973b), etc. Likewise, Robinson (2016) identifies at least 11 domains of practices “that participants use to form/ascribe social actions(s)”: the domain of turn-taking, the domain of membership categorization, the domain of sequence organization, the domain of repair, the domain of social status, including epistemic status, etc. For the researcher, each of these domains of practice (with their collections of exemplars, their technical terms, their specific discretization of the continuous stream of conducts) functions as “a set of lenses usable when analyzing data” (Robinson, 2016). Instead of approaching mundane human interactions as an inexplicably meaningful gestalt, this pre-existing body of works provides tools to efficiently describe “what is going on” without having to rely on rudimentary terms that heavily gloss (and, consequently, obscure) the local seen-but-unnoticed processes, methods, activities, etc. which form “social facts” (Garfinkel, 1967).

Yet, the existing literature about human-human interactions is *more* than helpful to conversation analysts who focus on human-robot interactions. To the best of our knowledge, this body of literature is currently the conceptual and empirical basis of each and every of these “HRI” researchers’ analyses (this work included). In other words, interactional phenomena reported by conversation analysts studying robots were, originally, described and labeled based on detailed studies of “human-human” interactions. Two possibilities arise in this state of affairs. Either some of the phenomena observable between robots and humans are *identical or homologous* (Mondémé, 2022; Pika et al., 2018) to those observed between humans, in the sense that they are naturally organized ordinary activities¹⁷¹ whose relevant features (for the participants themselves as they are immersed in their practical activities) are the same; or, conversation analysis has not yet built *sui generis* categories for human-robot interactions. In this last hypothesis, EMCA-oriented studies on human-robot interactions would constitute *metaphorical or analogical* (Mondémé, 2022) *glosses* that systematically index human-human processes to account for activities specific to human-robot or even human-machine settings. Rather than producing an autonomous description, as faithful as possible to the here-and-now or the “just thisness” (Garfinkel & Wieder, 1992) of human-robot settings, analysts would merely report some degree of similarity between an activity involving a robot and another documented set of human-human practices: these settings would be “similar enough” as defined emically by the analysts. Under this assumption (to which the analyses produced in this work are vulnerable), no matter the meticulousness of their descriptions, researchers would forcibly categorize what is observed in human-human boxes and risk “losing the phenomenon” (Garfinkel, 2002).

Some researchers who worked on interactions between humans and non-human entities have already taken note of this risk. Tisserand & Baldauf-Quilliatre (2023) suggest that

¹⁷¹ That is, those “ethnomethodological case materials” (Lynch & Garfinkel, 2022) which would be put between ticked brackets in Garfinkel’s (2002) notation: the local organization displayed by participants themselves, before any formal analysis is produced by an observer to account for these practices.

conducts regularly observed during the very first instants of co-presence between a robot and a human – which superficially appear to fit the typical practices involved in an “opening sequence” – may, in fact, be closer to a form of “testing” (e.g., as described by Tuncer et al., 2023) of the robot’s abilities. In accordance, studying interactions between baboons, Mondada & Meguerditchian (2022) occasionally speak of “hindquarters presentation” rather than “greetings” to avoid imposing on the interaction a category “formulated as such for and by humans” (Mondada & Meguerditchian, 2022 – our translation).

8.2.3. Analogically similar or *sui generis* practices?

The heart of the matter concerns what “comes into view” (Garfinkel, 1967) when humans and robots interact. Mainly, do human participants display background expectations, relevance rules, methods, categories, etc. typical of “life as usual” (Garfinkel, 1967) in exclusively “human-human” interactions? Or, instead, do they display entirely specific, *sui generis*, practices? This methodological issue arises each time categories or conceptual tools built to account for human-human interactions are applied to interactions involving non-humans. As mentioned previously, it was notably put forward by Pika et al. (2018) then Mondémé (2022) – in terms of “analogy” versus “homology” – regarding the extension of turn-taking mechanisms (as defined by Sacks et al., 1974) to non-human primates.

In the case of human-robot encounters, three possibilities can be considered for any observable conduct:

1. This conduct displays mechanisms (e.g., relevance rules – Robinson, 2016) or actions (requesting, complaining, etc.) that are *typical of humans’ activities as ordinary members of society*: it can be rigorously described using categories built from the observation of *human interactants*.
2. This conduct displays mechanisms or actions *typical of regular interactions with machines* (e.g., “in-the-wild Turing tests” – Ivarsson & Lindwall, 2023). It does not fit in analytical categories (Mondémé, 2022) like “greeting sequence,” “pre-closing,” “repair,” nor does it display mechanisms strictly homologous to those observable in *human-human conversation*, e.g., “turn-taking”.
3. This conduct is *produced as part of strictly idiosyncratic practices, designed for the co-present robot as a radically different kind of entity* – compared to other technological devices or human interactants. Whether they are built on the fly or incrementally after repeated interactions with this robot, these practices are *sui generis* in the sense that they do not fit typologies and mechanisms constructed from and for human-human or human-machine interactions. The locally organized activities observed by the researcher cannot be rigorously paired with previously documented activities; *and were not paired as such by local participants*. However, the degree to which current analyses of human-robot interactions provide examples of such practices (rather than

phenomena typically found in other human-machine interactions) is an empirical question beyond the scope of this list of theoretical possibilities.¹⁷²

To illustrate possibilities 1. and 2., let's consider a situation where a robot waves its hand in front of a group of humans. A short moment later, the humans achieve a similar-looking hand-waving gesture towards the robot. This situation may be analogically "paired" (Husserl, 2013, cited by Meyer, 2019) by the human participants with a typified "greeting sequence": in the sense where this situation and previously achieved human-human greetings "found phenomenologically a unity of similarity" (Husserl, 2013, cited by Meyer, 2019) for the situated actors themselves. Or, instead, the local organization and phenomenological experience of these conducts may be different from that which occurs during a "greeting" action: through their hand gesture, human participants may "do" something else than "greet the robot. For example, they may produce a form of "testing"¹⁷³ (Tisserand & Baldauf-Quilliatre, 2023; Tuncer et al., 2023) of the robot's "ability to recognise an action and produce an appropriate response" (Tuncer et al., 2023). Likewise, a detailed analysis may find that a human seizing a humanoid robot's hand is not "initiating a handshake (enacting the robot as a potential, intercorporeal, shaker)" (Tuncer et al., 2023) but is, instead, "testing for haptic sensitivity (enacting the robot as a machine with or without tactile capacities)" (Tuncer et al., 2023). Even more, what would constitute, in some human-human settings, an accountable interruption of an interlocutor's ongoing turn-at-talk might merely be an attempt to speed up the robot's pre-scripted speech – see for example section 3.3.3.2 of this manuscript for a transcript of a participant expertly optimizing the speed of his scripted check-in with a robot.

However, specifying the status of a particular strip of interaction (as typical or radically "new") does not eliminate methodological dilemmas. Even in the hypothesis of a total absence of overlap – in terms of praxeological organization and phenomenological experience¹⁷⁴ – between a human-robot encounter and documented practices among humans, one might contend that the preexisting analytical categories built and refined to account for interactions occurring between humans should still be used as *heuristic metaphors* (Mondémé, 2022) to describe, as a first step, what is happening with robots. Even more, as Mondémé (2022, 2023)

¹⁷² These three possibilities are heuristic and do not intend to draw a sharp distinction between dynamic statuses of a robot (Pelikan et al., 2022) – "that can be enacted, in one breath, as an agent and a thing" (Alač, 2016) – nor to deny the work undertaken by humans to (re-)categorize the robot (Ivarsson & Lindwall, 2023). This panorama aims at clarifying the type of discrepancies that can arise between locally organized human-robot settings and the categories available to report them.

¹⁷³ "In testing, participants appear less curious about the robot, they test its ability to recognise an action and produce an appropriate response; and if it does not respond appropriately and immediately, they turn away from the robot." (Tuncer et al., 2023)

¹⁷⁴ Note that non-formalized phenomena studied by EMCA researchers are not the same as the "things themselves" (Husserl, 2012) examined by phenomenologists. As Rawls (2002) remarks, "[...] when phenomenologists bracket, they bracket the typification, or concept, and attempt to discover what a phenomenon or perception consists of prior to the patterning action of the brain, or the patterning action of cultural typifications, depending on the kind of phenomenologist. [...] For Garfinkel, by contrast, it is not one's orientation toward the world, or the patterning action of typifications, that imposes an organization on an otherwise unorganized set of perceptions. At least insofar as the world is perceived in socially meaningful ways, it is because persons have, in interacting, organized that world for themselves such that the things in that world can be initially perceived as orderly, or are in fact orderly, because they have been constructed to be orderly. Events have been constructed so as to have witnessable recognizability." (A. W. Rawls, 2002). See also Lynch & Eisenmann (2022) on how "[p]henomenal fields, for Garfinkel, are formed through embodied work performed in concert with the actions of others".

advocates for the study of interspecies interactions, grouping sequences of actions that have the same “functional relevance” (Mondémé, 2022) may, in fact, help to uncover more encompassing mechanisms of sociality, beyond strictly human-human communication¹⁷⁵. In other words, the following methodological questions remain:

- 1) Should preexisting human-human categories be used *at all* to account for human-robot interactions (no matter if similarities are empirically demonstrable by the analyst)?
- 2) If so, and in the absence of more specific tools to account for practices observable in human-robot settings, should human-human categories be mobilized merely as “placeholders” until the analyst has managed to sufficiently specify “what is going on”? Or, instead, should these human-human categories be valued as heuristic devices facilitating the discovery of even more encompassing mechanisms of human interactions (Mondémé, 2022)?

¹⁷⁵ In the end, this questioning is a particular occurrence of debates regarding the “different routes to generalize findings” (Tavory, 2022) between conversation analysis, institutional conversation analysis, and ethnomethodology. It can be also connected with different views as to what constitutes “constructive analysis” and what constitutes ethnomethodology (Button et al., 2022), or, on a wider scale, can even be read as an occurrence of incommensurable approaches to science (Garfinkel & Wieder, 1992) between formal sciences and ethnomethodology (Garfinkel, 1996). In short, we argue that the challenge of choosing appropriate analytical categories for studying human-robot interactions boils down to balancing these two classic arguments:

- 1) Endlessly attempting to (over)specify the “just thisness” of every locally organized interaction will, ultimately, make any comparison or generalization impossible.
- 2) Yet, refusing to attend to the precise manner in which social settings are locally managed will lead to aggregate situations whose main features (from the point of view of the participants themselves) do not overlap.

9. REFERENCES

- Akrich, M. (1989). La construction d'un système socio-technique. Esquisse pour une anthropologie des techniques. *Anthropologie et Sociétés*, 13. <https://doi.org/10.7202/015076ar>
- Alač, M. (2016). Social robots: Things or agents? *AI & SOCIETY*, 31(4), 519–535. <https://doi.org/10.1007/s00146-015-0631-6>
- Alač, M., Gluzman, Y., Aflatoun, T., Bari, A., Jing, B., & Mozqueda, G. (2020). Talking to a Toaster: How Everyday Interactions with Digital Voice Assistants Resist a Return to the Individual. *Evental Aesthetics*, 9. https://eventalaesthetics.net/wp-content/uploads/2021/03/EAV9N1_2020_Alac_Toaster_3_53.pdf
- Albert, S., & de Ruiter, J. P. (2018). Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, 10(2), 279–313. <https://doi.org/10.1111/tops.12339>
- Albert, S., & Ruiter, J. (2018). Repair: The Interface Between Interaction and Cognition. *Topics in Cognitive Science*, 10, 279–313. <https://doi.org/10.1111/tops.12339>
- Aldrup, M. (2019). 'Well let me put it uhm the other way around maybe': Managing students' trouble displays in the CLIL classroom. *Classroom Discourse*, 10(1), 46–70. <https://doi.org/10.1080/19463014.2019.1567360>
- Antaki, C. (2008). Formulations in psychotherapy. In *Conversation Analysis and Psychotherapy (2008)* (Vol. 7, pp. 627–647). <https://doi.org/10.1017/CBO9780511490002.003>
- Antaki, C. (2011a). *Applied conversation analysis: Intervention and change in institutional talk*. Springer.
- Antaki, C. (2011b). Six Kinds of Applied Conversation Analysis. In C. Antaki (Ed.), *Applied Conversation Analysis: Intervention and Change in Institutional Talk* (pp. 1–14). Palgrave Macmillan UK. https://doi.org/10.1057/9780230316874_1
- Antaki, C. (2012). Conversation Analysis and the Study of Atypical Populations. In *The Handbook of Conversation Analysis* (pp. 534–550). <https://doi.org/10.1002/9781118325001.ch26>
- Antaki, C., Condor, S., & Levine, M. (1996). Social identities in talk: Speakers' own orientations. *British Journal of Social Psychology*, 35(4), 473–492. <https://doi.org/10.1111/j.2044-8309.1996.tb01109.x>
- Antaki, C., & Kent, A. (2012). Telling people what to do (and, sometimes, why): Contingency, entitlement and explanation in staff requests to adults with intellectual impairments. *Journal of Pragmatics*, 44(6), 876–889. <https://doi.org/10.1016/j.pragma.2012.03.014>
- Arias, K., Jeong, S., Park, H. W., & Breazeal, C. (2020). Toward Designing User-centered Idle Behaviors for Social Robots in the Home. *1st International Workshop on Designerly*

HRI Knowledge. Held in Conjunction with the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020).

