frontiers | Frontiers in Robotics and AI

TYPE Original Research
PUBLISHED 26 September 2025
DOI 10.3389/frobt.2025.1597276

OPEN ACCESS

EDITED BY
Manuel Giuliani,
Kempten University of Applied Sciences, Germany

REVIEWED BY
Frank Foerster,
University of Hertfordshire, United Kingdom
Haolin Fei,
Lancaster University, United Kingdom

*CORRESPONDENCE
Damien Rudaz,
daru@hum.ku.dk

RECEIVED 20 March 2025
ACCEPTED 11 July 2025
PUBLISHED 26 September 2025

CITATION
Rudaz D and Licoppe C (2025) Should robots display what they hear? Mishearing as a practical accomplishment.
Front. Robot. AI 12:1597276.
doi: 10.3389/frobt.2025.1597276

Should robots display what they hear? Mishearing as a practical accomplishment

Damien Rudaz[1]* and Christian Licoppe[2]

[1]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark, [2]Department of Economics and Social Sciences, Telecom Paris and Institut Polytechnique de Paris, Palaiseau, France

As a contribution to research on transparency and failures in human–robot interaction (HRI), our study investigates whether the informational ecology configured by publicly displaying a robot's automatic speech recognition (ASR) results is consequential in how miscommunications emerge and are dealt with. After a preliminary quantitative analysis of our participants' gaze behavior during an experiment where they interacted with a conversational robot, we rely on a micro-analytic approach to detail how the interpretation of this robot's conduct as inadequate was configured by what it displayed as having "heard" on its tablet. We investigate cases where an utterance or gesture by the robot was treated by participants as sequentially relevant only as long as they had not read the automatic speech recognition transcript but re-evaluated it as troublesome once they had read it. In doing so, we contribute to HRI by showing that systematically displaying an ASR transcript can play a crucial role in participants' interpretation of a co-constructed action (such as shaking hands with a robot) as having "failed". We demonstrate that "mistakes" and "errors" can be approached as practical accomplishments that emerge as such over the course of interaction rather than as social or technical phenomena pre-categorized by the researcher in reference to criteria exogenous to the activity being analyzed. In the end, while narrowing down on two video fragments, we find that this peculiar informational ecology did not merely impact how the robot was responded to. Instead, it modified the very definition of "mutual understanding" that was enacted and oriented to as relevant by the human participants in these fragments. Besides social robots, we caution that systematically providing such transcripts is a design decision not to be taken lightly; depending on the setting, it may have unintended consequences on interactions between humans and any form of conversational interface.

KEYWORDS

automatic speech recognition, errors and mistakes, transparency, action ascription, conversation analysis, ethnomethodology, repair, mishearing

1 Introduction

What should a social robot display about what it grasps from the world? Should this robot function as a complete "black box", or should it make public some of the information it uses to generate its actions? Ought it to display how many humans it currently sees, the confidence score attributed to the presence of each human, or even—each time a human speaks—what its automatic speech recognition system grasped from this speech? 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 a robot makes publicly available is one way (among many others) to select, process, and represent information it uses in local interactions.

The previous design dilemma is far from recent. A focus on what technological artifacts should describe, discretize, and communicate about their internal processes can be traced back at least to Norman (1988), for whom the surface visibility of action opportunities and of the state of a system were central concerns. As a direct extension of this line of inquiry, the question of the adequate degree of "transparency" for robots has recently received renewed attention. Namely, an emerging strand of studies focuses on what robots should display about their internal functioning (Chen, 2022; Fischer et al., 2018; Kong et al., 2024; Lee et al., 2023; Rossi and Rossi, 2024; Wortham and Theodorou, 2017). However, this question takes on a particular form with (intended to be) social robots: the act of clarifying what information is used by a robot to trigger specific behaviors on its part, or providing a real-time description of what this robot is currently "doing" (Fischer et al., 2018), seems, at first glance, capable of profoundly reconfiguring how humans interpret this robot's conduct—i.e., which social action (Tuncer et al., 2022) these humans take the robot to be currently performing and if this action is relevant, inadequate, or erroneous. Even more, for robots aimed at being conversational partners (Irfan et al., 2024; Uchida et al., 2024), different degrees of transparency appear likely to heavily modify the usual informational ecology in which human conversations are ordinarily held. In this regard, a blind spot persists in debates about social robots' transparency: the consequences of information made publicly available by a robot *on the relevance ascribed to its overall conduct* by humans.

Contributing to the literature on transparency and on failures in human–robot interaction (HRI), our study attempts to establish whether the informational ecology configured by displaying what a robot "hears" is consequential in the way miscommunications emerge and are dealt with compared with interactions that do not feature such information, including ordinary human conversation. To do so, we rely on a corpus of naturalistic and experimental interactions between humans and a version of the Pepper robot that displays the result of its "automatic speech recognition" (ASR) on the tablet placed on its torso. We preface our 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 (which exclusively displayed the transcript of what the robot heard, see Figure 1), while the robot's head and gestures were gazed at less and less. Once this preliminary quantitative picture of participants' gaze attention has been established, we use an ethnomethodological and conversation analytic approach to identify some local interactional phenomena aggregated in these eye-tracking data. Drawing on a detailed analysis of two video fragments, we show that the very emergence of "mistakes" or "errors" in the robot's situated conduct was configured by what it displayed as having "heard". Specifically, we explore how the "same" utterance or gesture of the robot (identical from an external and technical point of view) was treated by some participants as sequentially relevant only as long as they had not read the automatic speech recognition transcript, but was then re-evaluated as troublesome once they had read it. That is, we highlight how the automatic speech recognition transcript was used as a resource in framing the robot's embodied or verbal conduct as inadequate.

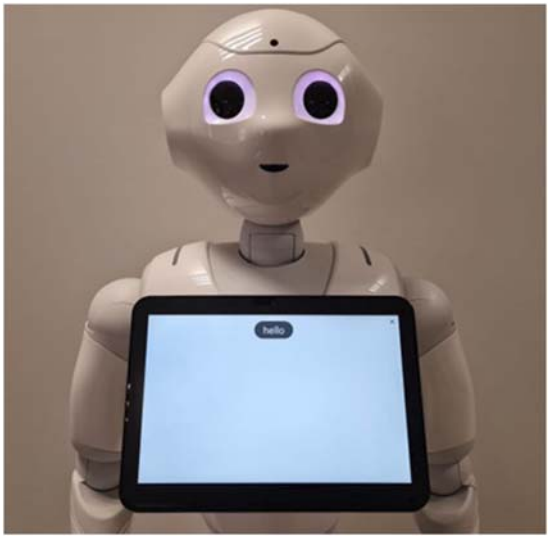

FIGURE 1
Default "speechbar" setting on a Pepper running its latest operating system (NAOqi 2.9). The robot displays "Hello" after a human pronounces this word and stops speaking for more than 200 ms.

The examined fragments constitute exemplars of situations where the robot did not properly "hear" what participants said, both from a technical point of view and as a situated phenomenon. Through their analysis, we emphasize that "mistakes" and "errors" can be approached as practical accomplishments that emerge as such over the course of human–robot interaction rather than as social or technical phenomena pre-categorized by the researcher in reference to criteria exogenous to the activity being analyzed. In doing so, we contribute to HRI by showing that systematically displaying an ASR transcript can play a crucial role in participants' interpretation of a co-constructed action (such as shaking hands with the robot) as having "failed". We observe that, as a specific form of transparency (Lee et al., 2023; Rossi and Rossi, 2024), an ASR transcript did not merely impact how the robot was responded to. Instead, it modified the very definition of "mutual understanding" that was enacted and oriented to as relevant by the human participants in our two fragments. We suggest that the automatic speech recognition transcript led these participants to enact a cognitivist definition of mutual understanding (as similar representations of the world) rather than an interactionist definition (as publicly mutually coordinated actions). Ultimately, besides social robots, we caution that systematically providing such transcripts is a design decision not to be taken lightly; depending on the setting, it may have unintended consequences on interactions between humans and any form of conversational interface.

2 Divergent informational ecologies

2.1 (Mis)hearing as a practical accomplishment

The following study relies on a long-standing ethnomethodological tradition that attends to the locally relevant features of a given setting as "accomplished" (Garfinkel, 1967; Pollner, 1974; Psathas, 1989); that is, as produced and maintained by co-present participants' mutual conduct. This "study policy" (Garfinkel, 1967) 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). From this interactionist (Pelikan et al., 2022), emergentist (Rasmussen, 2019), or dialogical (Linell, 1995; 1998) perspective, no property of a participant's conduct is, by itself and outside of any context, a "mistake" or an "error" (Peyrot, 1982; Zimmerman and Pollner, 2013). A conduct is constituted as troublesome only in and through the conduct of other participants to a setting, rather than as a "given", before any interaction occurs (Meyer, 2019; Pollner, 1974). According to this methodological standpoint, interactions do not feature ontologically "erroneous" actions that are merely detected by co-interactants: "errors" are emergent categories (Mondada, 2017; 2018) displayed in participants' orientation towards a local situation. It is within the immediate context of the interaction, and often through documented conversational methods (accounts, repairs, corrections, clarifications, etc.), that a "mistake" is constructed as such (Albert and Ruiter, 2018; Schegloff, 1992).

This interactionist stance implies a specific understanding of what it means to "(mis)hear" an interlocutor. 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 accountable phenomenon of miscommunication in itself: the only available resource for a co-present participant to detect and repair a potential "mishearing" is this interlocutor's embodied or verbal *response* to the previous turn (Mondada, 2011). If, in a noisy place, a participant happens to perceive absolutely nothing of their interlocutor's voice during a conversation, they may still fortuitously produce a second pair part of the type and form normatively expected (Kendrick et al., 2020) after the previous turn of this interlocutor (for example, by responding with a greeting to a greeting, or by laughing after a joke). In turn, this response may be treated, by co-present participants, as adequate to the action produced by the previous speaker. Should this occur, for all practical purposes, from the strictly emic perspective of participants involved in this ongoing conversation, these conversationalists' conduct is coordinated, and nothing stands out as "repairable" on the surface of interaction (Albert and Ruiter, 2018; Jefferson, 2017). In other words, by themselves, the cognitive processes of a co-participant are not pragmatically consequential phenomena: when a conversationalist is treated by co-present participants as "having (im)properly heard someone" or as "having understood an instruction", it is not this conversationalist's mental activity (as a presumed "private process"—(Wittgenstein, 1953)) that is being indexed as constituting "(im)proper hearing" or "proper understanding" (Coulter, 1999; 2008). This conversationalist's (in)adequate subsequent conduct *will be what "(mis)hearing" is* (Coulter, 2008; Ryle, 1949) rather than the mere evidence of a postulated underlying mental state.

2.2 Human–robot informational ecologies

A fundamental feature of ordinary 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; Garfinkel, 1963; Kristiansen and Rasmussen, 2021). As humans "do not carry MRI machines with them out in the world" (Kerrison, 2018), what happens in conversationalists' brains is never oriented to by their interlocutors as a relevant property of the setting. In most face-to-face interactions, humans' inner cognitive or physiological processes are rarely made systematically transparent to other interactants by being displayed on a screen: namely, in ordinary conversations, humans do not display real-time transcriptions of what they are "hearing", "perceiving", "picturing", etc. In these elementary conditions in which social interaction takes place, one parameter remains constant: whether for the involved actor or the researcher studying video data, participants for whom this would be a practical concern must always rely on "inferential procedures" (Deppermann, 2012) to establish relationships between "discourse and cognition" (Deppermann, 2012)—between what one's interlocutor says and what they may "think", "see", "feel", "perceive", etc[1]. Moreover, several works have noted that, while they are immersed in the practical urgency of their daily interactions, conversationalists are ordinarily seldom interested in detailing the exact phonemes, words, and gestures that they take their interlocutor to have "perceived", "processed", mentally "represented", etc., (Albert and Ruiter, 2018; Dingemanse et al., 2023; Mondada, 2011). Participants in a conversation are generally not concerned "with attaining absolute terminological precision as in certain scientific genres" (Linell and Lindström, 2016).