- Auer, P., Bauer, A., & Hörmeyer, I. (2020). How Can the 'Autonomous Speaker' Survive in Atypical Interaction? The Case of Anarthria and Aphasia. In R. Wilkinson, J. P. Rae, & G. Rasmussen (Eds.), *Atypical Interaction* (pp. 373–408). Springer International Publishing. https://doi.org/10.1007/978-3-030-28799-3_13
- Baddoura, R., & Venture, G. (2015). This Robot is Sociable: Close-up on the Gestures and Measured Motion of a Human Responding to a Proactive Robot. *International Journal of Social Robotics*, 7(4), 489–496. <https://doi.org/10.1007/s12369-015-0279-x>
- Battentier, A., & Kuipers, G. (2020). Technical Intermediaries and the Agency of Objects: How Sound Engineers Make Meaning in Live Music Production. *Biens Symboliques / Symbolic Goods*, 6. <https://doi.org/10.4000/bssg.438>
- Beach, W. A. (1994). Relevance and consequentially. *Western Journal of Communication (Includes Communication Reports)*, 58(1), 51–57.
- Beach, W. A., & Sigman, S. J. (1995). Conversation Analysis: "Okay" as a Clue for Understanding Consequentiality. In *The Consequentiality of Communication* (pp. 121–162). Routledge.
- Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, 3(2), 74. <https://doi.org/10.5898/JHRI.3.2.Beer>
- Ben-Youssef, A., Clavel, C., Essid, S., Bilac, M., Chamoux, M., & Lim, A. (2017). UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 464–472. <https://doi.org/10.1145/3136755.3136814>
- Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in child–robot interaction. *International Journal of Human-Computer Studies*, 69(7–8), 539–550. <https://doi.org/10.1016/j.ijhcs.2011.04.003>
- Bergmann, J. R. (1998). Introduction: Morality in Discourse. *Research on Language and Social Interaction*, 31. <http://www.tandfonline.com/doi/abs/10.1080/08351813.1998.9683594>
- Bergmann, J. R., & Drew, P. (2018). Introduction. Jefferson's "wild side" of conversation analysis. *Repairing the Broken Surface of Talk. Managing Problems in Speaking, Hearing, and Understanding in Conversation*, 1–26.
- Bertrand, R., & Goujon, A. (2017). (Dis)aligning for improving mutual understanding in talk-in-interaction. *Revue française de linguistique appliquée*, XXII(2), 53–70. Cairn.info. <https://doi.org/10.3917/rfla.222.0053>
- Billig, M. (2013). *Learn to Write Badly: How to Succeed in the Social Sciences* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139208833>

- Bilmes, J. (2014). Preference and the conversation analytic endeavor. *Journal of Pragmatics*, 64, 52–71. <https://doi.org/10.1016/j.pragma.2014.01.007>
- Bolden, G. B. (2012). Across languages and cultures: Brokering problems of understanding in conversational repair. *Language in Society*, 41(1), 97–121. Cambridge Core. <https://doi.org/10.1017/S0047404511000923>
- Bourdieu, P. (1977). *Outline of a Theory of Practice* (R. Nice, Trans.; 1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511812507>
- Bourdieu, P. (2000). *Pascalian meditations*. Stanford University Press.
- Brengman, M., De Gauquier, L., Willems, K., & Vanderborght, B. (2021). From stopping to shopping: An observational study comparing a humanoid service robot with a tablet service kiosk to attract and convert shoppers. *Journal of Business Research*, 134, 263–274. <https://doi.org/10.1016/j.jbusres.2021.05.025>
- Broadbent, E. (2017). Interactions With Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology*, 68(1), 627–652. <https://doi.org/10.1146/annurev-psych-010416-043958>
- Brown, B., Broth, M., & Vinkhuyzen, E. (2023). The Halting problem: Video analysis of self-driving cars in traffic. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3544548.3581045>
- Button, G. (1990). Going Up a Blind Alley: Conflating Conversation Analysis and Computational Modelling. In P. LUFF, N. GILBERT, & D. FROHLICH (Eds.), *Computers and Conversation* (pp. 67–90). Academic Press. <https://doi.org/10.1016/B978-0-08-050264-9.50009-9>
- Button, G., Lynch, M., & Sharrock, W. (2022). *Ethnomethodology, conversation analysis and constructive analysis: On formal structures of practical action*. Taylor & Francis.
- Button, G., & Sharrock, W. (1995). On simulacrum of conversation: Toward a clarification of the relevance of conversation analysis for human-computer interaction. In *The Social and Interactional Dimensions of Human-Computer Interfaces*. Cambridge University Press.
- Button, G., & Sharrock, W. (2016). In support of conversation analysis' radical agenda. *Discourse Studies*, 18(5), 610–620. <https://doi.org/10.1177/1461445616657955>
- Cao, X., Yamashita, N., & Ishida, T. (2016a). How Non-native Speakers Perceive Listening Comprehension Problems: Implications for Adaptive Support Technologies. In T. Yoshino, G.-D. Chen, G. Zurita, T. Yuizono, T. Inoue, & N. Baloian (Eds.), *Collaboration Technologies and Social Computing* (pp. 89–104). Springer Singapore.
- Cao, X., Yamashita, N., & Ishida, T. (2016b). Investigating the Impact of Automated Transcripts on Non-Native Speakers' Listening Comprehension. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 121–128. <https://doi.org/10.1145/2993148.2993161>

- Cappuccio, M. L. (2023). Dreyfus is right: Knowledge-that limits your skill. *Synthese*, 202(3), 85. <https://doi.org/10.1007/s11229-023-04248-6>
- Chen, M.-L., Yamashita, N., & Wang, H.-C. (2018). Beyond Lingua Franca: System-Facilitated Language Switching Diversifies Participation in Multiparty Multilingual Communication. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW). <https://doi.org/10.1145/3274303>
- Chevallier, M. (2023). Staging Paro: The care of making robot(s) care. *Social Studies of Science*, 53(5), 635–659. <https://doi.org/10.1177/03063127221126148>
- Clark, H., & Fischer, K. (2022). Social robots as depictions of social agents—Behavioral and Brain Sciences (forthcoming). *Behavioral and Brain Sciences*, 2022, 1–33.
- Clark, H. H., & Wasow, T. (1998). Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37(3), 201–242. <https://doi.org/10.1006/cogp.1998.0693>
- Clayman, S. E. (2015). Ethnomethodology, General. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 203–206). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.44020-1>
- Coeckelbergh, M. (2019). Skillful coping with and through technologies: Some challenges and avenues for a Dreyfus-inspired philosophy of technology. *AI & SOCIETY*, 34(2), 269–287. <https://doi.org/10.1007/s00146-018-0810-3>
- Coulon, A. (1993). *Ethnométhodologie et éducation*. Presses Universitaires de France; Cairn.info. <https://www.cairn.info/ethnomethodologie-et-education--9782130452362.htm>
- Dai, D. W., & Davey, M. (2023). On the Promise of Using Membership Categorization Analysis to Investigate Interactional Competence. *Applied Linguistics*, amad049. <https://doi.org/10.1093/applin/amad049>
- Dautenhahn, K. (2004). Robots we like to live with?!—A developmental perspective on a personalized, life-long robot companion. *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 17–22. <https://doi.org/10.1109/ROMAN.2004.1374720>
- De Stefani, E., & Mondada, L. (2018). Encounters in Public Space: How Acquainted Versus Unacquainted Persons Establish Social and Spatial Arrangements. *Research on Language and Social Interaction*, 51(3), 248–270. <https://doi.org/10.1080/08351813.2018.1485230>
- Dennis, A. (2003). Skepticist Philosophy as Ethnomethodology. *Philosophy of the Social Sciences*, 33(2), 151–173. <https://doi.org/10.1177/0048393103033002001>
- Deppermann, A. (2012). How does ‘cognition’ matter to the analysis of talk-in-interaction? *Language Sciences*, 34(6), 746–767. <https://doi.org/10.1016/j.langsci.2012.04.013>
- Deppermann, A. (2018). Inferential Practices in Social Interaction: A Conversation-Analytic Account. *Open Linguistics*, 4(1), 35–55. <https://doi.org/10.1515/opli-2018-0003>

- Deppermann, A., & Haugh, M. (2022a). Action Ascription in Social Interaction. In A. Deppermann & M. Haugh (Eds.), *Action Ascription in Interaction* (1st ed., pp. 3–28). Cambridge University Press. <https://doi.org/10.1017/9781108673419.001>
- Deppermann, A., & Haugh, M. (Eds.). (2022b). Introduction. In *Action Ascription in Interaction* (pp. 1–28). Cambridge University Press; Cambridge Core. <https://www.cambridge.org/core/product/D35F42F461781515EF3A6C57AFBC3DD8>
- Deppermann, A., Mondada, L., & Doehler, S. P. (2021). Early Responses: An Introduction. *Discourse Processes*, 58(4), 293–307. <https://doi.org/10.1080/0163853X.2021.1877516>
- Deppermann, A., & Schmidt, A. (2021). Micro-Sequential Coordination in Early Responses. *Discourse Processes*, 58(4), 372–396. <https://doi.org/10.1080/0163853X.2020.1842630>
- Deslyper, R. (2013). A “School of Self-Learning”? The Teaching of Popular Music and Formal Learning Theory. *Revue Française de Pédagogie*, 185(4), 49–58. Cairn.info. <https://doi.org/10.4000/rfp.4292>
- Drew, P. (1997). ‘Open’ class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics*, 28(1), 69–101. [https://doi.org/10.1016/S0378-2166\(97\)89759-7](https://doi.org/10.1016/S0378-2166(97)89759-7)
- Drew, P. (2017). The interface between pragmatics and conversation analysis. *Pragmatics and Its Interfaces*, 59–84.
- Drew, P., & Heritage, J. (1992). Talk at work: Interaction in institutional settings. (*No Title*).
- Drew, P., & Kendrick, K. H. (2018). Searching for Trouble: Recruiting Assistance Through Embodied Action. *Social Interaction. Video-Based Studies of Human Sociality*, 1(1). <https://doi.org/10.7146/si.v1i1.105496>
- Dreyfus, H. L. (1987). From Socrates to Expert Systems: The Limits of Calculative Rationality. *Bulletin of the American Academy of Arts and Sciences*, 40(4), 15. <https://doi.org/10.2307/3823297>
- Dreyfus, H. L. (1990). *Being-in-the-World: A Commentary on Heidegger’s Being in Time, Division I* (Vol. 102, Issue 2, pp. 290–293). Bradford.
- Dreyfus, H. L. (1992). *What computers still can’t do: A critique of artificial reason*. MIT press.
- Dreyfus, H. L. (2001). Phenomenological description versus rational reconstruction. *Revue internationale de philosophie*, 216(2), 181–196. Cairn.info. <https://doi.org/10.3917/rip.216.0181>
- Dreyfus, H. L. (2002). Intelligence without representation – Merleau-Ponty’s critique of mental representation The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*, 1(4), 367–383. <https://doi.org/10.1023/A:1021351606209>

- Dreyfus, H. L. (2005). Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise. *Proceedings and Addresses of the American Philosophical Association*, 79(2), 47–65. JSTOR.
- Dreyfus, H. L. (2013). The Myth of the Pervasiveness of the Mental. *Mind, Reason, and Being-in-the-World: The McDowell-Dreyfus Debate*, 15–40.
- Dreyfus, H. L. (2014). *Skillful coping: Essays on the phenomenology of everyday perception and action*. OUP Oxford.
- Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine*. Simon and Schuster.
- Dreyfus, H. L., & Dreyfus, S. E. (1992). What artificial experts can and cannot do. *AI & Society*, 6(1), 18–26. <https://doi.org/10.1007/BF02472766>
- Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral Vision: Expertise in Real World Contexts. *Organization Studies*, 26(5), 779–792. <https://doi.org/10.1177/0170840605053102>
- Dreyfus, H. L., & Kelly, S. D. (2007). Heterophenomenology: Heavy-handed sleight-of-hand. *Phenomenology and the Cognitive Sciences*, 6(1–2), 45–55. <https://doi.org/10.1007/s11097-006-9042-y>
- Dreyfus, S. E. (2004). The Five-Stage Model of Adult Skill Acquisition. *Bulletin of Science, Technology and Society*, 24(3), 177–181.
- Dreyfus, S. E. (2014). System 0: The overlooked explanation of expert intuition. In M. Sinclair (Ed.), *Handbook of Research Methods on Intuition*. Edward Elgar Publishing. <https://doi.org/10.4337/9781782545996.00009>
- Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Operations Research Center, University of California, Berkeley Berkeley, CA.
- Due, B. L. (2019). *Laughing at the robot: Incongruent robot actions as laughables*. <https://doi.org/10.18420/muc2019-ws-640>
- Due, B. L. (2023). Laughing at the robot: Three types of laughables when interacting with Pepper. In P. Lang (Ed.), *Interacting with Robots and Social Agents*.
- Due, B. L., & Licoppe, C. (2021). Video-Mediated Interaction (VMI): Introduction to a special issue on the multimodal accomplishment of VMI institutional activities. *Social Interaction. Video-Based Studies of Human Sociality*, 3(3). <https://doi.org/10.7146/si.v3i3.123836>
- Due, B. L., & Lüchow, L. (2022). VUI-Speak: There is Nothing Conversational about “Conversational User Interfaces.” In *Social Robots In Institutional Interaction*. Bielefeld University Press.
- Dupret, B. (2014). The Concept of Law: A Wittgensteinian Approach with Some Ethnomethodological Specifications. In *Concepts of Law: Comparative, Jurisprudential, and Social Science Perspectives*. Ashgate.

- Duranti, A. (2007). Agency in Language. In *A Companion to Linguistic Anthropology* (pp. 449–473). <https://doi.org/10.1002/9780470996522.ch20>
- Dynel, M. (2010). On “Revolutionary Road”: A Proposal for Extending the Gricean Model of Communication to Cover Multiple Hearers. *Lodz Papers in Pragmatics*, 6(2), 283–304. <https://doi.org/doi:10.2478/v10016-010-0014-x>
- Dynel, M. (2011). Revisiting Goffman’s postulates on participant statuses in verbal interaction. *Language and Linguistics Compass*, 5(7), 454–465. <https://doi.org/10.1111/j.1749-818X.2011.00286.x>
- Echenique, A., Yamashita, N., Kuzuoka, H., & Hautasaari, A. (2014). Effects of Video and Text Support on Grounding in Multilingual Multiparty Audio Conferencing. *Proceedings of the 5th ACM International Conference on Collaboration across Boundaries: Culture, Distance & Technology*, 73–81. <https://doi.org/10.1145/2631488.2631497>
- Edison, W., & Geissler, G. (2003). Measuring attitudes towards general technology: Antecedents, hypotheses and scale development. *Journal of Targeting, Measurement and Analysis for Marketing*, 12, 137–156. <https://doi.org/10.1057/palgrave.jt.5740104>
- Eisenmann, C., & Lynch, M. (2021). Introduction to Harold Garfinkel’s Ethnomethodological “Misreading” of Aron Gurwitsch on the Phenomenal Field. *Human Studies*, 44(1), 1–17. <https://doi.org/10.1007/s10746-020-09564-1>
- Emirbayer, M., & Maynard, D. W. (2011). Pragmatism and Ethnomethodology. *Qualitative Sociology*, 34(1), 221–261. <https://doi.org/10.1007/s11133-010-9183-8>
- Emmeche, C. (2001). *Does a robot have an Umwelt? Reflections on the qualitative biosemiotics of Jakob von Uexküll.*
- Enfield, N. J., & Sidnell, J. (2017). On the concept of action in the study of interaction. *Discourse Studies*, 19(5), 515–535. <https://doi.org/10.1177/1461445617730235>
- Enfield, N. J., & Sidnell, J. (2021). Intersubjectivity is activity plus accountability. In N. Gontier, A. Lock, & C. Sinha (Eds.), *The Oxford Handbook of Human Symbolic Evolution* (1st ed.). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198813781.013.25>
- Etemad-Sajadi, R., Soussan, A., & Schöpfer, T. (2022). How Ethical Issues Raised by Human–Robot Interaction can Impact the Intention to use the Robot? *International Journal of Social Robotics*, 14. <https://doi.org/10.1007/s12369-021-00857-8>
- F. Yamaoka, H. I., T. Kanda, & Hagita, N. (2007). How contingent should a lifelike robot be? The relationship between contingency and complexity. *Connection Science*, 19(2), 143–162. <https://doi.org/10.1080/09540090701371519>
- Fischer, J. E., Reeves, S., Porcheron, M., & Sikveland, R. O. (2019). Progressivity for Voice Interface Design. *Proceedings of the 1st International Conference on Conversational User Interfaces*. <https://doi.org/10.1145/3342775.3342788>
- Fischer, K. (2007). The Role of Users’ Concepts of the Robot in Human-Robot Spatial Instruction. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), *Spatial*