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, for 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"[2] where the result of the automatic speech recognition will be written once the robot hears no more speech for more than 200 ms (see Figure 1).

1 These basic observations are independent of whether "understanding" is defined as a mental process or an interactional achievement (Ryle, 1949). Neither the cognitivist nor the anti-cognitivist alternative is directly observable by participants in a conversation.

2 The technical documentation for the Pepper robot' "speechbar" and the ASR transcript it displays can be found on the website of Aldebaran (formerly Softbank Robotics), the company that designs and manufactures this robot: https://android.aldebaran.com/sdk/doc/pepper-sdk/ch4_api/conversation/conversation_feedbacks.html (retrieved 16 March 2025).

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 pronounced before any return action from the robot can be achieved. When treated as such, i.e., as intended by its designers[3], an automatic speech recognition transcript 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. Hence, this type of informational configuration currently appears to have no counterpart in interactions that do not comprise robots or conversational agents. There exists no so-called "human–human" face-to-face configuration involving a *scriptural resource* displaying precisely what words were *heard* by an interlocutor in real time[4]. These non-overlapping informational configurations (between interactions involving ASR transcripts and those that do not) belong to the wide set of parameters for which human–robot communication have been described as asymmetric (Frijns et al., 2023).

3 Related work

A common and long-documented challenge for roboticists and users alike is that robots frequently struggle to properly hear human speech when deployed in noisy environments outside the laboratory (Addlesee and Papaioannou, 2025; Foster et al., 2019). Several studies have explored these ASR failures in interactions with robots and vocal user interfaces (Förster et al., 2023; Gunson et al., 2022; Luger and Sellen, 2016) and how the resulting interactional trouble is handled by users (Fischer et al., 2019; Porcheron et al., 2018). However, to the best of our knowledge, no study has yet been undertaken about the specific impact of (erroneous) speech *transcripts* on human–robot interaction or on the perception of devices displaying such transcripts: robots, phones, computers, etc. While some research has begun to explore the role of live transcripts in human–human interaction (Echenique et al., 2014; Gao et al., 2014; Yao et al., 2011), informational configurations involving ASR transcripts displayed by robots or voice agents remain largely unexamined. 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—or to any recent voice-based conversational agent relying on generative AI that provides transcripts in real time.

3.1 Errors and mistakes in human–robot interaction

A rich body of literature directly or indirectly addresses the issue of categorizing types of "errors" or "mistakes" produced by robots during an activity (Honig and Oron-Gilad, 2018; Tian and Oviatt, 2021). This literature overwhelmingly adopts an "etic", exogenous, point of view on the notion of "errors" or "mistakes": "errors" are categorized as such by the researcher, independently of whether and how they emerged as errors for the participants themselves while they were involved in a situated activity with a robot (Honig and Oron-Gilad, 2018). This perspective facilitates the subsequent evaluation of the impact of robots' "failures" on the interaction itself (e.g., Gehle et al., 2015; Schütte et al., 2017), on human behavior (e.g., Giuliani et al., 2015; Hayes et al., 2016; Mirnig et al., 2015; 2017), or on perceptions of the robot (e.g., Ragni et al., 2016; Rossi et al., 2017; Salem et al., 2013; 2015; Tolmeijer et al., 2020). Nevertheless, identifying what constitutes an "error" from the point of view of participants is a challenge that must be regularly addressed even by works that do not specifically focus on this question (Tian and Oviatt, 2021). For example, it is a crucial criterion when asking annotators to code video interactions involving robots (Giuliani et al., 2015). Indeed, a behavior from a robot that constitutes an error from a technical point of view may be treated as entirely appropriate by the study participants; if unaddressed, this distinction may inadvertently become embedded in the coded data (Giuliani et al., 2015).

A branch of research within this literature on errors has extensively explored how a robot can detect that humans are treating its behavior as erroneous; for example, by examining multimodal cues produced by humans when a robot "makes a mistake" (e.g., Bremers et al., 2023; Giuliani et al., 2015; Hayes et al., 2016; Mirnig et al., 2015; 2017; Short et al., 2010). Such robot errors are sometimes intentionally introduced within an experimental context (Centeio Jorge et al., 2023; Mirnig et al., 2017) or observed when they occur in an unplanned way (Giuliani et al., 2015; Mirnig et al., 2015). Through these insights, these works contribute to the more general endeavor of clarifying relevant elements of a situation (including subtle multimodal human behaviors–Vinciarelli et al., 2009) that a robot must perceive and consider to respond adequately (Loth et al., 2015). That is, these investigations of "social signals" attempt to identify, among the potentially infinite number of properties of a setting that can be perceived and attended to, those relevant for the task at hand. Thus, a central contribution of the present article is an inversion of this perspective from which HRI studies are commonly conducted. Rather than focusing on what a robot must detect in human behavior to recognize when it "made a mistake", we explore the resources (provided by the robot or by the broader setting) that humans mobilize to construct robot behavior as an "error".

3 The Aldebaran documentation linked above explicitly frames the "speechbar" (including its public transcription of human speech) as a resource intended for humans to get a better grasp of the situation when interacting with Pepper.

4 Partly similar situations include, for example, speaking to someone connected to a polygraph (i.e., a lie detector) or to someone undergoing an MRI scan. 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.

3.2 Studies on transparency

This research on the emergence of errors connects with the "transparency" framework, which encompasses recent investigations into what aspects of their internal functioning robots should display (Chen, 2022; Fischer et al., 2018; Kong et al., 2024; Lee et al., 2023; Rossi and Rossi, 2024; Wortham and Theodorou, 2017). In this body of work, transparency is defined as "the visibility of underlying processes leading to a reduction of ambiguity regarding a behavior" (Bečková et al., 2024). "Transparency" is therefore distinguished from "explainability" (Verhagen et al., 2021; 2022), with the former emphasizing which aspects of a robot's processing are revealed to users (i.e., "what" the robot is doing) and the latter focusing on the reasons behind the robot's actions (i.e., "why" the robot is doing it—Verhagen et al., 2021).

Within this framework, several studies have explored the perception of humanoid or "social" robots depending on which properties of these robots' internal functioning were made publicly available in real time (Bečková et al., 2024; Fischer et al., 2018; Frijns et al., 2024; Mellmann et al., 2024). Most notably in regard to the informational configuration we will examine, K. Fischer et al. (2018) studied a medical robot that described its "current and upcoming actions" (e.g., "I am going to come closer now"). They found a positive impact of this type of transparency on participants' perceived comfort and trust in the robot. However, in a subsequent work on the same robot, Fischer (2018) suggested that "revealing the robot's real capabilities and underlying processing may hinder, rather than improve, interaction" as "it is especially the implicitness of information that makes human interactions so smooth and seamless". Mellmann et al. (2024) examined whether displaying a robot's goals and behavior impacted participants' perception of its intelligence, anthropomorphism, and agency. Their experimental setup used the Pepper robot's tablet: in one condition, participants interacted with a Pepper robot whose tablet provided a representation of the robot's perception of its environment, goals, and current actions (transparency condition). In another condition, participants interacted with a Pepper robot that did not display anything on its tablet (non-transparency condition). Several of Mellmann et al. (2024) findings lean in favor of higher transparency; however, the communication skills of the robot were reported as higher when the robot was not "transparent".

Overall, this recent body of work on transparency (only briefly outlined in this section) thoroughly analyzes the impact of different degrees of transparency on the way a social robot is perceived by humans (e.g., Angelopoulos et al., 2024; Frijns et al., 2024; Mellmann et al., 2024; Straten et al., 2020). Yet, to date, no study has provided a micro-analytic account of how publicly available transcripts emerge as a significant feature of situated human–robot interaction—if indeed they do. Specifically, one potential pragmatic consequence of social robots' transparency remains to be explored: its role as a resource on which humans may rely as they ascribe (relevant or inadequate) actions to the robot's conduct.

FIGURE 2
Naturalistic data collection setup for the Pepper robot (loading its chatbot), including an RGPD-compliant sign (left), a red stripe on the ground, and a camera.

4 Experimental design

4.1 Setup

4.1.1 Naturalistic and controlled data collections

To investigate the pragmatic impact of the ASR transcript in situated interactions, we base our analysis on our video corpus recorded in July 2022, Paris, at the Cité des Sciences et de l'Industrie, one of the biggest science museums in Europe. This dataset was collected in accordance with the General Data Protection Regulation (GDPR). Ethics approval was obtained from Paris-Saclay Université Research Ethics Board. The first half of this corpus features 100 naturally occurring interactions with a Pepper robot placed in the hall of the Cité des Sciences et de l'Industrie museum (see Figure 2). The second half of this corpus is composed of 108 controlled 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, whose speech was generated by a chatbot designed to "converse" on a wide variety of topics. The two video fragments examined in this work (see Section 7) were recorded during our controlled data collection; however, they constitute exemplars of typical orientations to the ASR transcript identified in an analysis of both the naturalistic and experimental parts of our corpus.

4.1.2 Instructions

For the naturalistic data collection, 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" and provided information in accordance with the

GDPR. 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).

For the controlled data collection, participants were asked to "speak with the robot in the room for 5 min maximum, and 2 min minimum". They were not given explanations regarding the conversational design of the robot, the meaning of the ASR transcript displayed on the tablet, or the robot's abilities. 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. In each interaction, only one participant wore eye-tracking glasses. The following analysis of the interactional consequences of the ASR transcript is therefore based on leisurely conversational interactions, i.e., encounters where *no task was predefined besides the conversation itself.*

4.2 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 one[5]. If any of these services returned a text input that matched one of the possible answers for the robot's chatbot, the robot synthesized this answer. Any utterance produced by a human followed by a silence of more than 200 ms was considered complete, transcribed on the robot's tablet for 3 s (see Figure 1), and responded to by the robot. In all recent Pepper robot models, this 200 ms silence threshold is the default setting used to determine when a user has finished speaking. Similarly, displaying the ASR transcript for 3 s is the default setting in commercial versions of the Pepper robot. We retained these values to ensure our findings remain relevant to Pepper robots currently in use.

After they produced an utterance, participants encountered one of the following behaviors from the robot:

- When the robot was provided with a result from the ASR, and could match a response (in its chatbot) with this specific ASR result, then the robot displayed the ASR transcript on the tablet and triggered its verbal and gestural answer at the same time.
- When the robot was provided with a result from the ASR but had no answer to this specific utterance, then the robot displayed the ASR transcript but indicated verbally that it did not know how to respond.
- When the robot was not provided with a result from the ASR for the sound it received, then the robot displayed a question mark "?" on the tablet and produced an open-class repair initiator (Drew, 1997): "Huh?", "What?", "Pardon?", "Excuse me?", etc.

Except for the ASR transcripts, the tablet was blank. Notably, these transcribed words displayed on the tablet were the sole basis on which the robot constructed a response to what a human had just said. Through this ASR transcript, participants had full access to the information from the outside world that the robot used to generate its verbal response—regardless of whether they oriented to the ASR in this way. No other information (gestures, tone, appearance of the interlocutor, etc.) was relied upon by the robot.

4.3 Robot's conversational design

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 from a language model, nor was it provided with the answers by an external API. However, the robot's responses covered a wide range of domains. Its chatbot could produce a variety of gestural and verbal responses on 70 widely defined topics: e.g., geography, sports, animals, personal information about the robot, which movement it could do, which songs it could sing, etc. The robot's software and conversational design were the same in the naturalistic and experimental data collections.

5 Methods

5.1 Ethnomethodological conversation analysis

The analytical perspective underpinning this research is Ethnomethodological Conversation Analysis (Mondada et al., 2020), hereafter EMCA. Researchers using this micro-analytic approach carefully examine how participants in a setting publicly produce and ascribe actions through the sequential (Sacks, 1995) organization of their talk, gestures, gaze, and overall bodily conduct (Goodwin, 1981; Mondada, 2016). A core tenet of EMCA is its emic orientation: the analysis remains anchored to what participants themselves demonstrably treat as relevant within the interaction rather than imposing external categories or assumptions (Mondada, 2018). Participants' actions are understood as being produced in a way that makes their meaning available to co-participants, who, in turn, display their own understanding through their subsequent actions (Garfinkel, 1967; vom Lehn, 2019). By attending to this publicly accomplished order of interaction (Schegloff, 2007), EMCA researchers describe the practices through which participants render intelligible to one another what they are doing, and what they take others to be doing. This provides a robust foundation for analyzing, for instance, how a human "user" orients to a feature of a robot's conduct (e.g., moving its arm) as accomplishing a specific type of action (Kendrick et al., 2020), e.g., as part of a request, a question, a greeting, or even as initiating a handshake. In essence, the EMCA stance treats the relevance of any detail of a setting as *indexical* (Garfinkel, 1967; Mondada, 2018), i.e., as

5 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.

emergent from participants' hearable and observable orientation to the situation, rather than being *a priori* significant. EMCA research routinely involves a rigorous analysis of video-recorded interactions (Heath et al., 2022; Mondada, 2019), allowing for the precise reconstruction (Ten Have, 2007) of how verbal and gestural conduct is coordinated and finely tuned to the setting (e.g., pointing to a specific feature on a technological device at a particular instant in the talk).

5.2 EMCA and human–robot interaction

In recent years, "human–robot interaction" has 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" (Due, 2023a; Krummheuer, 2015; Pelikan and Broth, 2016; Pitsch et al., 2013; Tuncer et al., 2023). Indeed, this approach is particularly well-suited for investigating how technological artifacts, such as robots, become consequential in situated interaction (Hutchby, 2001)—not by design alone (Lynch, 1995), but through the practical work of participants as they negotiate the intelligibility and relevance of the robot's conduct moment by moment. A crucial consequence of this "study policy" (Garfinkel, 1967) for human-robot interaction is that technical features of an artifact are only relevant insofar as they are treated as such by participants in the setting being observed by the researcher. In this context, what a robot "does" is not taken as a given and stable fact, but is instead grasped through the (sometimes rapidly shifting—Alač, 2016) orientations of the human interactants towards this entity.

5.3 EMCA and ASR transcripts

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. As mentioned previously, these mental processes are not "a discursive phenomenon, which is publicly displayed and collaboratively oriented to by the parties to a conversation" (Deppermann, 2012). To a certain degree, this cognitively agnostic stance extends to robots' inner processes: it does not matter to the EMCA analyst that the social action that a robot is taken to be doing (by participants *in situ*) was anticipated and programmed by its developers as such an action. For example, when a robot produces a waving gesture with its arm, and human participants respond to this conduct as a greeting wave, it is of little importance for the EMCA analyst whether this robot's waving gesture (and its triggering condition) was *also* labeled or categorized as a "greeting" in the robot's programming. Conversely, in this understanding, when a robot publicly transcribes what it hears (and is treated as doing so by co-present interlocutors), the direct phonological receipt (Svennevig, 2004) of what the robot heard becomes "observable-and-reportable" (Garfinkel, 1967) and, therefore, a potentially interactionally relevant property of the setting—both for the EMCA analyst and for co-present participants.

5.4 Eye-tracking as a complement to an ethnomethodological and conversation analytic approach

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 and Rasmussen, 2021). A major pitfall identified by Stukenbrock and Dao (2019) and discussed by Kristiansen and 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, is "specifically unavailable to co-participants and cannot for that very reason have any social significance for the participants in the interaction" (Kristiansen and 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 encouraging the analyst to provide *a priori* social relevance (Schegloff, 1993) to all gaze behaviors produced by participants. 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*. Rather than collecting eye-tracking data to bring "new analytic insights" (Kristiansen and 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. Hence, we compartmentalized, on one hand, the quantitative overview based on the ASR transcript (Section 6) and, on the other hand, the detailed study of interactional phenomena (Sections 7, 8). 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 explored the *social relevance* of this ASR transcript in constituting the robot's conduct as a "mistake" or an "error".

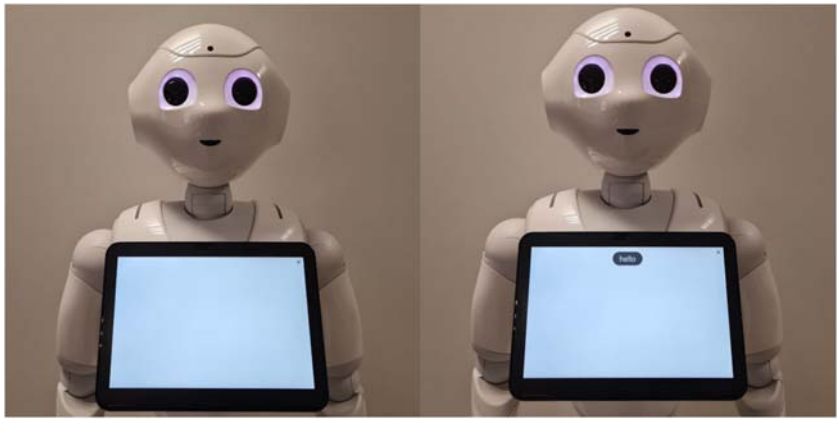

FIGURE 3
Pepper after hearing "Hello" in the No Transcript condition (left) and in the ASR Transcript condition (right).

6 Gaze data analysis

6.1 Conditions

During the experimental data collection, participants (N = 108) were randomly assigned to one of two conditions. We then randomly selected 22 participants from each condition (44 participants in total) for gaze behavior analysis. The total number of participants (44) was based on a practical estimate of how many participants' gaze fixations could be reasonably mapped, manually verified, and analyzed.

- In the "No Transcript" condition (see Figure 3), participants interacted with a robot whose tablet remained completely blank at all times (N = 22).
- In the "ASR Transcript" condition (see Figure 3), participants interacted with a robot displaying an automatic speech recognition (ASR) transcript on its tablet (N = 22).

Aside from the presence or absence of the ASR transcript, the autonomous robot's behavior was identical in both conditions. On average, the 44 participants conversed with the robot for 4 min and 24 s–that is, for 264 s (SD = 94)[6]. In each of these interaction recordings, only the segments relevant to our analysis were thoroughly transcribed in both their spoken and multimodal details (see Section 7).

6 For the sake of simplicity, and to obtain this rough estimate of interaction duration, we define the start of the interaction as the participants' first turn at talk and the end as their final turn at talk. However, identifying the precise onset of a coordinated interaction between the human and the robot would require a more fine-grained analysis (see Holthaus and Wachsmuth, 2021; Rudaz et al., 2023).

6.2 Data preparation

6.2.1 Timecoding of utterances and mapping of fixations

For each of the 44 randomly selected participants, we time-coded every moment during which a speaking turn was produced by a human (see Figure 4).

Meanwhile, a 2D view of the robot was created (see Figure 5). It was divided into several zones:

- Head
- Shoulders
- Arms
- Base
- ASR Transcript Zone (top of the robot's tablet)

We combined zones 1, 2, 3, and 4 under the category "Body and Head", to compare them with zone 5 "ASR Transcript Zone" (the top of the robot's tablet, where the ASR transcript appeared in the ASR Transcript condition). Each gaze fixation from participants was then mapped onto this 2D view[7] (see Figure 4). This provided us with a distribution of participants' fixations over the robot's visible features.

7 Some of the gaze points were mapped manually, while others were mapped using Tobii's automated algorithm. For the automatically mapped points, the "2D mapping layer" (see Figure 4) displays the confidence score: higher values on this graph indicate greater confidence in the algorithm's accuracy for that specific point. However, all recordings were systematically reviewed by a researcher to ensure the accurate placement of each gaze point on the 2D image. If a gaze point was found to be incorrectly mapped—regardless of the algorithm's confidence score—it was manually corrected by the researcher. The Tobii I-VT Attention filter was used for this task.

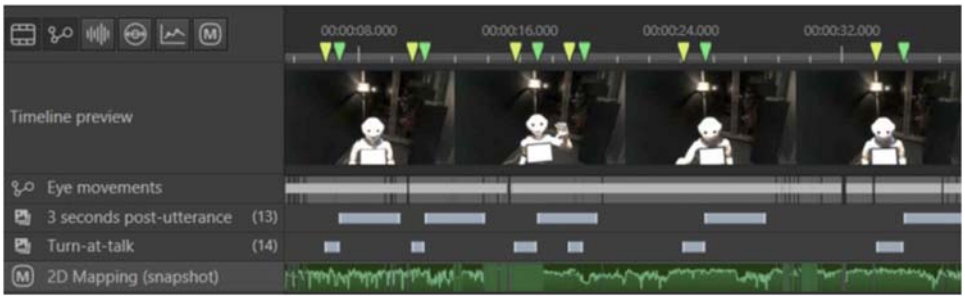

FIGURE 4
Speaking turns produced by a participant coded as Times of Interest (TOI) on Tobii Pro Lab. The "2D Mapping" layer displays each gaze fixation that was mapped onto a 2D picture of the robot.

FIGURE 5
Areas of Interest (AOI) set on a 2D picture of the robot on Tobii Pro Lab.

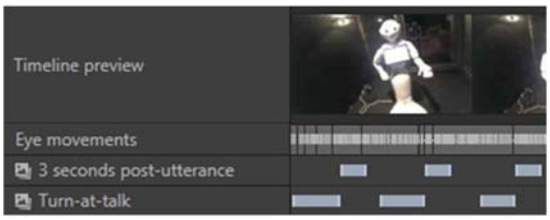

FIGURE 6
Example of timecoded utterances and post-utterance phases for a human participant.

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 s 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' conduct 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 3-s post-utterance phase. To facilitate a meaningful comparison (see Section 6.3), we then also extracted this post-utterance phase for the No Transcript condition. In both conditions, this phase of 3 s 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 words, *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 spoke for only a few minutes. Hence, for each participant, we considered their gaze behavior during the 3 s that followed each of the *first 20 speaking turns they addressed to the robot* (see Figure 6). In other words, in the ASR Transcript condition, we studied participants' gaze behavior during the first 60 s (3 s post-utterance multiplied by their first 20 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.

6.2.3 Binning eye-tracking data in fixed-duration time intervals

After extracting the cumulated 60 s of "post-utterance" time for each participant, we binned[8] them in 50 ms interval bins (for a

8 The R Studio script 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.

total of 1,200 bins per participant). Then, for each 50 ms 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 Section 6.4). That is, these steps served as a preparatory stage for the temporal and regression analyses conducted subsequently: they ensured that the descriptive and inferential statistics rely on the same pre-processed data set.

Segmenting gaze activity (or the overall embodied course of action of participants) into 50 ms bins is widely used in gaze research (e.g., Andrist et al., 2015; Ito and Knoeferle, 2022; Klein Selle et al., 2021; Orenes et al., 2019; Venker and Kover, 2015). Although these concerns ultimately proved not to be directly relevant to the analyses presented in this study, this resolution of 50 ms is fine enough to capture the quick timing of gaze coordination reported in conversation-analytic work[9], yet coarse enough to smooth tracker jitter and minor sampling-rate differences. Accordingly, because 50 ms bins are routinely adopted in eye-tracking research, we defined our bin width *a priori* in alignment with this common practice. Additionally, the shorter bins facilitated the generation of more readable visualizations of gaze evolution over time (see Figure 9).