Cognition V Reasoning, Action, Interaction (Vol. 4387, pp. 76–89). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75666-8_5

- Fischer, K. (2011). Interpersonal Variation in Understanding Robots as Social Actors. *Proceedings of the 6th International Conference on Human-Robot Interaction*, 53–60. <https://doi.org/10.1145/1957656.1957672>
- Fischer, K. (2021). Tracking Anthropomorphizing Behavior in Human-Robot Interaction. *J. Hum.-Robot Interact.*, 11(1). <https://doi.org/10.1145/3442677>
- Fishman, P. M. (1978). Interaction: The work women do. *Social Problems*, 25(4), 397–406. <https://doi.org/10.1525/sp.1978.25.4.03a00050>
- Fitzgerald, R., Housley, W., & Butler, C. W. (2009). Omnirelevance and interactional context. *Australian Journal of Communication*, 36(3), 45.
- Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(1), 10. <https://doi.org/10.1057/s41599-020-0494-4>
- Förster, F., Romeo, M., Holthaus, P., Wood, L. J., Dondrup, C., Fischer, J. E., Liza, F. F., Kaszuba, S., Hough, J., Nettet, B., Hernández García, D., Kontogiorgos, D., Williams, J., Özkan, E. E., Barnard, P., Berumen, G., Price, D., Cobb, S., Wiltschko, M., ... Kapetanios, E. (2023). Working with troubles and failures in conversation between humans and robots: Workshop report. *Frontiers in Robotics and AI*, 10, 1202306. <https://doi.org/10.3389/frobt.2023.1202306>
- Frid, E., Bresin, R., & Alexanderson, S. (2018). Perception of Mechanical Sounds Inherent to Expressive Gestures of a NAO Robot—Implications for Movement Sonification of Humanoids. *Proceedings of the 15th Sound and Music Computing Conference*. <https://doi.org/10.5281/zenodo.1422499>
- Fronemann, N., Pollmann, K., & Loh, W. (2022). Should my robot know what's best for me? Human–robot interaction between user experience and ethical design. *AI & SOCIETY*, 37(2), 517–533. <https://doi.org/10.1007/s00146-021-01210-3>
- Fuentes-Moraleda, L., Díaz-Pérez, P., Orea-Giner, A., Mazón, A. M., & Villacé-Molinero, T. (2020). Interaction between hotel service robots and humans: A hotel-specific Service Robot Acceptance Model (sRAM). *Tourism Management Perspectives*, 36, 100751. <https://doi.org/10.1016/j.tmp.2020.100751>
- Fuentes-Moraleda, L., Díaz-Pérez, P., Orea-Giner, A., Muñoz-Mazón, A., & Villacé-Molinero, T. (2020). Interaction between hotel service robots and humans: A hotel-specific Service Robot Acceptance Model (sRAM). *Tourism Management Perspectives*, 36, 100751. <https://doi.org/10.1016/j.tmp.2020.100751>
- Gao, G., Yamashita, N., Hautasaari, A., Echenique, A., & Fussell, S. R. (2014). Effects of public vs. Private automated transcripts on multiparty communication between native and non-native english speakers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. <https://api.semanticscholar.org/CorpusID:14193352>

- Gardner, R. (2007). "Broken" starts: Bricolage in turn starts in second language talk. In *Language Learning and Teaching as Social Inter-Action*. Palgrave Macmillan. https://link.springer.com/chapter/10.1057/9780230591240_5
- Gardner, R. (2015). Conversation Analysis and orientation to learning. *Journal of Applied Linguistics and Professional Practice*, 229–244. <https://doi.org/10.1558/japl.v5i3.229>
- Garfinkel, H. (1963). A conception of, and experiments with, "trust" as a condition of stable concerted actions. In *Motivation and Social Interaction: Cognitive Determinants*. Ronald Press.
- Garfinkel, H. (1964). Studies of the Routine Grounds of Everyday Activities. *Social Problems*, 11(3), 225–250. JSTOR. <https://doi.org/10.2307/798722>
- Garfinkel, H. (1967). *Studies in Ethnomethodology*. Polity Press.
- Garfinkel, H. (1991). Respecification: Evidence for locally produced, naturally accountable phenomena of order*, logic, reason, meaning, method, etc. in and as of the essential haecceity of immortal ordinary society (I): an announcement of studies. In *Ethnomethodology and the Human Sciences*. Cambridge University Press.
- Garfinkel, H. (1996). Ethnomethodology's Program. *Social Psychology Quarterly*, 59(1), 5. <https://doi.org/10.2307/2787116>
- Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Rowman & Littlefield Publishers.
- Garfinkel, H. (2006). *Seeing sociologically: The routine grounds of social action*. Paradigm.
- Garfinkel, H. (2019). Notes on language games as a source of methods for studying the formal properties of linguistic events¹. *European Journal of Social Theory*, 22(2), 148–174. <https://doi.org/10.1177/1368431018824733>
- Garfinkel, H., & Sacks, H. (1970). On formal structures of practical action. In *Theoretical Sociology: Perspectives and Developments*. Appleton-Century-Crofts.
- Garfinkel, H., & Wieder, D. L. (1992). Two incommensurable, asymmetrically alternate technologies of social analysis. In *Text in Context: Contributions to Ethnomethodology*. Sage Publications.
- Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Inc.
- Glas, D. F., Wada, K., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2017). Personal Greetings: Personalizing Robot Utterances Based on Novelty of Observed Behavior. *International Journal of Social Robotics*, 9(2), 181–198. <https://doi.org/10.1007/s12369-016-0385-4>
- Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3), 213–231.
- Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

- Goffman, E. (1967). *Interaction Ritual: Essays in Face-to-Face Behavior*. Aldine Publishing Company.
- Goffman, E. (1979a). Footing. *Semiotica*, 25(1–2), 1–30.
- Goffman, E. (1979b). Frame Analysis: An Essay on the Organization of Experience. *Philosophy and Phenomenological Research*, 39(4), 601–602.
- Goffman, E. (1983). The Interaction Order: American Sociological Association, 1982 Presidential Address. *American Sociological Review*, 48(1), 1. <https://doi.org/10.2307/2095141>
- Goodwin, C. (1981). *Conversational Organization: Interaction between Speakers and Hearers*. Irvington Publishers.
- Goodwin, C. (2007). Interactive footing. *Studies in Interactional Sociolinguistics*, 24, 16.
- Goody, J. (1975). *Literacy in traditional societies*. Cambridge University Press.
- Goody, J. (1977). *The domestication of the savage mind*. Cambridge University Press.
- Greiffenhagen, C., Xu, X., & Reeves, S. (2023). The Work to Make Facial Recognition Work. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–30. <https://doi.org/10.1145/3579531>
- Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Speech Acts* (pp. 41–58). BRILL. https://doi.org/10.1163/9789004368811_003
- Griffiths, S., Barnes, R., Britten, N., & Wilkinson, R. (2015). Multiple repair sequences in everyday conversations involving people with Parkinson's disease. *International Journal of Language & Communication Disorders*, 50(6), 814–829. <https://doi.org/10.1111/1460-6984.12178>
- Gurwitsch, A. (1964). *The Field of Consciousness*. Duquesne University Press.
- Hand, L. C., & Catlaw, T. J. (2019). Accomplishing the Public Encounter: A Case for Ethnomethodology in Public Administration Research. *Perspectives on Public Management and Governance*, 2(2), 125–137. <https://doi.org/10.1093/ppmgov/gvz004>
- Harjunpää, K. (2022). Repetition and prosodic matching in responding to pets' vocalizations. *Langage et société*, 176(2), 69–102. Cairn.info. <https://doi.org/10.3917/lis.176.0071>
- Harjunpää, K., Mondada, L., & Svinhufvud, K. (2018). The Coordinated Entry into Service Encounters in Food Shops: Managing Interactional Space, Availability, and Service During Openings. *Research on Language and Social Interaction*, 51(3), 271–291. <https://doi.org/10.1080/08351813.2018.1485231>
- Have, P. ten. (2005). The Notion of Member is the Heart of the Matter: On the Role of Membership Knowledge in Ethnomethodological Inquiry. *Historical Social Research / Historische Sozialforschung*, 30(1 (111)), 28–53. JSTOR.

- Have, P. ten. (2007). *Doing Conversation Analysis: A Practical Guide*. Sage Publications. <http://www.uk.sagepub.com/booksProdDesc.nav?prodId=Book229124>
- Heap, J. L. (1990). Applied ethnomethodology: Looking for the local rationality of reading activities. *Human Studies*, 13(1), 39–72. <https://doi.org/10.1007/BF00143040>
- Heath, C., Hindmarsh, J., & Luff, P. (2022). *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. <https://doi.org/10.4135/9781526435385>
- Heath, C., Lehn, D. V., & Osborne, J. (2005). Interaction and interactives: Collaboration and participation with computer-based exhibits. *Public Understanding of Science*, 14(1), 91–101. <https://doi.org/10.1177/0963662505047343>
- Heath, C., & Lehn, D. vom. (2004). Configuring reception: (Dis-)regarding the “spectator” in museums and galleries. *Theory, Culture & Society*, 21. <https://journals.sagepub.com/doi/abs/10.1177/0263276404047415>
- Heenan, B., Greenberg, S., Aghel-Manesh, S., & Sharlin, E. (2014). Designing social greetings in human robot interaction. *Proceedings of the 2014 Conference on Designing Interactive Systems*, 855–864. <https://doi.org/10.1145/2598510.2598513>
- Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing Acceptance of Assistive Social Agent Technology by Older Adults: The Almere Model. *International Journal of Social Robotics*, 2(4), 361–375. <https://doi.org/10.1007/s12369-010-0068-5>
- Heesen, R., & Fröhlich, M. (2022). Revisiting the human ‘interaction engine’: Comparative approaches to social action coordination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1859), 20210092. <https://doi.org/10.1098/rstb.2021.0092>
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, trans.). New York: Harper & Row.
- Heidegger, M. (1996). *Being and Time* (J. Macquarrie & E. Robinson, Trans.; Issue 56). SUNY Press.
- Henschel, A., Laban, G., & Cross, E. S. (2021). What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports*, 2(1), 9–19. <https://doi.org/10.1007/s43154-020-00035-0>
- Heritage, J. (1985). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson (Ed.), *Structures of Social Action* (pp. 299–345). Cambridge University Press; Cambridge Core. <https://doi.org/10.1017/CBO9780511665868.020>
- Heritage, J. (1989). Current developments in conversation analysis. In *Conversation: An Interdisciplinary Perspective*. Multilingual Matters.
- Heritage, J. (1998). *Conversation analysis and institutional talk: Analyzing distinctive turn-taking systems*. 3–17.
- Heritage, J. (2001). Goffman, Garfinkel and Conversation Analysis. In *Discourse Theory and Practices*. SAGE.

- Heritage, J. (2005). Conversation analysis and institutional talk. In R. Sanders & K. Fitch (Eds.), *Handbook of Language and Social Interaction* (pp. 103–146). Erlbaum.
- Heritage, J. (2007). Intersubjectivity and progressivity in person (and place) reference. In N. J. Enfield & T. Stivers (Eds.), *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives* (pp. 255–280). Cambridge University Press; Cambridge Core. <https://doi.org/10.1017/CBO9780511486746.012>
- Heritage, J. (2013a). *Garfinkel and ethnomethodology*. John Wiley & Sons.
- Heritage, J. (2013b). Turn-initial position and some of its occupants. *Journal of Pragmatics*, 57, 331–337. <https://doi.org/10.1016/j.pragma.2013.08.025>
- Heritage, J. (2018). The ubiquity of epistemics: A rebuttal to the ‘epistemics of epistemics’ group. *Discourse Studies*, 20(1), 14–56. <https://doi.org/10.1177/1461445617734342>
- Heritage, J., & Clayman, S. E. (2012). Melvin Pollner: A view from the suburbs. *American Sociologist*, 43. <http://link.springer.com/article/10.1007/s12108-012-9148-3>
- Heritage, J., & Watson, R. (1980). Formulations as conversational objects. *Semiotica*, 3, 245–262.
- Hirschauer, S. (2007). Putting things into words. Ethnographic description and the silence of the social. *Human Studies*, 29(4), 413–441. <https://doi.org/10.1007/s10746-007-9041-1>
- Hoeppe, G. (2023). Learning from Harold Garfinkel’s Studies of Work in the Sciences. *Michael Lynch (Ed.), Harold Garfinkel: Studies of Work in the Science*. New York: Routledge 2022, 207 S., Kt., 40,90 €, 46(2), 120–129. <https://doi.org/10.1515/srsr-2023-2023>
- Hoey, E. (2017). Sequence recompletion: A practice for managing lapses in conversation. *Journal of Pragmatics*, 109, 47–63. <https://doi.org/10.1016/j.pragma.2016.12.008>
- Hoey, E. (2020a). Sacks, Silence, and Self-(de)selection. In R. J. Smith, R. Fitzgerald, & W. Housley (Eds.), *On Sacks: Methodology, Materials, and Inspirations* (1st ed.). Routledge. <https://doi.org/10.4324/9780429024849>
- Hoey, E. (2020b). Silence and social interaction. In E. Hoey, *When Conversation Lapses* (pp. 1–38). Oxford University Press. <https://doi.org/10.1093/oso/9780190947651.003.0001>
- Hoey, E., & Kendrick, K. (2017a). *Conversation Analysis*.
- Hoey, E., & Kendrick, K. H. (2017b). Conversation analysis. *Research Methods in Psycholinguistics: A Practical Guide*, 151–173.
- Hoffding, S. (2014). What is Skilled Coping?: Experts on Expertise. *Journal of Consciousness Studies*, 21(9–10), 49–73.
- Holthaus, P., Pitsch, K., & Wachsmuth, S. (2011). How Can I Help?: Spatial Attention Strategies for a Receptionist Robot. *International Journal of Social Robotics*, 3(4), 383–393. <https://doi.org/10.1007/s12369-011-0108-9>

- Holthaus, P., & Wachsmuth, S. (2021). It was a Pleasure Meeting You: Towards a Holistic Model of Human–Robot Encounters. *International Journal of Social Robotics*, 13(7), 1729–1745. <https://doi.org/10.1007/s12369-021-00759-9>
- Hopper, R. (2005). A cognitive agnostic in conversation analysis: When do strategies affect spoken interaction? In H. te Molder & J. Potter (Eds.), *Conversation and Cognition* (1st ed., pp. 134–158). Cambridge University Press. <https://doi.org/10.1017/CBO9780511489990.007>
- Houtkoop-Steenstra, H. (1995). Meeting Both Ends: Between Standardization and Recipient Design in Telephone Survey Interviews. In P. Ten Have & G. Psathas (Eds.), *Situated Order: Studies in the Social Organization of Talk and Embodied Activities* (pp. 91–106). International Institute for Ethnomethodology and Conversation Analysis, and University Press of America.
- Huettenrauch, H., Eklundh, K., Green, A., & Topp, E. (2006). Investigating Spatial Relationships in Human-Robot Interaction. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5052–5059. <https://doi.org/10.1109/IROS.2006.282535>
- Hughson, J., & Englis, D. (2002). Accounting for Experience: Phenomenological Argots and Sportive Life-Worlds. *Indo-Pacific Journal of Phenomenology*, 2(2), 1–10. <https://doi.org/10.1080/20797222.2002.11433877>
- Husserl, E. (2012). *Ideas* (0 ed.). Routledge. <https://doi.org/10.4324/9780203120330>
- Husserl, E. (2013). *Cartesian meditations: An introduction to phenomenology*. Springer Science & Business Media.
- Ivarsson, J., & Lindwall, O. (2023). Suspicious Minds: The Problem of Trust and Conversational Agents. *Computer Supported Cooperative Work (CSCW)*, 32(3), 545–571. <https://doi.org/10.1007/s10606-023-09465-8>
- Jackson, R. B., & Williams, T. (2021). A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI*, 8. <https://doi.org/10.3389/frobt.2021.687726>
- Jarske, S., Raudaskoski, S., & Kaipainen, K. (2020). The “social” of the socially interactive robot: Rethinking human-robot interaction through ethnomethodology. *Culturally Sustainable Social Robotics: Proceedings of Robophilosophy 2020*, 194–203. <https://doi.org/10.3233/FAIA200915>
- Jefferson, G. (1972). Side sequences. In *Studies in Social Interaction*. Free Press.
- Jefferson, G. (1982). *Two explorations of the organization of overlapping talk in conversation*. Tilburg University, Department of Language and Literature.
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In *Conversation Analysis: Studies from the First Generation*. John Benjamins.
- Jefferson, G. (2017). Remarks on non-correction in conversation. In P. Drew, J. Bergmann, & G. Jefferson (Eds.), *Repairing the Broken Surface of Talk: Managing Problems in*

Speaking, Hearing, and Understanding in Conversation (pp. 313–329). Oxford University Press.