6.3 Results

6.3.1 ASR transcript condition (N = 22)

In the ASR Transcript condition, during the 3 s that followed their utterances, participants' gaze was mostly directed towards the tablet (see Figure 7). On average, after 20 speaking turns, the average total fixation time per area of interest was:

- 27 s for the "ASR Transcript Zone"
- 18 s 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 Zone" 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 Zone" during this condition (see Figure 8).

The gap between the average time that participants spent gazing at the robot's "Body and Head" and "ASR Transcript Zone" 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 $p = 0.5374$).

6.3.2 No transcript condition (N = 22)

In the No Transcript condition, participants' gaze was mostly directed towards the robot's "Body and Head" during the 3 s that followed their first 20 utterances (see Figure 7) and barely towards the "ASR Transcript Zone" (which remained blank). On average, after 20 speaking turns, the total fixation time per area of interest was:

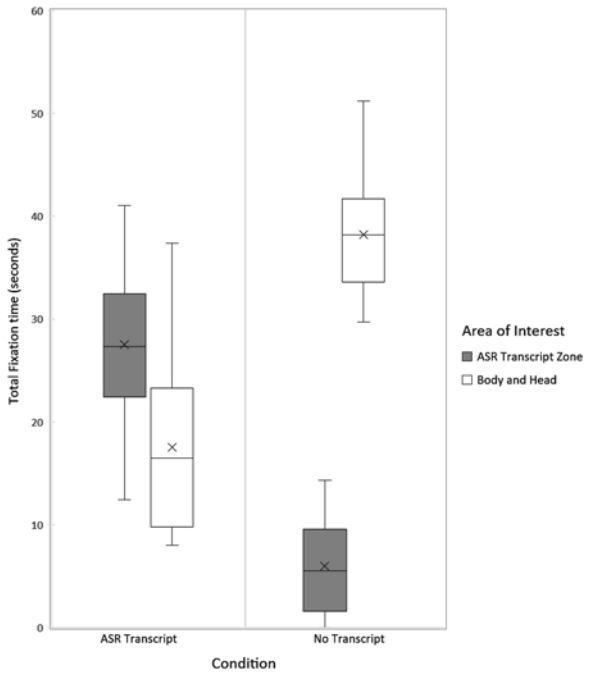

FIGURE 7
Distribution of participants' total fixation time on each area of interest in the No Transcript (N = 22) and ASR Transcript (N = 22) conditions, measured during the 3 s immediately following each human speaking turn.

- 6 s for the "ASR Transcript Zone"
- 38 s for the robot's "Body and Head".

The difference between the average time that participants spent gazing at the robot's "Body and Head" and 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 ($t(21) = -18.43$, $p < 0.0001$). Cohen's d indicates a large effect size ($d = 3.93$). Shapiro-Wilk Test suggests that the data is normally distributed ($p = 0.3044$). 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 8).

6.4 Evolution of post-utterance gaze fixations over the course of the interaction

Because the time participants spent looking at the "ASR Transcript Zone" (top of the robot's tablet) was minimal in the No Transcript condition, we then focused exclusively on the evolution of participants' gaze attention during the ASR Transcript condition.

9 To give one illustrative example of this temporality, Holler and Kendrick (2015) note that in the multi-person interaction settings they studied, "most frequent gaze shifts from current to next speaker were planned around 250 ms and observable around 50 ms prior to turn end."

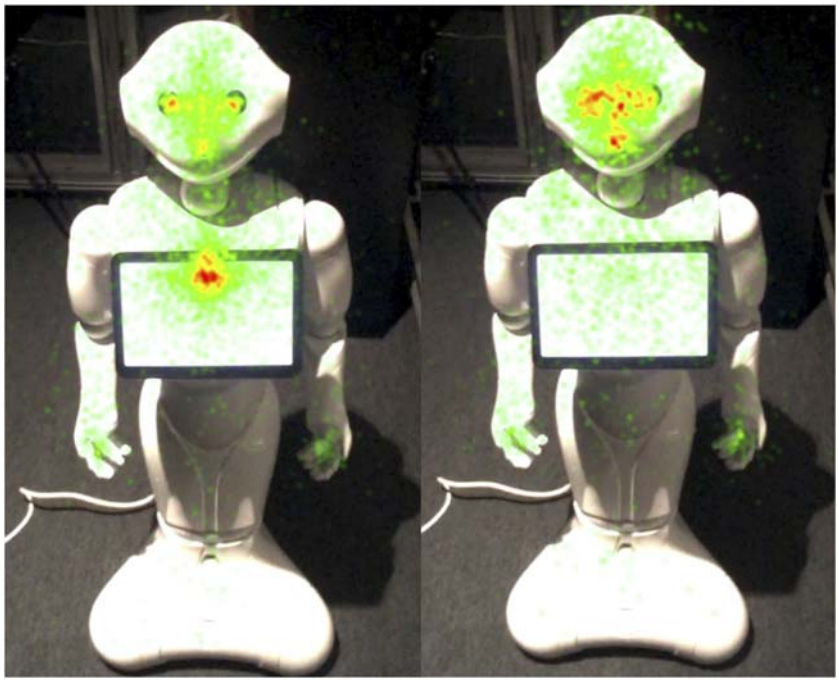

FIGURE 8
Heatmap of participants' fixations after they finished their speaking turn in the ASR Transcript condition (N = 22), left, and in the No Transcript condition (N = 22), right.

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 Zone" (versus the rest of the robot) over their first 60 s of "post-utterance" time (see Section 6.2.3). The results suggest that, in the ASR transcript condition, participants' gaze attention *gradually focused on the "ASR Transcript Zone" when they finished speaking* (see Figure 9). 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 Zone" after each utterance. Note that, to enhance readability and better visualize trends, fixation durations in Figure 9 were smoothed using a generalized additive model (GAM) with a cubic regression spline basis (see footnote 8 for the script used to generate Figure 9). However, all statistical analyses were conducted on the unsmoothed data.

To verify this observed trend, we conducted a linear regression analysis with the bin number (representing consecutive 50 ms intervals[10]) as the predictor and the mean gaze duration as the outcome. The regression model was significant, $F(1,1198) = 18.37$, $p < 0.001$, explaining approximately 1.5% of the variance in gaze duration (adjusted $R^2 = 0.014$). For every 50 ms increase in time (one bin), participants gazed at the "ASR Transcript Zone" for an additional 0.0023 ms on average: $b = 0.0023$, $t(1198) = 4.286$, $p < 0.001$.

A complementary way of investigating the progressive focus of participants' gaze on the "ASR Transcript Zone" was to analyze the difference between the time they spent fixating on this area 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 Zone" and the "Body and Head" of the robot for each 50 ms interval (i.e., for each bin)[11]. 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, $F(1,1198) = 48.61$, $p < 0.001$, explaining approximately 3.9% of the variance in the gaze difference (adjusted $R^2 = 0.038$). For every 50 ms increase in time (one bin), the difference in gaze duration between the "ASR Transcript Zone" and the "Body and Head" increased by 0.0066 ms on average: $b = 0.0066$, $t(1198) = 6.972$, $p < 0.001$. Hence, the post-utterance gradual focus on the

10 In other words, for each bin, we computed the average fixation duration on the "ASR Transcript Zone" across all 22 participants, resulting in a single mean value per bin.

11 In this case, for each bin, we subtracted the mean fixation duration on the "Body and Head" zone (averaged across participants) from the mean fixation duration on the "ASR Transcript Zone" (also averaged across participants).

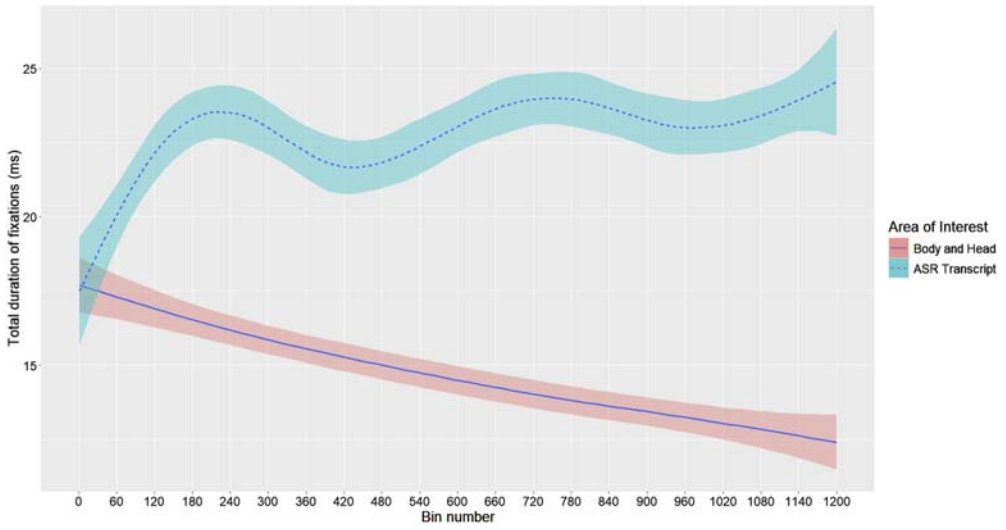

FIGURE 9
Post-utterance fixations over time by area of interest (N = 22). The x-axis is divided into 50 ms bins; a major tick every 60 bins (i.e., 3 s) marks the start of a new "post-utterance phase"—i.e., the 3-s period following each participant's utterance during which the ASR transcript was visible. The plot therefore displays 20 successive 3-s post-utterance phases concatenated in time order, while the speaking turns between these post-utterance phases are omitted. Solid (ASR Transcript) and dashed (Body and Head) lines show the GAM-smoothed mean total fixation duration per bin. The shaded bands around each line indicate ±1 standard error (SE) of these means.

"ASR Transcript Zone" is statistically significant but minimal: the model's R^2 value suggests that the bin number (i.e., each passing 50 ms) explains only a small portion of the variance in gaze time.

Note that, for the previous time-trend analyses, we first averaged fixation duration across all participants within each 50 ms bin (bin n°1 = 0–50 ms…up to bin n°1,200 = 59,950–60,000 ms). This produced a single series of 1,200 mean values that served as the outcome variable in the linear-regression tests reported above. In other words, for the previous linear regressions, *the unit of analysis was the individual time bin, aggregated across all participants*[12].

- For the first regression ("ASR Transcript Zone" only trend), the y-values were the mean fixation durations on the "ASR Transcript Zone" for each bin (1–1,200), and the x-values were the bin numbers.
- For the second regression ("Tablet" vs. "Body and Head" difference), the y-values were the difference in mean fixation durations between the "ASR Transcript Zone" and the "Body and Head" of the robot, averaged across participants per bin. The x-values were again the bin numbers.

12 Consequently, participant-level differences are not modelled in these particular tests. These results should be interpreted as showing a reliable overall shift of gaze toward the ASR Transcript Zone across the cohort, but further analyses that incorporate random effects for participants would be required to draw conclusions about individual variation.

6.5 Discussion: gaze analysis as a preliminary overview

The analysis of participants' gaze behavior provides three main findings:

- When an ASR transcript was displayed, participants gazed at it more than any other element of the robot or of the setting, after they completed a speaking turn.
- This post-utterance focus on the ASR transcript very slightly increased during the interaction at the expense of other parts of the robot (its gestures, its head movements, etc.). After they finished speaking, and while the robot started to respond verbally and gesturally, participants gazed more and more at the robot's tablet, and less and less at the robot's head or body.
- When no transcript was displayed (when the robot's tablet remained blank), participants barely gazed at the tablet.

The previous observations are significant on their own. Yet, this pronounced focus on the transcript begs the question of its *local relevance* (as a resource, as a remarkable phenomenon, etc.) in situated interaction. 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 and 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. This is where a micro-analytic EMCA perspective comes into play, to grasp the endogenous organization of interaction of which the ASR transcript was a potentially relevant feature[13]. 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, was the ASR consequential in how certain actions were "ascribed" (Levinson, 2012) to the robot, once it started to respond? 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 Qualitative fragments

The following fragments were selected from our corpus for three reasons:

1. They are exemplar cases. They display, in an acute form, a common orientation towards the ASR transcript that partially explains (from an emic perspective) the gaze data detailed above. That is, they bring into view some recurring practices that played out behind our quantitative results: i.e., typical practices (indexing the ASR transcript) observable during the interactions that took place in the previously described experiment. In our experimental data, in the condition where the ASR was present, 23 interactions (out of 54 recorded) contained at least one demonstrable reference to the ASR transcript by a participant; often several times within the same interaction. In particular, 9 of these 23 interactions contained at least one instance of the phenomenon we point to in the following excerpts: namely, the retrospective re-evaluation of the robot's action relying on the ASR transcript[14].
2. They feature (what is treated by participants as) mishearings or misunderstandings from the robot. As such, these troublesome exchanges are more likely to reveal participants' orientation towards the ASR transcript compared to perfectly smooth interactions (Förster et al., 2023). 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 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 (Due, 2023b; Garfinkel, 2002).
3. They display "re-evaluations" of an action previously ascribed to the robot. These fragments were selected as cases in which the meaning of the robot's conduct is hearably negotiated between interlocutors. As mentioned above, many participants explicitly relied on the ASR transcript as a resource in (publicly and visibly) ascribing an action to the robot; the following fragments are especially clear examples of such cases. In a different vocabulary, these interactions involving "re-evaluations" of the robot's actions functioned as "perspicuous settings" (Garfinkel, 2002). In particular, they allowed an observer to grasp what action the robot was initially ascribed by a participant before the ASR transcript was introduced as a relevant parameter of the setting by another participant.

These fragments were transcribed following Mondada's multimodal transcription conventions for embodied conduct (Mondada, 2016) and Jefferson's conventions for verbal conduct (Jefferson, 2004). Transcription conventions are provided at the end of this paper. In each fragment, PA1 is the participant wearing the eye-tracking glasses.

7.1 The ASR transcript as a resource for third parties in contesting the relevance of (embodied or verbal) responses from the robot

7.1.1 Fragment 1: using the ASR transcript in contesting the definition of the situation as one of intersubjectivity

The first fragment is presented below (Figure 10).

7.1.2 Analysis

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, although 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 (L.5). 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. He points out (L.7) that ROB heard "laver" ("cleaning") instead of "lever" ("looking up") and provides a possible cause for ROB's mishearing ("you do not speak loud enough", L.7). Significantly, in doing so,

13 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: we untangle some of the interactional phenomena intertwined in the statistical results.

14 However, since interactions varied in length, etc., this number—while not meaningless—does not lend itself easily to strict or comparable quantification. We rely on the EMCA convention of not emphasizing such figures, as doing so would shift attention away from the phenomenon as "one species" of interactional practice (Ten Have, 2007), regardless of its recurrence.

```
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 he/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 are not speaking loud enough by "clean up" he/it meant "look up"
   pa1    -->@turns towards ROB-->
8.        (2.0)
9. PA1    il@a marqué laver?#
          he/it wrote "clean up"?
   pa1   -->@turns towards PA2-->
   fig                  #Figure 10
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?
```

FIGURE 10
Fragment 1. Multimodal transcript of participants' conduct following Mondada's and Jefferson's conventions. Bold lines represent participants' talk, while the lines immediately below capture embodied conduct. Content displayed on the robot's tablet is highlighted in orange. This fragment is described in detail in the following subsection.

PA2 indexes the content of the ASR transcript (L.7) a few seconds after this transcript disappeared from ROB's tablet (L.4). This delay indicates that PA2 has read the ASR transcript *before* a form of trouble was publicly manifested by PA1 (L.5) and not only *in reaction* to such trouble.

Once details about what was written on the ASR transcript are 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 up?" (L.9, Figure 11), PA1 manifests she is dealing with new properties of the situation, on which she did not previously rely to interpret ROB's conduct. This reframing is observable when, after this turn, PA1 immediately casts the robot's previous answer as inadequate by repeating her own previous question (L.11) while insisting on the term which was stated, by PA2, to be mistranscribed ("can you > look < up?"). This re-evaluation of ROB's conduct illustrates how much the social significance of verbal and embodied responses from the robot can be contingent upon what the robot transcribes on its tablet.

Crucially, here, this re-evaluation is not about the *type* of response that ROB does. In this fragment, ROB's conduct rejects a request: at this level of description, on the surface of the interaction, ROB produces an adequate second pair part. However, this action (of refusing a request) is re-evaluated as *indexed to the wrong speaking turn.* Therefore, what is revised here is not the "action" in a generic sense. Rather, what is reconsidered is this action as activity-relevant, i.e., its status as a response to what was really requested by human participants. ROB's response is not problematized as a response to a request but *as a response to a specific request.*

Finally, although in this case it is the bystander who monitored and verbalized the robot's mistranscription of the main speaker's utterance, referencing the ASR transcript is not exclusively a bystander's role in our corpus of multiparty interactions. In the condition where the ASR transcript was available, 23 out of 54

FIGURE 11
PA1 points at ROB's tablet while asking PA2 for confirmation about what it previously displayed.

recorded interactions featured at least one demonstrable reference to the transcript by a participant (See Section 7), often multiple times within a single interaction. In 8 of these cases, a bystander initiated at least one speaking turn referring to the transcript. The fragment selected here was chosen for reasons of analytical clarity: the bystander's denial of the main speaker's interpretation allows us to grasp in minute detail how participants negotiated the issue of "what action the robot had been performing." This case aligns with prior research documenting the consequentiality of bystanders' contributions in multiparty human–robot interactions (see, e.g., Krummheuer, 2015; Pitsch, 2020; Rudaz and Licoppe, 2024).

7.1.3 Fragment 2: using the ASR transcript in accounting for a breakdown in interaction after it arose

The second fragment is presented below (Figure 12).

7.1.4 Analysis

This fragment illustrates a nominal use of the ASR transcript as a resource to identify, account for, or repair trouble in interaction. From a purely technical point of view, it begins with a speech recognition error by the robot, which displays "can you close your hand" on its tablet, although PA1 asked "can I shake your hand?". PA1 interprets ROB's subsequent left arm gesture as the initiation of a handshake (L.8, Figure 13), reconsiders this course of action when ROB lowers its left arm and raises its right one, frees her right hand from the object she was holding (L.10), and is finally confronted with ROB's closed hand when she approaches her right hand (L.11 and L.13, Figure 14). After this long sequence of visible struggle to coordinate between PA1 and ROB, PA2 produces an account of ROB's understanding of the situation (L.13). Over the course of the experiment, PA2 had regularly monitored the transcripts displayed on the robot's tablet (not in transcript). However, as this specific scene unfolded between PA1 and ROB, he had remained silent until its completion. Prefacing his account with a "no but" (L.13), he states that the robot "understood 'close the hand'", 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.

On the interactional surface (independently of the question of whether PA1 is pretending the robot is collaborating to achieve a

```
1. PA1    £(.).hhh@je *peux t'serrer * la main?
                      can I shake your hand?
   pa1 -->£gazes at ROB's head-->
   pa1            @leans forward-->
   pa1               *opens her hand*
2.        (1.0)@(0.4)%(0.2)£(0.1)
   pa1      -->@
   rob            -->%displays "tu peux fermer la main"-->
                              "can you close your hand"
   pa1                  -->£gazes at belly screen-->
3. ROB    quelle $ main?

          which hand?
   rob          $lowers and raises arms-->
4.        (0.1)£(0.1)%(0.1)*(0.3)
   pa1      -->£gazes at ROB's right arm-->
   pa1                    *extends arm towards ROB's right arm-->
   rob            -->%
5. PA1    eu::: (.) la gau$che/
                    the left one
   rob                 -->$
6.        (0.1)£    (1.4)%(0.1)    £(0.2)
   pa1      -->£gazes at ROB's head£gazes 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)
   pa1         £gazes at ROB's right hand-->
   pa1               *extends right arm towards ROB's right hand-->
   rob                    $retracts right arm-->
   fig                   #Figure 11
9.        (0.4)*(0.1)£(0.1)
   pa1         *retracts right arm->
   pa1               £gazes at belly screen-->
10.       (0.2)$(0.2)£*(0.1)+(0.4)
   rob      -->$extends left arm->
   pa1           -->£gazes at ROB's left hand->
   pa1            -->*switches eye-tracking recorder to right hand-->
   rob                   +gazes at its left arm-->
11.PA1    ah.*$okai
          oh okai
   pa1    -->*extends left arm towards ROB's left hand-->
   rob     -->$closes its left hand-->
12.       (0.3)%(0.1)£(0.1)
   rob      -->%
   pa1           -->£gazes bellyscreen
13.PA2    non*      mais# il a compris      *$fermer la main
          no        but he/it understood      close the hand
   pa1    -->*touches top of ROB's left hand*
   rob                                  -->$lowers left arm-->
   fig                #Figure 12
14.       (0.2)$(0.3)+(0.5)
   rob      -->$
   rob           -->+gazes PA1>>
15.PA1    nan::* mais@£    c'est s- @       £oui
          no     but       it's             yes
   pa1         *leans head towards PAR2>>
   pa1              @----shrugs----@shrugs-->
   pa1            -->£gazes ROB's right arm£gazes belly screen>>
16.PA1    d'accord@ serrer la main(.)ok/* (.) ok/ ca va
          all right shake the hand ok ok it's all right
   pa1         -->@
```

FIGURE 12
Fragment 2. Multimodal transcript of participants' conduct following Mondada's and Jefferson's conventions. Bold lines represent participants' talk, while the lines immediately below capture embodied conduct. Content displayed on the robot's tablet is highlighted in orange. This fragment is described in detail in the following subsection.

mutual handshake), this fragment therefore displays two different treatments of what is (for an external observer) the "same" gesture produced by the robot. The first treatment, by PA2, indexes 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 "what is going on" in this situation by (publicly and demonstrably) relying exclusively on the robot's verbal and gestural conduct. However, PA2 treats these two interpretations as unable to coexist. By prefacing his turn with a "no" (Lerner and Kitzinger, 2019), L.13, PA2 does more than accounting for PA1's trouble to properly complete her handshake gesture with the robot. PA2's turn belongs to a category of "no-prefaced rejection

FIGURE 13
PA1 extends her hand towards ROB's right hand.

of the trouble source" that, following Lerner and Kitzinger (2019), "explicitly disallow[s] the trouble source as having been mistaken". That is, PA2 negates that a mutually understood handshake was ever present, i.e., that the apparently coordinated gestures of ROB and PA1 manifested an intersubjectively shared reality between them. 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. 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 then. The public availability of the ASR was central in the re-evaluation of the situation as a misunderstanding rather than as a (non-smooth) mutually coordinated handshake.

8 Discussion: the practical relevance of the ASR transcript

8.1 Troublesome ASR transcripts could override the verbal and gestural response of the robot

In the video fragments analyzed above, the robot's ASR transcript was used as a resource unlike any typically found in ordinary human face-to-face interaction. The resulting informational configuration was consequential in shaping the timing and composition of repairs initiated by human participants. It impacted the terms participants treated as troublesome in previous turns (as words "misheard" or "misunderstood" by the robot, e.g., Fragment 2, L.13) and even which syllables were stressed in their repairs (e.g., Fragment 1, L.11). For the two groups of participants we examined, in the presence of a systematic ASR transcript, what "interacting with a robot" consists of practically was pulled further away from the concrete processes that constitute human–human interaction: the transcript modified the ordinary practices that were

FIGURE 14
PA1 touches ROB's left hand while PA2 states that ROB understood "close the hand". The two photos are different angles of the same moment.

(or could be) produced, and the pragmatically relevant properties of the interaction[15].