Jones, R. A. (2017). What makes a robot 'social'? *Social Studies of Science*, 47(4), 556–579. <https://doi.org/10.1177/0306312717704722>

Kaipainen, K., Ahtinen, A., & Hiltunen, A. (2018). Nice surprise, more present than a machine: Experiences evoked by a social robot for guidance and edutainment at a city service point. *Proceedings of the 22nd International Academic Mindtrek Conference*. <https://api.semanticscholar.org/CorpusID:53036513>

Kamino, W., & Sabanovic, S. (2023). *Coffee, Tea, Robots?: The Performative Staging of Service Robots in "Robot Cafes" in Japan*. 183–191. <https://doi.org/10.1145/3568162.3576967>

Kato, Y., Kanda, T., & Ishiguro, H. (2015). May I Help You? Design of Human-like Polite Approaching Behavior. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 35–42. <https://doi.org/10.1145/2696454.2696463>

Keevallik, L. (2018). Sequence Initiation or Self-Talk? Commenting on the Surroundings While Mucking out a Sheep Stable. *Research on Language and Social Interaction*, 51(3), 313–328. <https://doi.org/10.1080/08351813.2018.1485233>

Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*. (pp. xi, 292). Cambridge University Press.

Kendrick, K. H. (2015). The intersection of turn-taking and repair: The timing of other-initiations of repair in conversation. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00250>

Kendrick, K. H. (2017). Using Conversation Analysis in the Lab. *Research on Language and Social Interaction*, 50(1), 1–11. <https://doi.org/10.1080/08351813.2017.1267911>

Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., Hoey, E., Hoymann, G., Manrique, E., Rossi, G., & Levinson, S. C. (2020). Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics*, 168, 119–138. <https://doi.org/10.1016/j.pragma.2020.06.009>

Kerrison, A. T. (2018). *We're All Behind You: The Co-Construction of Turns and Sequences-at-Cheering*. Ulster University.

Kidwell, M. (2005). Gaze as Social Control: How Very Young Children Differentiate "The Look" From a "Mere Look" by Their Adult Caregivers. *Research on Language & Social Interaction*, 38(4), 417–449. https://doi.org/10.1207/s15327973rlsi3804_2

Kim, H. (2002). The form and function of next-turn repetition in English conversation. *LANGUAGE RESEARCH-SEOUL*, 38(1), 51–82.

Kim, K.-M. (1999). The Management of Temporality: Ethnomethodology as Historical Reconstruction of Practical Action. *The Sociological Quarterly*, 40(3), 505–523. <https://doi.org/10.1111/j.1533-8525.1999.tb01731.x>

- Krägeloh, C. U., Bharatharaj, J., Sasthan Kutty, S. K., Nirmala, P. R., & Huang, L. (2019). Questionnaires to Measure Acceptability of Social Robots: A Critical Review. *Robotics*, 8(4). <https://doi.org/10.3390/robotics8040088>
- Kristiansen, E. D., & Rasmussen, G. (2021). Eye-tracking Recordings as Data in EMCA Studies: Exploring Possibilities and Limitations. *Social Interaction: Video-Based Studies of Human Sociality*, 4. <https://tidsskrift.dk/socialinteraction/article/view/121776>
- Krummheuer, A. L. (2015a). Technical Agency in Practice: The enactment of artefacts as conversation partners, actants and opponents. *PsychNology J.*, 13, 179–202.
- Krummheuer, A. L. (2015b). Users, Bystanders and Agents: Participation Roles in Human-Agent Interaction. In *Human-Computer Interaction—INTERACT 2015*. Springer International Publishing. https://doi.org/10.1007/978-3-319-22723-8_19
- Küttner, U.-A. (2020). Tying Sequences Together with the [That's + Wh-Clause] Format: On (Retro-)Sequential Junctures in Conversation. *Research on Language and Social Interaction*, 53(2), 247–270. <https://doi.org/10.1080/08351813.2020.1739422>
- Laforest, M. (2009). Complaining in front of a witness: Aspects of blaming others for their behaviour in multi-party family interactions. *Complaining in Interaction*, 41(12), 2452–2464. <https://doi.org/10.1016/j.pragma.2008.09.043>
- Laurier, E., Mazé, R., & Lundin, J. (2005). Putting the Dog Back in the Park: Animal and Human Mind-in-Action. *Mind Culture and Activity*, 13, 2–24. https://doi.org/10.1207/s15327884mca1301_2
- Lee, M. K., Kiesler, S., & Forlizzi, J. (2010). Receptionist or information kiosk: How do people talk with a robot? *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work - CSCW '10*, 31. <https://doi.org/10.1145/1718918.1718927>
- Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010* (p. 210). <https://doi.org/10.1109/HRI.2010.5453195>
- Lehn, D., Heath, C., & Hindmarsh, J. (2001). Exhibiting Interaction: Conduct and Collaboration in Museums and Galleries. *Symbolic Interaction*, 24, 189–216. <https://doi.org/10.1525/si.2001.24.2.189>
- Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, 5(2), 291–308. <https://doi.org/10.1007/s12369-013-0178-y>
- Leiter, B. (1996). Heidegger and the Theory of Adjudication. *Yale LJ*, 106, 253.
- Lempert, M. (2013). No ordinary ethics. *Anthropological Theory*, 13(4), 370–393. <https://doi.org/10.1177/1463499613505571>
- Lerner, G. H. (2002). Turn-sharing: The choral co-production of talk-in-interaction. In *The Language of Turn and Sequence*. Oxford University Press USA.

- Lerner, G. H., & Kitzinger, C. (2019). Well-Prefacing in the Organization of Self-Initiated Repair. *Research on Language and Social Interaction*, 52(1), 1–19. <https://doi.org/10.1080/08351813.2019.1572376>
- Levinson, S. C. (1983). *Pragmatics*. Cambridge university press.
- Levinson, S. C. (1987). Minimization and conversational inference. In J. Verschueren & M. Bertuccelli Papi (Eds.), *Pragmatics & Beyond Companion Series* (Vol. 5, p. 61). John Benjamins Publishing Company. <https://doi.org/10.1075/pbcs.5.10lev>
- Levinson, S. C. (1988). Putting linguistics on a proper footing: Explorations in Goffman's concepts of participation. In *Erving Goffman: Exploring the interaction order*. (pp. 161–227). Northeastern University Press.
- Levinson, S. C. (2012). Action Formation and Ascription. In *The Handbook of Conversation Analysis* (pp. 101–130). <https://doi.org/10.1002/9781118325001.ch6>
- Levinson, S. C. (2020a). *On technologies of the intellect: Goody Lecture 2020*.
- Levinson, S. C. (2020b). On the human "interaction engine". In *Roots of human sociality* (pp. 39–69). Routledge.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00731>
- Lieberman, K. (1980). Ambiguity and gratuitous concurrence in inter-cultural communication. *Human Studies*, 3(1), 65–85. <https://doi.org/10.1007/BF02331801>
- Licoppe, C. (2017). Skype appearances, multiple greetings and 'coucou': The sequential organization of video-mediated conversation openings. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 27(3), 351–386. <https://doi.org/10.1075/prag.27.3.03lic>
- Licoppe, C., & Rollet, N. (2020). « Je dois y aller ». Analyses de séquences de clôtures entre humains et robot. *Réseaux*, N°220-221(2), 151. <https://doi.org/10.3917/res.220.0151>
- Ligthart, M. E. U., Neerinx, M. A., & Hindriks, K. V. (2022). Getting acquainted: First steps for child-robot relationship formation. *Frontiers in Robotics and AI*, 9. <https://www.frontiersin.org/articles/10.3389/frobt.2022.853665>
- Lindemann, G. (2005). The Analysis of the Borders of the Social World: A Challenge for Sociological Theory. *Journal for the Theory of Social Behaviour*, 35, 69–98. <https://doi.org/10.1111/j.0021-8308.2005.00264.x>
- Linell, P. (2004). *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.
- Linell, P., & Lindström, J. (2016). Partial intersubjectivity and sufficient understandings for current practical purposes: On a specialized practice in Swedish conversation. *Nordic Journal of Linguistics*, 39(2), 113–133. Cambridge Core. <https://doi.org/10.1017/S0332586516000081>

- Lipp, B. (2022). Caring for robots: How care comes to matter in human-machine interfacing. *Social Studies of Science*, 03063127221081446. <https://doi.org/10.1177/03063127221081446>
- List, C. (2019). Levels: Descriptive, Explanatory, and Ontological. *Noûs*, 53(4), 852–883. <https://doi.org/10.1111/nous.12241>
- Liu, P., Glas, D. F., Kanda, T., & Ishiguro, H. (2016). Data-Driven HRI: Learning Social Behaviors by Example From Human–Human Interaction. *IEEE Transactions on Robotics*, 32(4), 988–1008. <https://doi.org/10.1109/TRO.2016.2588880>
- Livingston, E. (2008). Context and detail in studies of the witnessable social order: Puzzles, maps, checkers, and geometry. *Journal of Pragmatics*, 40(5), 840–862. <https://doi.org/10.1016/j.pragma.2007.09.009>
- Livingston, E. (2017). *Ethnographies of reason*. Routledge.
- Luhmann, N. (2013). *A sociological theory of law*. Routledge.
- Lynch, M. (2000). The ethnomethodological foundations of conversation analysis. *Text - Interdisciplinary Journal for the Study of Discourse*, 20(4). <https://doi.org/10.1515/text.1.2000.20.4.517>
- Lynch, M. (2016). Radical ethnomethodology. *Position Paper for Workshop on Radical Ethnomethodology, Manchester Metropolitan University, UK*, 22(23), 6.
- Lynch, M. (2019). Garfinkel, Sacks and Formal Structures: Collaborative Origins, Divergences and the History of Ethnomethodology and Conversation Analysis. *Human Studies*, 42(2), 183–198. <https://doi.org/10.1007/s10746-019-09510-w>
- Lynch, M. (2022). Comment on Martin Hammersley, “Is ‘Representation’ a Folk Term?” *Philosophy of the Social Sciences*, 52(4), 258–267. <https://doi.org/10.1177/00483931221091555>
- Lynch, M., & Eisenmann, C. (2022). Transposing Gestalt Phenomena from Visual Fields to Practical and Interactional Work: Garfinkel’s and Sacks’ Social Praxeology. *Philosophia Scientiæ*, 26–3(3), 95–122. Cairn.info. <https://doi.org/10.4000/philosophiascien>
- Lynch, M., & Garfinkel, H. (2022). *Harold Garfinkel: Studies of Work in the Sciences* (1st ed.). Routledge. <https://doi.org/10.4324/9781003172611>
- Macbeth, D. (2018). Does ethnomethodological CA have a “soft underbelly”? *Ethnographic Studies*, 15. <https://zenodo.org/record/1475761>
- Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, 15(1), 19–44. <https://doi.org/10.1080/00437956.1959.11659682>
- Mahdi, H., Akgun, S. A., Saleh, S., & Dautenhahn, K. (2022). A survey on the design and evolution of social robots—Past, present and future. *Robotics and Autonomous Systems*, 156, 104193. <https://doi.org/10.1016/j.robot.2022.104193>

- Makatchev, M., Lee, M. K., & Simmons, R. (2009). Relating initial turns of human-robot dialogues to discourse. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction - HRI '09*, 321. <https://doi.org/10.1145/1514095.1514196>
- Maynard, D. W. (2003). *Bad news, good news: Conversational order in everyday talk and clinical settings*. University of Chicago Press.
- Maynard, D. W. (2006). Cognition on the ground. *Discourse Studies*, 8. <https://journals.sagepub.com/doi/abs/10.1177/1461445606059560>
- Maynard, D. W. (2011). On “Interactional Semantics” and Problems of Meaning. *Human Studies - HUM STUD*, 34, 199–207. <https://doi.org/10.1007/s10746-011-9188-7>
- Maynard, D. W., & Heritage, J. (2023). Ethnomethodology’s Legacies and Prospects. *Annual Review of Sociology*, 49. <https://doi.org/10.1146/annurev-soc-020321-033738>
- McDermott, D. (1976). Artificial Intelligence Meets Natural Stupidity. *SIGART Bull.*, 57, 4–9. <https://doi.org/10.1145/1045339.1045340>
- Meier Zu Verl, C., & Meyer, C. (2024). Ethnomethodological ethnography: Historical, conceptual, and methodological foundations. *Qualitative Research*, 24(1), 11–31. <https://doi.org/10.1177/14687941221129798>
- Meyer, C. (2019). Ethnomethodology’s Culture. *Human Studies*, 42(2), 281–303. <https://doi.org/10.1007/s10746-019-09515-5>
- Mitchell, M. (2021). Why AI is Harder than We Think. *Proceedings of the Genetic and Evolutionary Computation Conference*, 3. <https://doi.org/10.1145/3449639.3465421>
- Mlynář, J., & Arminen, I. (2023). Respecifying social change: The obsolescence of practices and the transience of technology. *Frontiers in Sociology*, 8. <https://doi.org/10.3389/fsoc.2023.1222734>
- Mondada, L. (2002). Pratiques de transcription et effets de catégorisation. *Cahiers de Praxématique*, 39, 45–75. <https://doi.org/10.4000/praxematique.1835>
- Mondada, L. (2009). Emergent focused interactions in public places: A systematic analysis of the multimodal achievement of a common interactional space. *Journal of Pragmatics*, 41(10), 1977–1997. <https://doi.org/10.1016/j.pragma.2008.09.019>
- Mondada, L. (2011). Understanding as an embodied, situated and sequential achievement in interaction. *Journal of Pragmatics*, 43(2), 542–552. <https://doi.org/10.1016/j.pragma.2010.08.019>
- Mondada, L. (2014). Instructions in the operating room: How the surgeon directs their assistant’s hands. *Discourse Studies*, 16(2), 131–161. JSTOR.
- Mondada, L. (2015). Ouverture et préouverture des réunions visiophoniques. *Réseaux*, 194, 39–84. <https://doi.org/10.3917/res.194.0039>
- Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, 20(3), 336–366. https://doi.org/10.1111/josl.1_12177

- Mondada, L. (2017). Walking and talking together: Questions/answers and mobile participation in guided visits. *Social Science Information*, 56(2), 220–253. <https://doi.org/10.1177/0539018417694777>
- Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Social Interaction*, 51(1), 85–106. <https://doi.org/10.1080/08351813.2018.1413878>
- Mondada, L. (2019). Transcribing silent actions: A multimodal approach of sequence organization. *Social Interaction. Video-Based Studies of Human Sociality*, 2(1). <https://doi.org/10.7146/si.v2i1.113150>
- Mondada, L. (2021). How Early can Embodied Responses be? Issues in Time and Sequentiality. *Discourse Processes*, 58(4), 397–418. <https://doi.org/10.1080/0163853X.2020.1871561>
- Mondada, L. (2024). Micro-sequentiality. In A. Gubina, E. Hoey, & C. Wesley Raymond (Eds.), *Encyclopedia of Terminology for Conversation Analysis and Interactional Linguistics*. International Society for Conversation Analysis (ISCA).
- Mondada, L., Bänninger, J., Bouaouina, S. A., Camus, L., Gauthier, G., Hänggi, P., Koda, M., Svensson, H., & Tekin, B. S. (2020). Human sociality in the times of the Covid-19 pandemic: A systematic examination of change in greetings. *Journal of Sociolinguistics*, 24(4), 441–468. <https://doi.org/10.1111/josl.12433>
- Mondada, L., & Meguerditchian, A. (2022). Initiating interaction and the offering of greetings between baboons: Sequential organization and mutual physical disposition. *Langage et Société*, 176(2), 127–160. Cairn.info. <https://doi.org/10.3917/lis.176.0129>
- Mondémé, C. (2016a). Extending the Notion of “social order” to Human/animal Interaction. An Ethnomethodological Approach. *L'Année Sociologique*, 66(2), 319–350. Cairn.info.
- Mondémé, C. (2016b). Extension de la question de « l'ordre social » aux interactions hommes / animaux. Une approche ethnométhodologique. *L'Année sociologique*, 66(2), 319–350. Cairn.info. <https://doi.org/10.3917/anso.162.0319>
- Mondémé, C. (2019). *La socialité interspécifique: Une analyse multimodale des interactions homme-chien*. LL, Lambert-Lucas.
- Mondémé, C. (2022). Why study turn-taking sequences in interspecies interactions? *Journal for the Theory of Social Behaviour*, 52(1), 67–85. <https://doi.org/10.1111/jtsb.12295>
- Mondémé, C. (2023). Sequence organization in human–animal interaction. An exploration of two canonical sequences. *Journal of Pragmatics*, 214, 73–88. <https://doi.org/10.1016/j.pragma.2023.06.006>
- Moore, R. J., & Arar, R. (2018). Conversational UX Design: An Introduction. In R. J. Moore, M. H. Szymanski, R. Arar, & G.-J. Ren (Eds.), *Studies in Conversational UX Design* (pp. 1–16). Springer International Publishing. https://doi.org/10.1007/978-3-319-95579-7_1

- Morita, E., & Takagi, T. (2018). Marking “commitment to undertaking of the task at hand”: Initiating responses with eeto in Japanese conversation. *Journal of Pragmatics*, 124, 31–49. <https://doi.org/10.1016/j.pragma.2017.12.002>
- Mortensen, K., & Hazel, S. (2014). Moving into interaction—Social practices for initiating encounters at a help desk. *Journal of Pragmatics*, 62, 46–67. <https://doi.org/10.1016/j.pragma.2013.11.009>
- Mubin, O., Shahid, S., Van De Sande, E., Krahmer, E., Swerts, M., Bartneck, C., & Feijs, L. (2010). Using child-robot interaction to investigate the user acceptance of constrained and artificial languages. *19th International Symposium in Robot and Human Interactive Communication*, 588–593. <https://doi.org/10.1109/ROMAN.2010.5598731>
- Muhle, F. (2015). Encounters at the borders of the social world. Theoretical and methodological considerations on a new type of sociological research. In *Yearbook 2013 of the Institute for Advanced Studies on Science, Technology and Society*.
- Nguyen, H., & Nguyen, T.-M. (2022). Conversation Analysis and Membership Categorization Analysis. In *The Routledge Handbook of Second Language Acquisition and Sociolinguistics*. Routledge. <https://doi.org/10.4324/9781003017325-24>
- Nishizaka, A. (1995). The interactive constitution of interculturality: How to be a Japanese with words. *Human Studies*, 18, 301–326.
- Nöth, W. (2001). *Semiosis and the Umwelt of a robot*.
- Oloff, F. (2009). *Contribution à l'étude systématique de l'organisation des tours de parole: Les chevauchements en français et en allemand*.
- Pelikan, H. (2020). Intermediate-Level Knowledge: A Conversation Analysis Perspective. In *First international workshop on Designerly HRI Knowledge. Held in conjunction with the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020)*. <http://hridesign.eu/assets/pdf/Pelikan.pdf>
- Pelikan, H. (2021). Why autonomous driving is so hard: The social dimension of traffic. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 81–85.
- Pelikan, H. (2023). Transcribing human–robot interaction. In P. Haddington, T. Eilittä, A. Kamunen, L. Kohonen-Aho, T. Oittinen, I. Rautiainen, & A. Vatanen, *Ethnomethodological Conversation Analysis in Motion* (1st ed., pp. 42–62). Routledge. <https://doi.org/10.4324/9781003424888-4>
- Pelikan, H., & Broth, M. (2016). Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4921–4932. <https://doi.org/10.1145/2858036.2858478>
- Pelikan, H., Broth, M., & Keevallik, L. (2020). “Are You Sad, Cozmo?": How Humans Make Sense of a Home Robot's Emotion Displays. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 461–470). Association for Computing Machinery. <https://doi.org/10.1145/3319502.3374814>

- Pelikan, H., Broth, M., & Keevallik, L. (2022). When a Robot Comes to Life: The Interactional Achievement of Agency as a Transient Phenomenon. *Social Interaction. Video-Based Studies of Human Sociality*, 5(3). <https://doi.org/10.7146/si.v5i3.129915>
- Pelikan, H., Reeves, S., & Cantarutti, M. N. (2024). Encountering Autonomous Robots on Public Streets. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 561–571. <https://doi.org/10.1145/3610977.3634936>
- Peña, A. (2010). The Dreyfus model of clinical problem-solving skills acquisition: A critical perspective. *Medical Education Online*, 15(1), 4846. <https://doi.org/10.3402/meo.v15i0.4846>
- Perrin, L., Deshaies, D., & Paradis, C. (2003). Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35(12), 1843–1860. [https://doi.org/10.1016/S0378-2166\(03\)00117-6](https://doi.org/10.1016/S0378-2166(03)00117-6)
- Peyrot, M. (1982a). Understanding ethnomethodology: A remedy for some common misconceptions. *Human Studies*, 5(1), 261–283. <https://doi.org/10.1007/BF02127681>
- Peyrot, M. (1982b). Understanding ethnomethodology: A remedy for some common misconceptions. *Human Studies*, 5(1), 261–283. <https://doi.org/10.1007/BF02127681>
- Pickup, A. (2016). Critical Inquiry as Virtuous Truth-Telling: Implications of Phronesis and Parrhesia. *Critical Questions in Education*, 7(3), 178–193.
- Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: Bridging the gap between human and animal communication. *Proceedings of the Royal Society B: Biological Sciences*, 285(1880), 20180598. <https://doi.org/10.1098/rspb.2018.0598>
- Pillet-Shore, D. (2012). Greeting: Displaying Stance Through Prosodic Recipient Design. *Research on Language & Social Interaction*, 45(4), 375–398. <https://doi.org/10.1080/08351813.2012.724994>
- Pillet-Shore, D. (2018). How to Begin. *Research on Language and Social Interaction*, 51(3), 213–231. <https://doi.org/10.1080/08351813.2018.1485224>
- Pillet-Shore, D. M. (2008). *Coming together: Creating and maintaining social relationships through the openings of face-to-face interactions*. University of California, Los Angeles.
- Pilnick, A., O'Brien, R., Beeke, S., Goldberg, S., & Harwood, R. (2021). Avoiding repair, maintaining face: Responding to hard-to-interpret talk from people living with dementia in the acute hospital. *Social Science & Medicine*, 282, 114156. <https://doi.org/10.1016/j.socscimed.2021.114156>
- Pitsch, K. (2016). Limits and opportunities for mathematizing communicational conduct for social robotics in the real world? Toward enabling a robot to make use of the human's competences: Response to the Question 2: "Are there limits to mathematization?". In Gesa Lindemann, (this volume), *Social interaction with robots—three questions*. *AI & SOCIETY*, 31(4), 587–593. <https://doi.org/10.1007/s00146-015-0629-0>

- Pitsch, K. (2020). Answering a robot's questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot. *Réseaux*, 220–221(2–3), 113–150. Cairn.info.
- Pitsch, K., & Koch, B. (2010). How infants perceive the toy robot Pleo. An exploratory case study on infant-robot-interaction. *HRI 2010*.
- Pitsch, K., Kuzuoka, H., Suzuki, Y., Sussenbach, L., Luff, P., & Heath, C. (2009). "The first five seconds": Contingent stepwise entry into an interaction as a means to secure sustained engagement in HRI. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 985–991. <https://doi.org/10.1109/ROMAN.2009.5326167>
- Pitsch, K., Vollmer, A.-L., & Mühlig, M. (2013). Robot feedback shapes the tutor's presentation: How a robot's online gaze strategies lead to micro-adaptation of the human's conduct. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 14(2), 268–296. <https://doi.org/10.1075/is.14.2.06pit>
- Pitsch, K., Vollmer, A.-L., Rohlfing, K. J., Fritsch, J., & Wrede, B. (2014). Tutoring in adult-child interaction: On the loop of the tutor's action modification and the recipient's gaze. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 15(1), 55–98. <https://doi.org/10.1075/is.15.1.03pit>
- Polak, P., & Krzanowski, R. (2020). Ethics in autonomous robots as philosophy in silico: The study case of phronetic machine ethics. *Logos i Ethos*, 52.
- Pollner, M. (1974). Sociological and common sense models of the labelling process. In *Ethnomethodology: Selected Readings*. Penguin.
- Pollner, M. (2012). The End(s) of Ethnomethodology. *The American Sociologist*, 43(1), 7–20. <https://doi.org/10.1007/s12108-011-9144-z>
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3173574.3174214>
- Psathas, G. (1980). Approaches to the Study of the World of Everyday Life. *Human Studies*, 3(1), 3–17. JSTOR.
- Psathas, G. (1989). *Phenomenology and sociology: Theory and research* (University Press of America).
- Psathas, G., & Anderson, T. (1990). The 'practices' of transcription in conversation analysis. *Semiotica*, 78(1–2), 75–100.
- Randall, D., Rouncefield, M., & Tolmie, P. (2021). Ethnography, CSCW and Ethnomethodology. *Computer Supported Cooperative Work (CSCW)*, 30(2), 189–214. <https://doi.org/10.1007/s10606-020-09388-8>
- Rasmussen, G. (2019). Emergentism. In J. S. Damico & M. J. Ball (Eds.), *The SAGE Encyclopedia of Human Communication Sciences and Disorders* (pp. 684–685). SAGE Publications. <https://doi.org/10.4135/9781483380810.n228>

- Rawls, A. (2003). Harold Garfinkel. *The Blackwell Companion to Major Contemporary Social Theorists*, 122–153.
- Rawls, A. W. (2002). Editor's introduction. In *Ethnomethodology's Program: Working Out Durkheim's Aphorism*. Rowman & Littlefield.
- Reeve, C. D. C. (1992). *Practices of reason: Aristotle's Nicomachean ethics* (Vol. 105, Issue 2, pp. 411–412). Oxford University Press.
- Reeves, S. (2022). *Navigating Incommensurability Between Ethnomethodology, Conversation Analysis, and Artificial Intelligence*. <https://doi.org/10.48550/ARXIV.2206.11899>
- Relieu, M. (2007). La téléprésence, ou l'autre visiophonie. *Réseaux*, 144(5), 183–223. Cairn.info.
- Relieu, M., Sahin, M., & Francillon, A. (2020). Une approche configurationnelle des leures conversationnels. *Réseaux*, N°220-221(2), 81. <https://doi.org/10.3917/res.220.0081>
- Relieu, M., Salembier, P., & Theureau, J. (2004). Introduction au numéro spécial «activité et action/cognition située». *Activités*, 1(1–2).
- Riddoch, K. A., & Cross, Emily. S. (2021). “Hit the Robot on the Head With This Mallet” – Making a Case for Including More Open Questions in HRI Research. *Frontiers in Robotics and AI*, 8, 603510. <https://doi.org/10.3389/frobt.2021.603510>
- Roberts, F. (2004). Speaking to and for Animals in a Veterinary Clinic: A Practice for Managing Interpersonal Interaction. *Research on Language and Social Interaction*, 37(4), 421–446. https://doi.org/10.1207/s15327973rlsi3704_2
- Robinson, J. D. (1998). Getting Down to Business Talk, Gaze, and Body Orientation During Openings of Doctor-Patient Consultations. *Human Communication Research*, 25(1), 97–123. <https://doi.org/10.1111/j.1468-2958.1998.tb00438.x>
- Robinson, J. D. (2016). Accountability in Social Interaction. In J. D. Robinson (Ed.), *Accountability in Social Interaction* (pp. 1–44). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780190210557.003.0001>
- Robles, J. S. (2015). Morality in discourse. In *International Encyclopedia of Language and Social Interaction* (pp. 132–137). John Wiley & Sons.
- Rollet, N., & Clavel, C. (2020). “Talk to you later”: Doing social robotics with conversation analysis. Towards the development of an automatic system for the prediction of disengagement. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 21(2), 268–292. <https://doi.org/10.1075/is.19001.roll>
- Rollet, N., Jain, V., Licoppe, C., & Devillers, L. (2017). Towards Interactional Symbiosis: Epistemic Balance and Co-presence in a Quantified Self Experiment. In L. Gamberini, A. Spagnoli, G. Jacucci, B. Blankertz, & J. Freeman (Eds.), *Symbiotic Interaction* (pp. 143–154). Springer International Publishing.
- Rossano, F. (2013). Sequence organization and timing of bonobo mother-infant interactions. *Interaction Studies*, 14(2), 160–189.