Note that 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' conduct (e.g., the sudden initiation of a repair sequence although the robot's conduct was treated as relevant so far) is only fully understandable if the presence of an ASR transcript in front of them is taken into account. If a robot were to attempt to interpret participants' reactions without this crucial contextual element of "what it displayed on its tablet", it would 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 (e.g., 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. These observations therefore add to the list of key parameters that embodied agents should perceive and consider (Giuliani et al., 2015; Mirnig et al., 2017; Vinciarelli et al., 2009) to better understand what human participants are indexing in their speaking turns–thereby helping the robot to identify when its actions are being framed as a "mistake" by human participants (Frijns et al., 2024; Giuliani et al., 2015; Mirnig et al., 2017).

8.2 Impact on the "sequential plasticity" of the robot's turns

Could it be that some fundamental conditions or encompassing mechanisms of human sociality (Mondémé, 2022) were highlighted or altered by the unusual informational ecology visible in the previous fragments? Garfinkel's early experiments (Eisenmann et al., 2023; Garfinkel, 1967; Ivarsson and 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 prior conduct. 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 roboticists who programmed this robot, the result of an unforeseen sequence of events in which the robot did not, in fact, adapt to the humans—see, e.g., Pelikan et al. (2020), Rudaz et al. (2023), Rudaz and Licoppe (2024), and Tuncer et al. (2023). When considering this regularly "fortuitous" (Tuncer et al., 2023) character of a 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. For example, in Fragments 1 and 2, the ASR transcript contributed to removing the vagueness of the robot's response "as an answer-to-the-question" (Garfinkel, 1967—see Fragment 1) and "as an answer-to-the-request" (see Fragment 2): the exact phonological form (Svennevig, 2004) of what the robot was responding to was apparent to the main human speaker or, at least, to a bystander.

The resulting informational configuration narrowed down the range of behaviors from the robot which could be directly treated as (adequate) *social actions*—that is, as responsive to the situated interaction and as "making relevant a set of potential next actions" (Tuncer et al., 2022). For example, 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 et al., 2022) or to produce other practices—well documented in HRI—through which a robot is ordinarily maintained as a competent interactant by human participants (e.g., Rudaz and Licoppe, 2024). If, in interaction, each action "tests the hypothesis a participant has about a co-participant's response to her/his action" (vom Lehn, 2019), these "experiments in miniature" (vom Lehn, 2019) took a very different form in these human-robot interactions where the robot featured a transcript of what it "heard".

15 A possible hypothesis is that the ASR transcript may have modified "the limits of what [the parties in a conversation] will seek to bring to determinacy" (Liberman, 1980): that is, what could be "passed" over (Garfinkel, 1967; Liberman, 1980) and left tacit as non-consequential "for current practical purposes" (Albert and Ruiter, 2018; Linell and Lindström, 2016; Schutz, 1972).

8.3 Does the ASR transcript enforce a cognitivist definition of mutual understanding?

Based on the previous discussion, the orientation of the previous participants toward the ASR transcript can be described in two ways.

1. A safe manner of verbalizing these participants' practices is to say that it became a "members' problem" (Garfinkel, 1967) that the ASR transcript displayed an adequate receipt of their previous turn. The performance that was expected from the robot was not merely gestural and verbal. It was also an "auditory" performance. For all practical purposes, it was relevant for them that the robot displayed a too-dissimilar transcript of what they said.
2. 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 these participants (and by many others in our larger corpus) towards the ASR transcript: they were led to enact a cognitivist definition of mutual understanding. In other words, these participants publicly cared about what was inside the robot's "head" or "algorithm" when they treated the situation as repairable.

In this second interpretation, as a technological artifact, the ASR transcript facilitated a specific orientation to human–human understanding (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 enforced a definition of "understanding"[16] as sharing a "similar mental representation about the world" (Albert and Ruiter, 2018) rather than, as it is generally described in conversation analytic works, being "related to the next action achieved by the co-participant" (Mondada, 2011—see also Coulter, 2008; Dingemanse et al., 2023; Mondémé, 2022). It favored an orientation to progressivity in conversation as based on a cognitively shared reality (Searle, 1969), rather than as "being able to 'go on' with each other" (Sterponi and Fasulo, 2010). From a different theoretical perspective, the ASR transcript led participants' "theory of mind" of the robot (Angelopoulos et al., 2024; De Graaf and Malle, 2017; Wortham et al., 2016) to take a central place in their publicly displayed interpretation of the robot's actions. In particular, the *intention* of the robot (De Graaf and Malle, 2017; Malle and Knobe, 1997) to produce the action it was originally taken to be producing (a handshake, a refusal to a specific request) became a central concern.

For the bystanders in fragments 1 and 2, it did not only matter that the robot produced verbal and gestural answers to what was just said, but *it also mattered that these verbal and gestural answers stemmed from an entity that had properly (displayed to have) heard what had been said*. In order to ascribe actions to that entity, in order to grasp what it was "doing", an element internal to that entity mattered (what phonological reconstruction it had produced of the previous talk), possibly more than its verbal and gestural response. In other words, displaying the (in)adequate phonological reconstruction of the utterance that a turn responded to regularly impacted the relevance of this turn as an adequate response–it played a key role in participants' framing of that conduct as a "failure", a "mistake", a "misunderstanding", i.e., as a troublesome behavior. The ASR transcript rendered practically feasible a cognitivist definition of mutual understanding "as a mental process" (Ryle, 1949; Wittgenstein, 1953), i.e., as "grasping what is in the other's mind" (Shotter, 1996). On this view, participants displayed 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. This claim is partially reinforced by the comparison with our participants' conduct in the second condition of the experiment, where no ASR transcript was available. In this "No Transcript" condition, we found no sequence during which the robots' exact hearing of what a human participant just said emerged as a necessary condition to continue with the ongoing activity. This is strikingly different from what we examined in the previous video excerpts.

Nevertheless, this interpretation (that participants were led to enact a cognitivist definition of mutual understanding) inevitably goes beyond what our data can substantiate based on participants' observable and hearable orientations to the setting. Since what is in the robot's "head" (i.e., what is processed by its algorithms) can only be documented through the ASR transcript displayed on the tablet, it is ultimately impossible to rigorously distinguish a strict orientation to the transcript from an orientation to the robot's internal states. From a strictly observable standpoint, what can be conclusively demonstrated is that the content displayed on the robot's tablet was recurrently relevant for participants—for all practical purposes—in progressing or repairing the interaction. In mundane conversation, participants ordinarily 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 and Lindström, 2016; Schutz, 1972). However, in the presence of an ASR transcript, each time "what was heard by the robot" differed from what was publicly said by the human, this information was accessible by said human. Consequently, a minimal reading of our data is that the presence of an ASR transcript constantly threatened to trouble participants' practical sense of what was a disregardable non-significant mismatch to reciprocal alignment (Shotter, 1996; Sterponi and Fasulo, 2010) or, conversely, what was a locally relevant misunderstanding which should be addressed–even at the cost of a "time out" (Heritage, 2007) in the current course of action (Bolden, 2012; Schegloff et al., 1977).

16 Ivarsson and 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 believe they are analytically valid categories emerging from the fragments under study and from our larger corpus.

8.4 Implications for design: ASR transcripts and "conversational moves" available to a robot

Some design considerations can be drawn from the empirical observations above. 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 (in leisurely conversational interactions, rather than heavily task-oriented contexts) unavoidably complicates the accomplishment of "conversational moves" that rely on the absence of an internalist window on "what was heard".

For example, the conversational move that Liberman (1980) names "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" 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—as evidenced at a later point in the conversation. 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. In the case of a mishearing, the conversational "move" of gratuitous concurrence (e.g., responding "yes" to a sentence that one did not hear entirely) is therefore not replicable in the same conditions for a robot displaying a transcript, 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 (e.g., by laughing although it did not hear what its interlocutor said, by tentatively producing the change-of-state token "oh", etc.,) would require that the human either does not publicly attend to the ASR transcript as a relevant property of the interaction or orients to it as something else than a transcript.

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) as humans to maintain the sequential relevance of its turns and, through this, its perceived conversational competence.

8.5 Limitations and implications for future research

The current study examined only one configuration of transcripts: the ASR transcript appeared on a tablet positioned on the robot's torso, for a fixed duration of 3 s, immediately after the robot detected the end of a human utterance. This design involved a single criterion for triggering the ASR transcript display—namely, the recognition of human speech by the robot. Future research could explore alternative configurations, including:

1. Displaying the transcript only when the ASR confidence score falls below a certain threshold (i.e., when the human is likely to have been "misheard").
2. Displaying the transcript alongside the corresponding ASR confidence score (to highlight when the robot is likely to have mistranscribed what was just pronounced).
3. Displaying transcripts selectively, based on the type of human turn.

With regard to the third point, a promising alternative would be to display an ASR transcript only after specific kinds of utterances, rather than systematically after every turn. This would build on the parallel between the interactional function of written transcripts and verbal "displays of hearing" (Svennevig, 2004) occasionally produced in human talk, such as repeating one's interlocutor's utterance. Indeed, humans do not routinely produce such displays of hearing after every turn of their interlocutor. On the contrary, displays of hearing are accomplished in response to specific local contingencies, where they play a meaningful role in coordinating action. As Svennevig (2004) notes, conversationalists only "sometimes need to display explicitly that they have registered a piece of information". This "need" arises mainly when an interlocutor's 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): for example, *when a participant is being provided with someone's name, a date, or a schedule*. A possible design would thus be to display an ASR transcript exclusively after these types of turns[17].

9 Conclusion: vagueness as a core feature of interaction

In the previous fragments, what locally emerged as interactional trouble (a "failure" to co-produce a handshake, a "mishearing" or "misunderstanding" on the robot's part, etc.,) was not a pre-given. That "something went wrong" in the ongoing activity was not merely discovered by participants: it was publicly displayed and enforced as an accountable phenomenon. In our two examples, this interactional work was carried out by bystanders standing behind the participant speaking directly to the robot: they challenged the main speaker's visible orientation toward the interaction as requiring no repair. Crucially, in

17 Of course, rather than following a simplistic rule, an optimal solution would be to display an ASR transcript *only when it is interactionally relevant for the co-present members themselves*. However, the local applicability of such a generic guideline for an artificial agent remains uncertain: producing a transcript "only when situationally relevant" confronts the classic problem of relevance and commonsense understanding in AI (McCarthy and Hayes, 1981; Minsky, 1988; Mitchell, 2021).

treating the robot's conduct as inadequate, these bystanders heavily indexed the information provided by the automatic speech recognition transcript—even when the robot's embodied and verbal conduct had previously been treated as entirely relevant. In sum, while our preliminary quantitative analysis revealed that participants' gaze frequently focused on the automatic speech recognition transcript in our corpus (see Section 6) at the expense of other parts of the robot, a qualitative micro-analytic approach indicates that, in situated cases, this systematic transcript also served as a pivotal local resource for some participants in publicly accounting for the robot's conduct as either relevant or irrelevant. From this perspective, following Fischer's (2018) suggestion, transparency could be re-evaluated in terms of opportunities for action provided to human participants.

Provided the analysis we outlined in this work holds true, the previous fragments exemplify in various ways that attempting "to produce unambiguousness" (Meyer, 2019) can "unnecessarily complicate the practical situation" (Meyer, 2019)[18]. Transcriptions or descriptions are not "views from nowhere" (Nagel, 1989), merely standing next to the action: they are part of the setting, available to be interpreted as constituents of the ongoing action of the participant producing them (Mair et al., 2021). As Garfinkel (1967) mentions, "members' accounts, of every sort, in all logical modes, with all their uses, and for every method for their assembly are constituent features of the settings they make observable". When treated as such (i.e., when interpreted by participants as "displays of hearing"—Svennevig, 2004), automatic speech recognition transcripts impacted what actions were ascribed to the robot (Levinson, 2012). To clarify the generalizability of the preceding analyses, we suggest that future quantitative-oriented research examine the overall quality and smoothness of conversational interactions with voice agents that display ASR transcripts compared to agents that do not; for example, by employing the Human–Robot Interaction Conversational User Enjoyment Scale developed by Irfan et al. (2025). Such a comprehensive approach may help determine to what extent ASR transcripts serve as a "lesser evil"—potentially prompting more repair sequences while preventing deeper misunderstandings—or, conversely, whether these transcripts disrupt the usual inner workings of human conversation too severely.