- Rossi, G. (2020). Other-repetition in conversation across languages: Bringing prosody into pragmatic typology. *Language in Society*, 49(4), 495–520. Cambridge Core. <https://doi.org/10.1017/S0047404520000251>
- Rudaz, D., & Licoppe, C. (2024). “Playing the Robot’s Advocate”: Bystanders’ Descriptions of a Robot’s Conduct in Public Settings. *Discourse and Communication*, 18(4).
- Rudaz, D., Tatarian, K., Stower, R., & Licoppe, C. (2023). From Inanimate Object to Agent: Impact of Pre-Beginnings on the Emergence of Greetings with a Robot. *J. Hum.-Robot Interact.*, 12(3). <https://doi.org/10.1145/3575806>
- Sabanovic, S., & Chang, W. (2016). Socializing robots: Constructing robotic sociality in the design and use of the assistive robot PARO. *AI & SOCIETY*, 31. <https://doi.org/10.1007/s00146-015-0636-1>
- Sacks, H. (1968). *Lectures on conversation. Vol. 2. Gail Jefferson (toim.)*. Cambridge: Blackwell Publishers.
- Sacks, H. (1975). Everyone has to lie. In *Sociocultural Dimensions of Language Use*. Academic Press.
- Sacks, H. (1984a). Notes on methodology. In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press.
- Sacks, H. (1984b). On doing ‘being ordinary.’ *Structures of Social Action: Studies in Conversation Analysis*, 413–429.
- Sacks, H. (1995). *Lectures on conversation: Volumes I & II* (G. Jefferson, Ed.; 1. publ. in one paperback volume). Blackwell.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4), 696. <https://doi.org/10.2307/412243>
- Sarrica, M., Brondi, S., & Fortunati, L. (2019). How many facets does a “social robot” have? A review of scientific and popular definitions online. *Information Technology & People*, 33(1), 1–21. <https://doi.org/10.1108/ITP-04-2018-0203>
- Sawyer, K. (2015). Creativity and dialogue: The improvisational nature of conversational interaction. In *The Routledge Handbook of Language and Creativity* (pp. 78–91). Routledge.
- Scheffler, J., & Pitsch, K. (2020). Pre-beginnings in Human-Robot Encounters: Dealing with time delay. *Proceedings of the 18th European Conference on Computer-Supported Cooperative Work*. https://doi.org/10.18420/ecscw2020_p02
- Schegloff, E. A. (1968). Sequencing in Conversational Openings. *American Anthropologist*, 70(6), 1075–1095. <https://doi.org/10.1525/aa.1968.70.6.02a00030>
- Schegloff, E. A. (1979a). Identification and recognition in telephone conversation openings. In G. Psathas (Ed.), *Everyday Language: Studies in Ethnomethodology* (pp. 23–78). Irvington Publishers, Inc.

- Schegloff, E. A. (1979b). The Relevance of Repair to Syntax-for-Conversation. In *Syntax and Semantics* (Vol. 12, pp. 261–286). https://doi.org/10.1163/9789004368897_012
- Schegloff, E. A. (1980). Preliminaries to Preliminaries: “Can I Ask You a Question?” *Sociological Inquiry*, 50(3–4), 104–152. <https://doi.org/10.1111/j.1475-682X.1980.tb00018.x>
- Schegloff, E. A. (1984). On some gestures’ relation to talk. In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press.
- Schegloff, E. A. (1986). The routine as achievement. *Human Studies*, 9(2), 111–151. <https://doi.org/10.1007/BF00148124>
- Schegloff, E. A. (1988). Goffman and the analysis of conversation. In *Erving Goffman: Exploring the Interaction Order*. Polity Press.
- Schegloff, E. A. (1991). Reflections on talk and social structure. In *Talk and Social Structure: Studies in Ethnomethodology and Conversation Analysis*. Polity Press.
- Schegloff, E. A. (1992a). On talk and its institutional occasions. *Talk at Work: Interaction in Institutional Settings*, 101, 101–134.
- Schegloff, E. A. (1992b). Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, 97, 1295–1345.
- Schegloff, E. A. (1992c). Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, 97, 1295–1345.
- Schegloff, E. A. (1993). Reflections on Quantification in the Study of Conversation. *Research on Language & Social Interaction*, 26(1), 99–128. https://doi.org/10.1207/s15327973rlsi2601_5
- Schegloff, E. A. (1996a). Confirming Allusions: Toward an Empirical Account of Action. *American Journal of Sociology*, 102(1), 161–216. JSTOR.
- Schegloff, E. A. (1996b). Confirming Allusions: Toward an Empirical Account of Action. *American Journal of Sociology*, 102(1), 161–216. JSTOR.
- Schegloff, E. A. (1996c). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (1st ed., pp. 52–133). Cambridge University Press. <https://doi.org/10.1017/CBO9780511620874.002>
- Schegloff, E. A. (1997). Whose Text? Whose Context? *Discourse & Society*, 8(2), 165–187. <https://doi.org/10.1177/0957926597008002002>
- Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1), 1–63. Cambridge Core. <https://doi.org/10.1017/S0047404500001019>
- Schegloff, E. A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511791208>

- Schegloff, E. A. (2010a). Commentary on Stivers and Rossano: "Mobilizing Response." *Research on Language & Social Interaction*, 43(1), 38–48. <https://doi.org/10.1080/08351810903471282>
- Schegloff, E. A. (2010b). Some Other "Uh(m)"s. *Discourse Processes*, 47(2), 130–174. <https://doi.org/10.1080/01638530903223380>
- Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 53(2), 361–382. JSTOR. <https://doi.org/10.2307/413107>
- Schegloff, E. A., & Sacks, H. (1973a). Opening up Closings. *Semiotica*, 8(4). <https://doi.org/10.1515/semi.1973.8.4.289>
- Schegloff, E. A., & Sacks, H. (1973b). Opening up Closings. *Semiotica*, 8(4). <https://doi.org/10.1515/semi.1973.8.4.289>
- Schiffrin, D. (1993). Speaking for another in sociolinguistic interviews. In D. Tannen (Ed.), *Framing in discourse* (pp. 231–263). Oxford University Press.
- Schneider, W. L. (2013). *Grundlagen der soziologischen Theorie: Band 2: Garfinkel-RC-Habermas-Luhmann*. Springer-Verlag.
- Schulz, T., Soma, R., & Holthaus, P. (2021). Movement acts in breakdown situations: How a robot's recovery procedure affects participants' opinions. *Paladyn, Journal of Behavioral Robotics*, 12(1), 336–355. <https://doi.org/10.1515/pjbr-2021-0027>
- Schutz, A. (1972). Common-Sense and Scientific Interpretation of Human Action. In M. Natanson (Ed.), *Collected Papers I: The Problem of Social Reality* (pp. 3–47). Springer Netherlands. https://doi.org/10.1007/978-94-010-2851-6_1
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press; Cambridge Core. <https://doi.org/10.1017/CBO9781139173438>
- Searle, J. R. (1975). Indirect Speech Acts. In P. Cole & J. L. Morgan (Eds.), *Speech Acts* (pp. 59–82). BRILL. https://doi.org/10.1163/9789004368811_004
- Seo, M.-S., & Koshik, I. (2010). A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics*, 42(8), 2219–2239. <https://doi.org/10.1016/j.pragma.2010.01.021>
- Seuren, L. M., Wherton, J., Greenhalgh, T., & Shaw, S. E. (2021). Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics*, 172, 63–78. <https://doi.org/10.1016/j.pragma.2020.11.005>
- Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, 23(8–9), 1099–1103. <https://doi.org/10.1016/j.neunet.2010.08.014>
- Shotter, J. (1996). 'Now I can go on:.' Wittgenstein and our embodied embeddedness in the 'Hurly-Burly' of life. *Human Studies*, 19(4), 385–407. <https://doi.org/10.1007/BF00188850>

- Sidnell, J. (2005). Talk and practical epistemology. *Talk and Practical Epistemology*, 1–272.
- Sidnell, J. (2010). The ordinary ethics of everyday talk. In *Ordinary Ethics: Anthropology, Language, and Action*. Fordham University Press.
- Sidnell, J. (2012). Basic conversation analytic methods. *The Handbook of Conversation Analysis*, 77–99.
- Sidnell, J. (2017). Action in interaction is conduct under a description. *Language in Society*, 46(3), 313–337. Cambridge Core. <https://doi.org/10.1017/S0047404517000173>
- Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language*, 67, 101178. <https://doi.org/10.1016/j.csl.2020.101178>
- Stanley, S., Smith, R. J., Ford, E., & Jones, J. (2020). Making something out of nothing: Breaching everyday life by standing still in a public place. *The Sociological Review*, 68(6), 1250–1272. <https://doi.org/10.1177/0038026120940616>
- Sterponi, L., & Fasulo, A. (2010). “How to Go On”: Intersubjectivity and Progressivity in the Communication of a Child with Autism. *Ethos*, 38, 116–142. <https://doi.org/10.1111/j.1548-1352.2009.01084.x>
- Stevanovic, M. (2023). Accountability and interactional inequality: The management of problems of interaction as a matter of cultural ideals and ideologies. *Frontiers in Sociology*, 8. <https://doi.org/10.3389/fsoc.2023.1204086>
- Stichter, M. (2016). Practical Skills and Practical Wisdom in Virtue. *Australasian Journal of Philosophy*, 94(3), 435–448. <https://doi.org/10.1080/00048402.2015.1074257>
- Stivers, T., Mondada, L., & Steensig, J. (2011). Knowledge, morality and affiliation in social interaction. In J. Steensig, L. Mondada, & T. Stivers (Eds.), *The Morality of Knowledge in Conversation* (pp. 3–24). Cambridge University Press; Cambridge Core. <https://doi.org/10.1017/CBO9780511921674.002>
- Stivers, T., & Robinson, J. D. (2006). A Preference for Progressivity in Interaction. *Language in Society*, 35(3), 367–392. JSTOR.
- Stivers, T., & Rossano, F. (2010a). A Scalar View of Response Relevance. *Research on Language and Social Interaction*, 43(1), 49–56. <https://doi.org/10.1080/08351810903471381>
- Stivers, T., & Rossano, F. (2010b). Mobilizing Response. *Research on Language & Social Interaction*, 43(1), 3–31. <https://doi.org/10.1080/08351810903471258>
- Stivers, T., Rossi, G., & Chalfoun, A. (2023). Ambiguities in Action Ascription. *Social Forces*, 101(3), 1552–1579. <https://doi.org/10.1093/sf/soac021>
- Stokoe, E. (2013). The (in) authenticity of simulated talk: Comparing role-played and actual interaction and the implications for communication training. *Research on Language & Social Interaction*, 46(2), 165–185.

- Stokoe, E., & Attenborough, F. (2014). Ethnomethodological methods for identity and culture: Conversation analysis and membership categorization. In *Researching Identity and Interculturality*. Routledge.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9781315816883-7/ethnomethodological-methods-identity-culture-elizabeth-stokoe-frederick-attenborough>
- Stokoe, E., & Smithson, J. (2001). Making Gender Relevant: Conversation Analysis and Gender Categories in Interaction. *Discourse & Society - DISCOURSE SOCIETY*, 12, 217–244. <https://doi.org/10.1177/0957926501012002005>
- Stommel, W., de Rijk, L., & Boumans, R. (2022). “Pepper, what do you mean?” Miscommunication and repair in robot-led survey interaction. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 385–392. <https://doi.org/10.1109/RO-MAN53752.2022.9900528>
- Straub, I. (2016). ‘It looks like a human!’ The interrelation of social presence, interaction and agency ascription: A case study about the effects of an android robot on social agency ascription. *AI & SOCIETY*, 31(4), 553–571. <https://doi.org/10.1007/s00146-015-0632-5>
- Straub, I., Nishio, S., & Ishiguro, H. (2012). From an object to a subject—Transitions of an android robot into a social being. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 821–826. <https://doi.org/10.1109/ROMAN.2012.6343853>
- Stukenbrock, A., & Dao, A. N. (2019). Joint Attention in Passing: What Dual Mobile Eye Tracking Reveals About Gaze in Coordinating Embodied Activities at a Market. In E. Reber & C. Gerhardt (Eds.), *Embodied Activities in Face-to-face and Mediated Settings: Social Encounters in Time and Space* (pp. 177–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-97325-8_6
- Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. (pp. xii, 203). Cambridge University Press.
- Sudnow, D. (1993). *Ways of the hand: The organization of improvised conduct*. MIT Press.
- Svennevig, J. (2004). Other-repetition as display of hearing, understanding and emotional stance. *Discourse Studies*, 6(4), 489–516. <https://doi.org/10.1177/1461445604046591>
- Tannen, D. (2004). Talking the Dog: Framing Pets as Interactional Resources in Family Discourse. *Research on Language and Social Interaction*, 37(4), 399–420. https://doi.org/10.1207/s15327973rlsi3704_1
- Tatarian, K., Stower, R., Rudaz, D., Chamoux, M., Kappas, A., & Chetouani, M. (2021). How does Modality Matter? Investigating the Synthesis and Effects of Multi-modal Robot Behavior on Social Intelligence. *International Journal of Social Robotics*. <https://doi.org/10.1007/s12369-021-00839-w>
- Tavory, I. (2022). Occam’s Razor and the Challenges of Generalization in Ethnomethodology. In D. W. Maynard & J. Heritage (Eds.), *The Ethnomethodology Program: Legacies and*

Prospects (p. 0). Oxford University Press.
<https://doi.org/10.1093/oso/9780190854409.003.0016>

- Tedeschi, E. (2016). Animals, Humans and Sociability. *Italian Sociological Review*, 6(2), 151.
<https://doi.org/10.13136/isr.v6i2.130>
- Ten Have, P. (1990). Methodological Issues in Conversation Analysis. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 27(1), 23–51.
<https://doi.org/10.1177/075910639002700102>
- Ten Have, P. (2007). *Doing Conversation Analysis*. SAGE Publications, Ltd.
<https://doi.org/10.4135/9781849208895>
- Ten Have, P. (2016). Ethnomethodology. In K. B. Jensen, E. W. Rothenbuhler, J. D. Pooley, & R. T. Craig (Eds.), *The International Encyclopedia of Communication Theory and Philosophy* (1st ed., pp. 1–12). Wiley.
<https://doi.org/10.1002/9781118766804.wbiect010>
- Tennent, H., Moore, D., Jung, M., & Ju, W. (2017). Good vibrations: How consequential sounds affect perception of robotic arms. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 928–935.
<https://doi.org/10.1109/ROMAN.2017.8172414>
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Thellman, S., Lundberg, J., Arvola, M., & Ziemke, T. (2017). What Is It Like to Be a Bot?: Toward More Immediate Wizard-of-Oz Control in Social Human-Robot Interaction. *Proceedings of the 5th International Conference on Human Agent Interaction*, 435–438. <https://doi.org/10.1145/3125739.3132580>
- Theureau, J. (2004). L'hypothèse de la cognition (ou action) située et la tradition d'analyse du travail de l'ergonomie de langue française. *Activites*, 01(2).
<https://doi.org/10.4000/activites.1219>
- Tisserand, L., Armetta, F., Baldauf-Quilliatre, H., Bouquin, A., Hassas, S., & Lefort, M. (2023). *Sequential annotations for naturally-occurring HRI: First insights*.
<https://doi.org/10.48550/ARXIV.2308.15097>
- Tisserand, L., & Baldauf-Quilliatre, H. (2023, June 26). *Beyond the speech recognition, (doing) being perceived by a robot*. Paper presented at the International Conference on Conversation Analysis, Brisbane, Australia.
- Tuncer, S., Gillet, S., & Leite, I. (2022a). Robot-Mediated Inclusive Processes in Groups of Children: From Gaze Aversion to Mutual Smiling Gaze. *Frontiers in Robotics and AI*, 9. <https://doi.org/10.3389/frobt.2022.729146>
- Tuncer, S., Gillet, S., & Leite, I. (2022b). Robot-Mediated Inclusive Processes in Groups of Children: From Gaze Aversion to Mutual Smiling Gaze. *Frontiers in Robotics and AI*, 9, 729146. <https://doi.org/10.3389/frobt.2022.729146>

- Tuncer, S., Licoppe, C., Luff, P., & Heath, C. (2023). Recipient design in human–robot interaction: The emergent assessment of a robot’s competence. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-022-01608-7>
- Turner, R. (2017). Words, utterances, and activities. In *Everyday Life* (pp. 169–187). Routledge.
- Turowetz, J., Hollander, M. M., & Maynard, D. W. (2016). Ethnomethodology and Social Phenomenology. In *Handbooks of Sociology and Social Research*. Springer. https://link.springer.com/chapter/10.1007/978-3-319-32250-6_19
- Turowetz, J. J., & Maynard, D. W. (2010). Morality in the Social Interactional and Discursive World of Everyday Life. In S. Hitlin & S. Vaisey (Eds.), *Handbook of the Sociology of Morality* (pp. 503–526). Springer New York. https://doi.org/10.1007/978-1-4419-6896-8_27
- Uexküll, T. von. (1982). Introduction: Meaning and science in Jakob von Uexküll’s concept of biology. *Semiotica*, 42(1), 1–24.
- Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, 31(3), 361–380. <https://doi.org/10.1177/0162243905285847>
- Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12), 1743–1759. <https://doi.org/10.1016/j.imavis.2008.11.007>
- vom Lehn, D. (2019). From Garfinkel’s ‘Experiments in Miniature’ to the Ethnomethodological Analysis of Interaction. *Human Studies*, 42(2), 305–326. <https://doi.org/10.1007/s10746-019-09496-5>
- Von Uexküll, T. (1988). Jakob von Uexküll’s Umwelt Theory. *The Semiotic Web*, 129–158.
- Walker, G. (2003). Doing a rushthrough’: A phonetic resource for holding the turn in everyday conversation. *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona/Causal Productions*, 50.
- Weatherall, A., & Edmonds, D. M. (2018). Speakers formulating their talk as interruptive. *Journal of Pragmatics*, 123, 11–23. <https://doi.org/10.1016/j.pragma.2017.11.008>
- Weiss, A., Igelsböck, J., Tscheligi, M., Bauer, A., Kühnlenz, K., Wollherr, D., & Buss, M. (2010). Robots asking for directions: The willingness of passers-by to support robots. *Proceeding of the 5th ACM/IEEE International Conference on Human-Robot Interaction - HRI ’10*, 23. <https://doi.org/10.1145/1734454.1734468>
- West, C., & Fenstermaker, S. (1993). Power, inequality and the accomplishment of gender: An ethnomethodological view. In *Theory on Gender/Feminism on Theory*. Aldine.
- West, C., & Zimmerman, D. H. (1987). Doing Gender. *Gender and Society*, 1(2), 125–151. JSTOR.

- West, C., & Zimmerman, D. H. (2009). Accounting for Doing Gender. *Gender & Society*, 23(1), 112–122. <https://doi.org/10.1177/0891243208326529>
- Wieder, D. L., & Pratt, S. (1990). On being a recognizable Indian among Indians. In *Cultural communication and intercultural contact*. (pp. 45–64). Lawrence Erlbaum Associates, Inc.
- Williams, J. (2001). Phenomenology in Sociology. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 11361–11363). Elsevier. <https://doi.org/10.1016/B0-08-043076-7/01939-2>
- Wittgenstein, L. (1953). *Philosophical Investigations* (Vol. 17, Issue 69, p. 362). Wiley-Blackwell.
- Wittgenstein, L. (1969). *On Certainty* (ed. Anscombe and von Wright) (Issue 323, pp. 453–457). Harper Torchbooks.
- Wooffitt, R. (1992). *Telling tales of the unexpected: The organization of factual discourse*. (pp. xii, 217). Harvester Wheatsheaf.
- Yamauchi, Y., & Hiramoto, T. (2016). Reflexivity of Routines: An Ethnomethodological Investigation of Initial Service Encounters at Sushi Bars in Tokyo. *Organization Studies*, 37(10), 1473–1499. <https://doi.org/10.1177/0170840616634125>
- Yanco, H. A., & Drury, J. (2004). Classifying human-robot interaction: An updated taxonomy. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 3, 2841–2846. <https://doi.org/10.1109/ICSMC.2004.1400763>
- Yang, F., Gao, Y., Ma, R., Zojaji, S., Castellano, G., & Peters, C. (2021). A dataset of human and robot approach behaviors into small free-standing conversational groups. *PLOS ONE*, 16(2), e0247364. <https://doi.org/10.1371/journal.pone.0247364>
- Yao, L., Pan, Y., & Jiang, D. (2011a). Effects of Automated Transcription Delay on Non-Native Speakers' Comprehension in Real-Time Computer-Mediated Communication. *Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part I*, 207–214.
- Yao, L., Pan, Y., & Jiang, D. (2011b). Effects of Automated Transcription Delay on Non-native Speakers' Comprehension in Real-Time Computer-Mediated Communication. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2011* (pp. 207–214). Springer Berlin Heidelberg.
- Zimmerman, D. H., & Pollner, M. (2013). The Everyday World as a Phenomenon'. *People and Information: Pergamon General Psychology Series*, 6, 33.

10. APPENDIX

10.1. List of figures and tables

10.1.1. List of figures

Figure 1.1. Pepper's dimensions	11
Figure 1.2. Simplified layout diagram of Pepper's sensors	11
Figure 1.3. Pepper's degrees of freedom	12
Figure 2.1. Images 1 & 2 – HUM asks ROB if it knows how to dance (img. 1), then laughs after ROB starts dancing (img. 2).	20
Figure 3.1. Range of applicability of an interactionist definition of social agency to observational data.....	45
Figure 3.2. Image 1 – KAR slaps ROB.....	50
Figure 3.3. Image 2 – ROB gazes on the left, KAR seizes ROB's head with both hands.....	50
Figure 3.4. Image 3 – KAR turns ROB's head towards him.....	51
Figure 4.1. "Engagement Intention Flow Diagram". This diagram summarizes the criteria for the identification of humans "Seeking Engagement", "Interested" and "Not Interested" (Image courtesy of Miriam Bilac, used with permission).	63
Figure 4.2. The serving robot Plato of Aldebaran in its current shape (2023)	67
Figure 4.3. Transcription of a customer's conduct when the waiter was not responding to her requests (showcased during a PowerPoint presentation).....	68
Figure 4.4. Step-by-step representation of typical ways of "departing from a table" for waiters.	70
Figure 4.5. Examples of the mode of representation used on Figma for designing the robot's behavior in human-robot interactions. Used with permission.....	73
Figure 4.6. Zoom in on the expected behaviors from the robot in each "step"	74
Figure 4.7. The sound "uhm" transcribed in "L&H" phonetic alphabet. The "+", "%" and "~" symbols conflict with their use within the QiChat syntax.....	78
Figure 4.8. Example of a potential continuation turn in QiChat, provided in former versions of Pepper and Nao's documentation	80
Figure 4.9. A two-second (2000ms) pause during a speaking turn of the robot, composed in QiChat language. Although it is scripted as a pause (i.e., as a silence occurring when the turn is still ongoing in the script), this silence may be treated as a "gap" by human listeners (i.e., as a slot to produce a response).	82
Figure 4.10. Structure of a dialog tree described in the official documentation available online.	83

Figure 4.11. Screen from a “rule” taken from a conversational application made for Pepper that uses the QiChat language. Upon hearing various requests about the meaning of its name, Pepper explains what it means and why it was given this name.....	84
Figure 5.1. Distribution of all greetings produced by participants between “activation steps”, including multiple greetings by the same participants. Total first greetings = 62, total greetings overall = 85.....	99
Figure 5.2. Image 1.1 – Participant tilts her head after the robot does not reply to her greeting.....	100
Figure 5.3. Image 2.1 – Participant adjusts his clothes after positioning himself in front of the robot	102
Figure 5.4. Images 2.2 and 2.3 – Participant steps towards the robot after mutual gaze is established, then, after a few seconds of silence, steps back.....	103
Figure 5.5. Images 2.4 to 2.7 – Participant starts to extend his hand towards the robot, then modifies the trajectory of his hand to align with the robot’s waving gesture	104
Figure 5.6. Images 3.1 and 3.2 – The robot gazes at the participant, she stops her swinging movement and establishes mutual gaze	105
Figure 5.7. Images 3.3 and 3.4 – Participant’s widened smile when uttering her second “hello”	106
Figure 5.8. Images 4.1 to 4.4 – The participant catches up with the robot’s waving gesture, then retracts her gesture immediately after the robot retracts its arm	108
Figure 5.9. Image 5.1 – Participant approaches the robot while producing “self-talk”	110
Figure 5.10. Image 5.2 – Participant laughs after the robot achieves mutual gaze	110
Figure 5.11. Image 5.3 – Participant’s second laugh during the waving gesture of the robot	111
Figure 6.1. Image 1.1 – ISA greets ROB.....	122
Figure 6.2. Image 1.2 – ISA gazes at OLI after ROB does not respond to her greeting.....	122
Figure 6.3. Image 1.3 – OLI lightly pushes ISA towards ROB	123
Figure 6.4. Image 2.1 & 2.2 – EMI gazes at ROB (image 2.1), then turns towards SAM and MIR after ROB utters “pardon?” (image 2.2)	125
Figure 6.5. Image 3.1 – ROB keeps dancing after TOM’s request. SYL and CLA are positioned behind TOM as bystanders.	128
Figure 6.6. Image 3.2 – TOM, SYL and CLA laugh after ROB’s utterance	129
Figure 6.7. Image 3.3 – After SYL and CLA describe ROB’s conduct as “trolling”, TOM torques towards them and produces an assessment about ROB’s conduct	130
Figure 6.8. Image 4.1 – WIL pushes EMM while he requests her to stop saying “I love you” to ROB.....	133
Figure 6.9. Image 4.2 – WIL requests ROB to tell EMM that it does not love her	133

Figure 6.10. Image 4.3 – WIL informs EMM that she must leave and that ROB said “go away”	134
Figure 7.1. Default speechbar setting on a recent Pepper (that is, running its latest operating system, NAOqi 2.9). The robot displays “hello” after a human pronounced this word and stopped speaking for more than 200ms.	149
Figure 7.2. Setup for the Pepper robot (loading its chatbot) at the start of the “Robots” exhibition, including an RGPD-compliant sign (left), a red stripe on the ground, and a camera.....	154
Figure 7.3. Screenshot of the content of the chatbot. On the left, some of the topics about which the robot could produce answers. On the right, some of the concepts it could recognize in humans’ speech.....	156
Figure 7.4. The robot received a strip of sound but could not identify any word within it....	158
Figure 7.5. Pepper after hearing “hello” in the No Transcript condition (left) and in the ASR Transcript condition (right)	161
Figure 7.6. Speaking turns produced by a participant coded as Times of Interest (TOI) on Tobii Pro Lab.	161
Figure 7.7. Areas of Interest (AOI) set on a 2D picture of the robot on Tobii Pro Lab	162
Figure 7.8. Example of timecoded utterances and post-utterance phases for a human participant	163
Figure 7.9. Mean total fixation time on areas of interest per condition (post-utterance)	164
Figure 7.10. Heatmap of participants' fixations after they finished their speaking turn in the ASR Transcript condition (N=22), left, and in the No Transcript condition (N=22), right.....	165
Figure 7.11. Post-utterance fixations over time per areas of interest	166
Figure 7.12. Image 1.1 – PA1 turns away from ROB after reading the ASR transcript	168
Figure 7.13. Image 1.2 & 1.3 – PA2 points at the robot’s tablet.....	169
Figure 7.14. Image 2.1 – PA1 raises a finger	171
Figure 7.15. Image 2.2 & 2.3 – PA1 extends and retracts his hand.....	172
Figure 7.16. Image 3.1 – PA1 asks a question to ROB	175
Figure 7.17. Image 3.2 – PA1 steps and turns towards PA2	176
Figure 7.18. Image 4.1 – PA1 points at ROB’s tablet	179
Figure 7.19. Image 4.2 – PA1 waves a finger in front of ROB’s tablet	179
Figure 7.20. Image 5.1 – PA1 extends his closed fist in front of ROB.....	181
Figure 7.21. Image 5.2 – PA1 opens his hand, palm facing the ceiling.....	182
Figure 7.22. Image 6.1 – Group of visitors looks at the ASR transcript.....	185
Figure 7.23. Image 6.2 – PA3 waves a finger in front of ROB’s head	186
Figure 7.24. Image 7.1 – PA1 extends her hand towards ROB’s right hand	188

Figure 7.25. Images 7.2 & 7.3 – PA1 touches ROB’s left hand while PA2 states that ROB understood “close the hand”..... 189

Figure 7.26. Image 8.1 – PA1 points at ROB’s tablet while asking PA2 for confirmation about what it previously displayed..... 192

10.1.2. List of tables

Table 5.1. Summary of First Greeting Occurrences 99

10.2. Ethical approval and consent forms for the INSEAD experiment



Étude SU – Interaction sociale avec un robot

Demande d'autorisation d'utilisation de l'image d'une personne

Je soussigné(e) :

Demeurant :

Autorise le centre de recherche INSEAD-Sorbonne Université et les chercheurs : Mohamed Chetouani, Karen Tatarian et Rebecca Stower, à utiliser les photos et enregistrements vidéos sur lesquels j'apparais et qui ont été réalisés dans le cadre de la passation de l'étude « Interaction sociale avec un robot » ce jour. Ces enregistrements pourront être utilisés dans le cadre de l'étude et à des fins d'analyses pour leurs travaux de recherche.

En cochant la case ci-contre, j'autorise également les chercheurs à utiliser ces images dans le cadre de présentations en conférence ou d'illustrations dans des publications, à condition que mon anonymat soit conservé.