Indeed, as a "feature", the—systematic—ASR transcripts might belong to a broad category of technologies that endeavor "to remedy a vagueness that is actually required" (Eisenmann et al., 2023), at least, in leisurely conversational interactions. In the two video fragments analyzed above, systematic and non-context-sensitive ASR transcripts hindered intersubjectivity enacted as a "situated, temporarily sustained and only partially shared experience" (Linell and Lindström, 2016). As Linell and Lindström (2016) suggest, experiencing intersubjectivity is dependent on the fact that, in human-human interaction, "a single contribution within a reasonably coherent sequence presupposes an understanding of the prior contribution(s)". Yet, in the previous fragments, the display of ASR transcripts removed this essential presupposition that one's previous contribution was understood: when the robot responded verbally, the ASR transcript could override the relevance of the robot's verbal response *as a response to a specific turn*. In this new configuration, the "weave of interactional moves" (Linell and Lindström, 2016) produced between humans and robots was less likely to offer the (superficial) coherence that is the bedrock of an experience of the ongoing interaction as intersubjectively shared.

18 Within a different theoretical framework, a systematically displayed ASR transcript constitutes an exception to a "maxim of minimization" (Levinson, 1987).

10 Transcription conventions

10.1 Transcription of talk follows Jefferson's 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 (Jefferson, 2004)

10.2 Embodied actions were transcribed using Mondada's multimodal transcription conventions

**	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 (Mondada, 2016).

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

```
PA1  Turn at talk from a human participant
     (PA1, PA2, etc.)
ROB  Turn at talk from the robot
pa1  Multimodal action from a participant
     (pa1, pa2, etc.)
rob  Multimodal action from the robot
fig  Screenshot of a transcribed event
£    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
```

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Paris-Saclay Université Research Ethics Board. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

DR: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing. CL: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors would like to thank the Cité des Sciences et de l'Industrie and its employees for graciously providing them with the necessary legal and material conditions for conducting this experiment. The material used in this manuscript was collected as part of Damien Rudaz's unpublished PhD research (Rudaz, 2024).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Addlesee, A., and Papaioannou, I. (2025). Building for speech: designing the next-generation of social robots for audio interaction. *Front. Robotics AI* 11, 1356477. doi:10.3389/frobt.2024.1356477

Alač, M. (2016). Social robots: things or agents? *AI and Soc.* 31 (4), 519–535. doi:10.1007/s00146-015-0631-6

Albert, S., and Ruiter, J. (2018). Repair: the interface between interaction and cognition. *Top. Cognitive Sci.* 10, 279–313. doi:10.1111/tops.12339

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., and Shaffer, D. (2015). Look together: analyzing gaze coordination with epistemic network analysis. *Front. Psychol.* 6, 1016. doi:10.3389/fpsyg.2015.01016

Angelopoulos, G., Lacroix, D., Wullenkord, R., Rossi, A., Rossi, S., and Eyssel, F. (2024). Measuring transparency in intelligent robots. doi:10.48550/arXiv.2408.16865

Bečková, I., Pócoš, Š., Belgiovine, G., Matarese, M., Sciutti, A., and Mazzola, C. (2024). A multi-modal explainability approach for human-aware robots in multi-party conversation. doi:10.48550/arXiv.2407.03340

Bolden, G. B. (2012). Across languages and cultures: brokering problems of understanding in conversational repair. *Lang. Soc.* 41 (1), 97–121. doi:10.1017/s0047404511000923

Bremers, A., Pabst, A., Parreira, M. T., and Ju, W. (2023). Using social cues to recognize task failures for hri: overview, state-of-the-art, and future directions. doi:10.48550/ARXIV.2301.11972

Centeio Jorge, C., Bouman, N. H., Jonker, C. M., and Tielman, M. L. (2023). Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Front. Robotics AI* 10, 1143723. doi:10.3389/frobt.2023.1143723

Chen, J. Y. C. (2022). Transparent human–agent communications. *Int. J. Human–Computer Interact.* 38 (18–20), 1737–1738. doi:10.1080/10447318.2022.2120173

Coulter, J. (1999). Discourse and mind. *Hum. Stud.* 22, 163–181. doi:10.1023/a:1005484316659

Coulter, J. (2008). Twenty-five theses against cognitivism. *Theory, Cult. and Soc.* 25 (2), 19–32. doi:10.1177/0263276407086789

De Graaf, M. M., and Malle, B. F. (2017). "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI fall symposium series*, 2017.

Dennis, A. (2003). Skepticist philosophy as ethnomethodology. *Philosophy Soc. Sci.* 33 (2), 151–173. doi:10.1177/0048393103033002001

Deppermann, A. (2012). How does 'cognition' matter to the analysis of talk-in-interaction? *Lang. Sci.* 34 (6), 746–767. doi:10.1016/j.langsci.2012.04.013

Deppermann, A. (2018). Inferential practices in social interaction: a conversation-analytic account. *Open Linguist.* 4 (1), 35–55. doi:10.1515/opli-2018-0003

Dingemanse, M., Liesenfeld, A., Rasenberg, M., Albert, S., Ameka, F. K., Birhane, A., et al. (2023). Beyond single-mindedness: a figure-ground reversal for the cognitive sciences. *Cognitive Sci.* 47 (1), e13230. doi:10.1111/cogs.13230

Drew, P. (1997). Open' class repair initiators in response to sequential sources of troubles in conversation. *J. Pragmat.* 28 (1), 69–101. doi:10.1016/s0378-2166(97)89759-7

Due, B. L. (2023a). "Laughing at the robot: three types of laughables when interacting with pepper," in *Interacting with robots and social agents*. Editor P. Lang, ACII.

Due, B. L. (2023b). "The practical accomplishment of living with visual impairment," in *The practical accomplishment of everyday activities without sight*. 1st ed. (London: Routledge), 1–25. Available online at: https://www.taylorfrancis.com/books/9781003156819/chapters/10.4324/9781003156819-1 (Accessed January 10, 2025).

Echenique, A., Yamashita, N., Kuzuoka, H., and Hautasaari, A. (2014). "Effects of video and text support on grounding in multilingual multiparty audio conferencing," in *Proceedings of the 5th ACM international conference on collaboration across boundaries: culture, distance and technology* (New York, NY, USA: CABS '14. Association for Computing Machinery), 73–81. doi:10.1145/2631488.2631497

Eisenmann, C., Mlynář, J., Turowetz, J., and Rawls, A. W. (2023). "machine down": making sense of human–computer interaction—Garfinkel's research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance. *AI and Soc.* doi:10.1007/s00146-023-01793-z

Fischer, K. (2018). "When transparent does not mean explainable," in *Explainable robotic systems—workshop in conjunction with HRI 2018*.

Fischer, K., Weigelin, H. M., and Bodenhagen, L. (2018). Increasing trust in human–robot medical interactions: effects of transparency and adaptability. *Paladyn, J. Behav. Robotics* 9 (1), 95–109. doi:10.1515/pjbr-2018-0007

Fischer, J. E., Reeves, S., Porcheron, M., and Sikveland, R. O. (2019). "Progressivity for voice interface design," in *Proceedings of the 1st international conference on conversational user interfaces* (New York, NY, USA: Association for Computing Machinery). doi:10.1145/3342775.3342788

Förster, F., Romeo, M., Holthaus, P., Wood, L. J., Dondrup, C., Fischer, J. E., et al. (2023). Working with troubles and failures in conversation between humans and robots: workshop report. *Front. Robotics AI* 10, 1202306. doi:10.3389/frobt.2023.1202306

Foster, M. E., Craenen, B., Deshmukh, A., Lemon, O., Bastianelli, E., Dondrup, C., et al. (2019). Mummer: socially intelligent human-robot interaction in public spaces. doi:10.48550/ARXIV.1909.06749

Frijns, H. A., Schürer, O., and Koeszegi, S. T. (2023). Communication models in human–robot interaction: an asymmetric MODel of ALterity in human–robot interaction (AMODAL-HRI). *Int. J. Soc. Robotics* 15 (3), 473–500. doi:10.1007/s12369-021-00785-7

Frijns, H. A., Hirschmanner, M., Sienkiewicz, B., Hönig, P., Indurkhya, B., and Vincze, M. (2024). Human-in-the-loop error detection in an object organization task with a social robot. *Front. Robotics AI* 11, 1356827. doi:10.3389/frobt.2024.1356827

Gao, G., Yamashita, N., Hautasaari, A., Echenique, A., and Fussell, S. R. (2014). Effects of public vs. private automated transcripts on multiparty communication between native and non-native english speakers. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 843–852. doi:10.1145/2556288.2557303

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. (1967). *Studies in ethnomethodology*. Cambridge: Polity Press.

Garfinkel, H. (2002). *Ethnomethodology's program: working out durkheim's aphorism*. Lanham, MD: Rowman and Littlefield Publishers.

Gehle, R., Pitsch, K., Dankert, T., and Wrede, S. (2015). "Trouble-based group dynamics in real-world HRI—reactions on unexpected next moves of a museum guide robot," in *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)* (IEEE), 407–412.

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. (2015). Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Front. Psychol.* 6, 931. doi:10.3389/fpsyg.2015.00931

Goodwin, C. (1981). *Conversational organization: interaction between speakers and hearers*. New York: Irvington Publishers.

Gunson, N., García, D. H., Sieińska, W., Dondrup, C., and Lemon, O. (2022). "Developing a social conversational robot for the hospital waiting room," in *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)* (IEEE), 1352–1357.

Hayes, C. J., Moosaei, M., and Riek, L. D. (2016). "Exploring implicit human responses to robot mistakes in a learning from demonstration task," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (IEEE), 246–252.

Heath, C., Hindmarsh, J., and Luff, P. (2022). Video in qualitative research: analysing social interaction in everyday life. 55 city road, London. Available online at: https://methods.sagepub.com/book/video-in-qualitative-research.

Heritage, J. (1985). "A change-of-state token and aspects of its sequential placement," in *Structures of social action. Studies in emotion and social interaction*. Editor J. M. Atkinson (Cambridge: Cambridge University Press), 299–345. Available online at: https://www.cambridge.org/core/books/structures-of-social-action/changeofstate-token-and-aspects-of-its-sequential-placement/8AACF1699E4A7AE1B19D8248B6613405.

Heritage, J. (2007). "Intersubjectivity and progressivity in person (and place) reference," in *Person reference in interaction: Linguistic, cultural and social perspectives. Language culture and cognition*. Editors N. J. Enfield, and T. Stivers (Cambridge: Cambridge University Press), 255–280. Available online at: https://www.cambridge.org/core/books/person-reference-in-interaction/intersubjectivity-and-progressivity-in-person-and-place-reference/3BB0AC6C51893651C687063B0AE2EB5E.

Holler, J., and Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Front. Psychol.* 6, 98. doi:10.3389/fpsyg.2015.00098

Holthaus, P., and Wachsmuth, S. (2021). It was a pleasure meeting you: towards a holistic model of human–robot encounters. *Int. J. Soc. Robotics* 13 (7), 1729–1745. doi:10.1007/s12369-021-00759-9

Honig, S., and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: literature review and model development. *Front. Psychol.* 9, 861. doi:10.3389/fpsyg.2018.00861

Hopper, R. (2005). "A cognitive agnostic in conversation analysis: when do strategies affect spoken interaction?," in *Conversation and cognition*. Editors H. te Molder, and J. Potter, 1st edn (Cambridge University Press), 134–158. Available online at: https://www.cambridge.org/core/product/identifier/CBO9780511489990A015/type/book_part (Accessed January 17, 2022).