Fait à PARIS , le

Signature



The Business School
for the World®

**Centre Multidisciplinaire des
Sciences Comportementales
Sorbonne Université-INSEAD**



FEUILLE DE CONSENTEMENT

Bienvenue au Centre Multidisciplinaire des Sciences Comportementales Sorbonne Université-INSEAD. Les chercheurs de l'étude à laquelle vous allez participer sont: Karen Tatarian (karen.tatarian@softbankrobotics.com) et Rebecca Stower (r.stower@jacobs-university.de)

Dans cette étude, vous devrez interagir avec un robot social qui va se comporter comme un agent de voyage. Après, vous serez demandé de remplir les questionnaires sur votre impression du robot. La durée estimée de l'étude est de 30 minutes.

Dans cette étude, vous allez interagir avec un robot social qui va se comporter comme un agent de voyage. A la fin, on vous demandera de remplir un questionnaire pour mieux comprendre votre avis sur le robot. La durée estimée de l'étude est 30 minutes.

Nos études sont à visées académiques, et les résultats seront accessibles dans des publications scientifiques. Nous ne réalisons pas d'études pour le compte d'entreprises privées. Il n'existe aucun risque lié à cette étude autre que ceux de la vie de tous les jours.

Si vous terminez cette étude, vous recevrez €6. Par ailleurs, vous pourrez recevoir des informations sur les conclusions de l'étude si vous le souhaitez, ainsi que des références concernant le type de recherche auquel vous avez participé. Cependant, parce que vos réponses sont anonymes, nous ne pourrions vous renseigner que sur les résultats agrégés de l'étude, et non sur votre performance ou les performances de n'importe qui d'autre ayant participé.

Nous aimerions également votre permission de procéder à un enregistrement vidéo/audio de l'étude. Si nous avons besoin d'utiliser des citations ou des éléments susceptibles de vous identifier, nous vous demanderons alors votre permission avant de le faire. Le reste des données saisies sera complètement confidentiel, et ne sera jamais diffusé d'une manière qui permette de vous identifier.

Les données concernant cette étude seront conservées sous clé ou protégées par un mot de passe, et seront détruites dès lors qu'elles ne seront plus utilisées.

Votre participation à l'étude doit être entièrement volontaire, et vous avez la possibilité de vous retirer de l'étude à tout moment sans aucune pénalité.

Je déclare être majeur(e), et ayant lu et parfaitement compris les paragraphes ci-dessus, accepte de mon plein gré de participer à cette étude.

DATE :

NOM :

PRÉNOM :

SIGNATURE :

OUTCOME - INSEAD INSTITUTIONAL REVIEW BOARD

Your study "**Interacting with a Social Robot**" reference number **201955** has been approved. If this protocol is used in conjunction with any other human use, it must be re-reviewed. The IRB requests prompt notification of any complications or incidents of noncompliance, which may occur during any human use procedure. In case any new data is collected, all data including the consent forms, must be retained for a minimum of three years past the completion of this research. Additional requirement may be imposed by your funding agency, your department, or other entities.

Please use the reference number stated above for any further correspondence, amendment relating to this study.

Best regards,

The INSEAD Institutional Review Board

10.3. Ethical approval and consent forms for the Cité des Sciences experiment



Avis donné par les membres du Comité d’Ethique pour la Recherche (CER) de l’Université Paris-Saclay

Numéro de dossier: 412
Titre de l’étude : Robot d’information en situation naturelle
Date de l’étude : réception de l’avis ..
Demandeur de l’étude : LICOPPE Christian & RUDAZ Damien, Prof. & Doctorant resp. à Télécom Paris et IPP, Lab I3
Date de réception de la demande : 31 mars 2022
Lieu(x) de l’étude : Cité des Sciences et de l’Industrie, 30 Av. Corentin Cariou, 75019 Paris
Date d’émission de l’avis : le 4 juillet 2022
Version d’avis : 2

Sur proposition des rapporteurs le comité adopte l’avis suivant :

Avis 1. Favorable.

La référence associée à cet avis est la suivante : CER-Paris-Saclay-2022-040

FEUILLE DE CONSENTEMENT (MINEUR)

En interagissant avec le robot « Pepper » dans le cadre de cette expérimentation "Interaction avec un Robot", votre enfant participera à une étude scientifique visant à personnaliser les comportements du robot en réaction à ses interlocuteurs.

Notre étude est à visée académique, et les résultats seront dévoilés dans des publications scientifiques. Il n'existe aucun risque lié à cette étude autre que ceux de la vie de tous les jours.

Vous pourrez recevoir des informations sur les conclusions de l'étude si vous le souhaitez, ainsi que des références concernant le type de recherche auquel votre enfant a participé. Cependant, parce que les réponses de votre enfant sont anonymes, nous ne pourrions vous renseigner que sur les résultats agrégés de l'étude, et non sur sa performance ou les performances de n'importe qui d'autre ayant participé.

Nous aimerions également votre permission de procéder à un enregistrement vidéo et audio de l'interaction de votre enfant avec notre robot. Ces enregistrements pourront être utilisés dans le cadre de l'étude et à des fins d'analyse pour des travaux de recherche. Le reste des données saisies sera dans tous les cas complètement confidentiel, et ne sera jamais diffusé d'une manière qui permette d'identifier votre enfant.

Les données concernant cette étude, incluant les enregistrements, seront conservées sous clé ou protégées par un mot de passe. Elles seront détruites deux ans après leur captation.

La participation de votre enfant à l'étude doit être entièrement volontaire, et vous avez la possibilité de vous retirer de l'étude à tout moment sans aucune pénalité.

Pour toute information sur ce dispositif, ou si vous souhaitez exercer vos droits d'accès, de rectification et d'opposition aux images qui concernent votre enfant, contactez damien.rudaz@telecom-paris.fr

Ayant lu et parfaitement compris les paragraphes ci-dessus, j'accepte librement et volontairement que mon enfant, dont je suis le représentant légal, participe à l'expérimentation "Interaction avec un robot"

Votre nom et prénom

Votre adresse email

Nom et prénom de l'enfant représenté

Valider

FEUILLE DE CONSENTEMENT

En interagissant avec le robot « Pepper » à proximité de vous, vous participez à une étude scientifique visant à personnaliser les comportements du robot en réaction à ses interlocuteurs.

Notre étude est à visée académique, et les résultats seront dévoilés dans des publications scientifiques. Il n'existe aucun risque lié à cette étude autre que ceux de la vie de tous les jours.

Vous pourrez recevoir des informations sur les conclusions de l'étude si vous le souhaitez, ainsi que des références concernant le type de recherche auquel vous avez participé. Cependant, parce que vos réponses sont anonymes, nous ne pourrions vous renseigner que sur les résultats agrégés de l'étude, et non sur votre performance ou les performances de n'importe qui d'autre ayant participé.

Nous aimerions également votre permission de procéder à un enregistrement vidéo et audio de votre interaction avec notre robot. Ces enregistrements pourront être utilisés dans le cadre de l'étude et à des fins d'analyse pour des travaux de recherche. Le reste des données saisies sera dans tous les cas complètement confidentiel, et ne sera jamais diffusé d'une manière qui permette de vous identifier.

Les données concernant cette étude, incluant les enregistrements, seront conservées sous clé ou protégées par un mot de passe. Elles seront détruites deux ans après leur captation.

Votre participation à l'étude doit être entièrement volontaire, et vous avez la possibilité de vous retirer de l'étude à tout moment sans aucune pénalité.

Pour toute information sur ce dispositif, ou si vous souhaitez exercer vos droits d'accès, de rectification et d'opposition aux images qui vous concernent, contactez damien.rudaz@telecom-paris.fr

Je déclare avoir l'âge légal de consentement, et ayant lu et parfaitement compris les paragraphes ci-dessus, accepte de mon plein gré de participer à cette étude.

* Identité

Nom

Prénom

Adresse email

Valider

DEMANDE D'AUTORISATION D'UTILISATION DE L'IMAGE D'UNE PERSONNE (MINEUR)

J'autorise les chercheurs Christian Licoppe, Marc Hulcelle et Damien Rudaz à utiliser les photos et enregistrements vidéo et audio sur lesquels mon enfant apparaît et qui ont été réalisés dans le cadre de la passation de l'étude "Interaction avec un robot", ce jour.

Ces enregistrements pourront être utilisés dans le cadre de l'étude et à des fins d'analyse dans leur travail de recherche. Le stockage de ces enregistrements est réalisé par Télécom Paris. Ils ne seront pas transférés à des tiers et seront effacés dans un délai maximum de 2 ans après leur captation.

Pour toute information sur ce dispositif, ou si je souhaite exercer mes droits d'accès, de rectification et d'opposition aux images qui concernent mon enfant, je peux contacter damien.rudaz@telecom-paris.fr

- (Optionnel)** En cochant la case ci-contre, j'autorise également les chercheurs à utiliser ces images dans le cadre de présentations en conférences scientifiques ou d'illustrations dans des publications, à condition que l'anonymat de mon enfant soit conservé.

* Identité du représentant légal de l'enfant

Nom

Prénom

Précédent

Terminé

DEMANDE D'AUTORISATION D'UTILISATION DE L'IMAGE D'UNE PERSONNE

J'autorise les chercheurs Christian Licoppe, Marc Hulcelle et Damien Rudaz à utiliser les photos et enregistrements vidéo et audio sur lesquels j'apparaît et qui ont été réalisés dans le cadre de la passation de l'étude "Interaction sociale avec un robot", ce jour.

Ces enregistrements pourront être utilisés dans le cadre de l'étude et à des fins d'analyse dans leur travail de recherche. Le stockage de ces enregistrements est réalisé par Télécom Paris. Ils ne seront pas transférés à des tiers et seront effacés dans un délai maximum de 2 ans après leur captation.

Pour toute information sur ce dispositif, ou si je souhaite exercer mes droits d'accès, de rectification et d'opposition aux images qui me concernent, je peux contacter damien.rudaz@telecom-paris.fr

Je déclare avoir l'âge légal de consentement, et ayant lu et parfaitement compris les paragraphes ci-dessus, accepte de mon plein gré de participer à cette étude.

(Optionnel) En cochant la case ci-contre, j'autorise également les chercheurs à utiliser ces images dans le cadre de présentations en conférences scientifiques ou d'illustrations dans des publications, à condition que mon anonymat soit conservé.

* Identité

Nom

Prénom



Cette zone est filmée



- Vos interactions avec le robot Pepper, matérialisées par des captations de votre image, de votre voix et du contenu de votre discours, sont enregistrées dans le but d'analyser les interactions humain-robot. Ceci à des fins de recherche scientifique, dans le cadre d'une étude portant sur la personnalisation de l'interaction avec un robot.
- Le stockage de ces données est réalisé par Telecom Paris. Elles ne seront pas transférées à des tiers et seront effacées dans un délai maximum de 2 ans après leur captation.
- Toute donnée personnelle captée sera anonymisée (visage, voix, nom, etc.).
- Pour toute information sur ce dispositif, ou si vous souhaitez exercer vos droits d'accès, de rectification et d'opposition aux images qui vous concernent, contactez Damien.Rudaz@telecom-paris.fr
- Si vous entrez dans cette zone, délimitée par les bandes rouges au sol, vous consentez à ce que vos données personnelles soient traitées pour la finalité décrite ci-dessus et en conformité avec le RGPD. Vous consentez de plus à ce que votre image soit captée pour les besoins de la finalité.



This area is being recorded



- Your interactions with the Pepper robot, materialized by captures of your image, of your voice and of the content of your speech, are recorded in order to analyse human-robot interactions. This is done for scientific purposes, as part of a study focused on the personalization of interactions with a robot.
- The storage of these data is carried out by Telecom Paris. These data will not be transferred to third parties and will be deleted after a maximum delay of two years after their capture.
- Any personal data collected will be anonymized (face, voice, name, etc.)
- For more information on this setup, or if you wish to exercise your right to access, modify or delete the images which concern you, contact Damien.Rudaz@telecom-paris.fr
- If you enter in this zone, delimited by the red strip on the ground, you consent to have your personal data processed for the purpose described above and in conformity with GDPR. Additionally, you consent to having your image captured as required for the purpose described above.

Titre : Comment les robots deviennent-ils sociaux ? Une spécification empirique de la (non-) émergence des robots comme agents sociaux

Mots clés : Ouvertures, Robots sociaux, Analyse Conversationnelle, Agent Social, HRI, EMCA

Résumé : Plutôt que d'aborder les interactions humain-robot (HRI) comme des boîtes noires, ce travail étudie les subtiles pratiques micro-interactionnelles à travers lesquelles un robot émerge comme "agent social" dans différents environnements. À partir de la perspective micro-analytique de l'Ethnométhodologie et de l'Analyse Conversationnelle (EMCA), nous examinons plusieurs larges corpus d'interactions impliquant le robot humanoïde Pepper. Cette exploration nous permet d'étendre la liste des processus interactionnels documentés prenant place durant la rencontre entre un humain et un robot (indexés à des environnements, des contextes séquentiels et des configurations spatiales spécifiques) au travers desquels un robot est traité, localement et momentanément, comme un "agent" dans une interaction sociale. Étudier la production du statut du robot comme un accomplissement

pratique nous mène à une respcification du travail interactionnel habituellement obscurci par l'usage commun du terme de robot "social".

Cependant, en lieu et place d'une perspective quiétiste visant exclusivement à clarifier des incohérences conceptuelles, notre approche répond à des problématiques de design, d'ergonomie et d'expérience utilisateur (UX) vis à vis des robots "sociaux". En somme, en faisant jour sur les pratiques localement organisées prenant place dans des contextes impliquant des robots, nous tentons de produire un autre niveau d'explication quant à "ce qui a fonctionné" et "ce qui n'a pas fonctionné" dans une interaction avec un robot : des explications basées sur les propriétés de l'interaction rendues pertinentes par les participants eux-mêmes lorsqu'ils sont immergés dans l'urgence pratique de ces interactions "humain-robot".

Title: How Do Robots Become "Social Robots"? An Empirical Specification of the (Non-) Emergence of Robots as Social Agents

Keywords: Openings, Social Robots, Conversation Analysis, Social Agent, HRI, EMCA

Abstract: As opposed to viewing the inner workings of human-robot interactions (HRI) as black boxes, this work investigates the finely tuned micro-interactional practices through which a robot emerges as a "social agent" in different settings. Using the micro-analytic approach of Ethnomethodological Conversation Analysis (EMCA), it examines several large corpora of encounters between humans and the humanoid robot Pepper. Their exploration allows us to broaden the list of documented interactional processes occurring during human-robot encounters (indexed to specific settings, sequential contexts, spatial configurations, etc.) by which a robot can be said to be, momentarily and locally, treated as an "agent" in a social interaction. Attending to the moment-by-moment production of the robot's status as a practical accomplishment

leads our inquiry to a respecification of the interactional work commonly glossed by the lay use of the term "social robot".

However, rather than merely a quietist attempt at clarifying conceptual mix-ups, our approach responds to design, ergonomic, and user experience (UX) concerns regarding "social" robots. That is, by attending to the locally organized practices taking place in human-robot encounters, we attempt to provide a different type of explanation as to "what went wrong" or "what went right" in an interaction with a robot: explanations based on the features made relevant by the participants themselves as they are practically immersed within the urgency of these ongoing human-robot interactions.