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology* 35 (2), 441–456. doi:10.1177/s0038038501000219

Irfan, B., Miniota, J., Thunberg, S., Lagerstedt, E., Kuoppamäki, S., Skantze, G., et al. (2025). Human-robot interaction conversational user enjoyment scale (HRI CUES). *IEEE Trans. Affect. Comput.* 1–17. doi:10.1109/TAFFC.2025.3590359

Irfan, B., Kuoppamäki, S., and Skantze, G. (2024). Recommendations for designing conversational companion robots with older adults through foundation models. *Front. Robotics AI* 11, 1363713. doi:10.3389/frobt.2024.1363713

Ito, A., and Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: comparison of different analysis methods. *Behav. Res. Methods* 55 (7), 3461–3493. doi:10.3758/s13428-022-01969-3

Ivarsson, J., and Lindwall, O. (2023). Suspicious minds: the problem of trust and conversational agents. *Comput. Support. Coop. Work (CSCW)* 32 (3), 545–571. doi:10.1007/s10606-023-09465-8

Jefferson, G. (2004). "Glossary of transcript symbols with an introduction," in *Conversation analysis: studies from the first generation* (John Benjamins).

Jefferson, G. (2017). *Repairing the broken surface of talk: managing problems in speaking, hearing, and understanding in conversation*. Oxford University Press. Available

online at: https://global.oup.com/academic/product/repairing-the-broken-surface-of-talk-9780190697969?q=Gail_Jefferson&lang=en&cc=nl.

Kendrick, K. H., Brown, P., Dingemanse, M., Floyd, S., Gipper, S., Hayano, K., et al. (2020). Sequence organization: a universal infrastructure for social action. *J. Pragmat.* 168, 119–138. doi: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*. United Kingdom: Ulster University. Available online at: https://ulster-staging.pure.elsevier.com/ws/portalfiles/portal/12689965/2018KerrisonATPhD.pdf

Klein Selle, N., Gamer, M., and Pertzov, Y. (2021). Gaze-pattern similarity at encoding may interfere with future memory. *Sci. Rep.* 11 (1), 7697. doi:10.1038/s41598-021-87258-z

Kong, X., Xing, Y., Tsourdos, A., Wang, Z., Guo, W., Perrusquia, A., et al. (2024). Explainable interface for human-autonomy teaming: a survey. doi:10.48550/ARXIV.2405.02583

Kristiansen, E. D., and Rasmussen, G. (2021). Eye-tracking recordings as data in EMCA studies: exploring possibilities and limitations. *Soc. Interact. Video-Based Stud. Hum. Sociality* 4. doi:10.7146/si.v4i4.121776

Krummheuer, A. L. (2015). "Users, bystanders and agents: participation roles in human-agent interaction," in *Human-computer interaction - INTERACT 2015* (Springer International Publishing).

Lee, M., Ruijten, P., Frank, L., and IJsselsteijn, W. (2023). "Here's looking at you, robot: the transparency conundrum in HRI," in *2023 32nd IEEE international conference on robot and human interactive communication (RO-MAN), Busan, Korea, republic of, 28 August 2023* (IEEE), 2120–2127. doi:10.1109/RO-MAN57019.2023.10309653

Lerner, G. H., and Kitzinger, C. (2019). Well-prefacing in the organization of self-initiated repair. *Res. Lang. Soc. Interact.* 52 (1), 1–19. doi:10.1080/08351813.2019.1572376

Levinson, S. C. (1987). "Minimization and conversational inference," in *Pragmatics and beyond companion series*. Editors J. Verschueren, and M. Bertuccelli Papi (Amsterdam: John Benjamins Publishing Company), 61. Available online at: https://benjamins.com/catalog/pbcs.5.10lev (Accessed October 2, 2023).

Levinson, S. C. (2012). "Action formation and ascription," in *The handbook of conversation analysis*, 101–130. doi:10.1002/9781118325001.ch6

Liberman, K. (1980). Ambiguity and gratuitous concurrence in inter-cultural communication. *Hum. Stud.* 3 (1), 65–85. doi:10.1007/bf02331801

Linell, P. (1995). "Troubles with mutualities: towards a dialogical theory of misunderstanding and miscommunication," in *Mutualities in dialogue 176*. Cambridge: Cambridge University Press, 213.

Linell, P. (1998). *Approaching dialogue: talk, interaction and contexts in dialogical perspectives*. John Benjamins Publishing.

Linell, P., and Lindström, J. (2016). Partial intersubjectivity and sufficient understandings for current practical purposes: on a specialized practice in Swedish conversation. *Nordic J. Linguistics* 39 (2), 113–133. doi:10.1017/s03325865 16000081

Loth, S., Jettka, K., Giuliani, M., and de Ruiter, J. P. (2015). Ghost-in-the-Machine reveals human social signals for human–robot interaction. *Front. Psychol.* 6, 1641. doi:10.3389/fpsyg.2015.01641

Luger, E., and Sellen, A. (2016). "Like having a really bad PA': the gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, (San Jose, CA: ACM), 5286–5297. doi:10.1145/2858036.2858288

Lynch, M. (1995). The idylls of the academy. *Soc. Stud. Sci.* 25 (3), 582–600. doi:10.1177/030631295025003008

Mair, M., Brooker, P., Dutton, W., and Sormani, P. (2021). Just what are we doing when we're describing AI? Harvey sacks, the commentator machine, and the descriptive politics of the new artificial intelligence. *Qual. Res.* 21 (3), 341–359. doi:10.1177/1468794120975988

Malle, B. F., and Knobe, J. (1997). The folk concept of intentionality. *J. Exp. Soc. Psychol.* 33 (2), 101–121. doi:10.1006/jesp.1996.1314

Maynard, D. W. (2006). Cognition on the ground. *Discourse Stud.* 8, 105–115. doi:10.1177/1461445606059560

McCarthy, J., and Hayes, P. J. (1981). "Some philosophical problems from the standpoint of artificial intelligence," in *Readings in artificial intelligence* (Elsevier), 431–450.

Mellmann, H., Arbuzova, P., Kontogiorgos, D., Yordanova, M., Haensel, J. X., Hafner, V. V., et al. (2024). "Effects of transparency in humanoid robots - a pilot study," in *Companion of the 2024 ACM/IEEE international conference on human-robot interaction* (Boulder,CO: ACM), 750–754. doi:10.1145/3610978.3640613

Meyer, C. (2019). Ethnomethodology's culture. *Hum. Stud.* 42 (2), 281–303. doi:10.1007/s10746-019-09515-5

Minsky, M. (1988). *Society of mind*. Simon & Schuster.

Mirnig, N., Giuliani, M., Stollnberger, G., Stadler, S., Buchner, R., and Tscheligi, M. (2015). "Impact of robot actions on social signals and reaction times in HRI error situations," in *Social robotics: 7th international conference, ICSR 2015, paris, France, October 26-30, 2015, proceedings 7, 2015* (Springer), 461–471.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. (2017). To err is robot: how humans assess and act toward an erroneous social robot. *Front. Robotics AI* 4, 21. doi:10.3389/frobt.2017.00021

Mitchell, M. (2021). "Why AI is harder than we think," in *Proceedings of the genetic and evolutionary computation conference* (New York, NY, USA: GECCO '21. Association for Computing Machinery). doi:10.1145/3449639.3465421

Mondada, L. (2011). Understanding as an embodied, situated and sequential achievement in interaction. *J. Pragmat.* 43 (2), 542–552. doi:10.1016/j.pragma.2010.08.019

Mondada, L. (2016). Challenges of multimodality: language and the body in social interaction. *J. Sociolinguistics* 20 (3), 336–366. doi:10.1111/josl.1_12177

Mondada, L. (2017). Walking and talking together: questions/answers and mobile participation in guided visits. *Soc. Sci. Inf.* 56 (2), 220–253. doi:10.1177/0539018417694777

Mondada, L. (2018). Multiple temporalities of language and body in interaction: challenges for transcribing multimodality. *Res. Lang. Soc. Interact.* 51, 85–106. doi:10.1080/08351813.2018.1413878

Mondada, L. (2019). Transcribing silent actions: a multimodal approach of sequence organization. *Soc. Interact. Video-Based Stud. Hum. Sociality* 2 (1). doi:10.7146/si.v2i1.113150

Mondada, L., Bänninger, J., Bouaouina, S. A., Camus, L., Gauthier, G., Hänggi, P., et al. (2020). Human sociality in the times of the Covid-19 pandemic: a systematic examination of change in greetings. *J. Sociolinguistics* 24 (4), 441–468. doi:10.1111/josl.12433

Mondémé, C. (2022). Why study turn-taking sequences in interspecies interactions? *J. Theory Soc. Behav.* 52 (1), 67–85. doi:10.1111/jtsb.12295

Nagel, T. (1989). *The view from nowhere*. Oxford University Press.

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

Orenes, I., García-Madruga, J. A., Gómez-Veiga, I., Espino, O., and Byrne, R. M. J. (2019). The comprehension of counterfactual conditionals: evidence from eye-tracking in the visual world paradigm. *Front. Psychol.* 10, 1172. doi:10.3389/fpsyg.2019.01172

Pelikan, H., and Broth, M. (2016). "Why that nao?: how humans adapt to a conventional humanoid robot in taking turns-at-talk," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, (San Jose, CA: ACM), 4921–4932. doi:10.1145/2858036.2858478

Pelikan, H., Broth, M., and 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* (New York, NY, USA: Association for Computing Machinery), 461–470. doi:10.1145/3319502.3374814

Pelikan, H., Broth, M., and Keevallik, L. (2022). When a robot comes to life: the interactional achievement of agency as a transient phenomenon. *Soc. Interact. Video-Based Stud. Hum. Sociality* 5 (3). doi:10.7146/si.v5i3.129915

Peyrot, M. (1982). Understanding ethnomethodology: a remedy for some common misconceptions. *Hum. Stud.* 5 (1), 261–283. doi:10.1007/bf02127681

Pitsch, K. (2020). Answering a robot's questions: participation dynamics of adult-child-groups in encounters with a museum guide robot. *Réseaux* 220–221, 113–150. doi:10.3917/res.220.0113

Pitsch, K., Vollmer, A.-L., and 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. *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.* 14 (2), 268–296. doi:10.1075/is.14.2.06pit

Pollner, M. (1974). "Sociological and common sense models of the labelling process," in *Ethnomethodology: selected readings* (Penguin).

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, (New York, NY: Association for Computing Machinery), 1–12. doi:10.1145/3173574.3174214

Psathas, G. (1989). *Phenomenology and sociology: theory and research*. Boston: University Press of America.

Ragni, M., Rudenko, A., Kuhnert, B., and Arras, K. O. (2016). "Errare humanum est: erroneous robots in human-robot interaction," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (New York, NY:IEEE), 501–506. doi:10.1109/ROMAN.2016.7745164

Rasmussen, G. (2019). "Emergentism," in *The SAGE encyclopedia of human communication sciences and disorders damico JS and ball MJ*. Thousand Oaks: SAGE Publications.

Relieu, M., Sahin, M., and Francillon, A. (2020). Une approche configurationnelle des leurres conversationnels. *Réseaux N°* (2), 81–111. doi:10.3917/res.220.0081

Rossi, A., and Rossi, S. (2024). "On the way to a transparent HRI," in *Adjunct proceedings of the 32nd ACM conference on user modeling, adaptation and personalization, cagliari Italy* (New York, NY: ACM), 215–219. doi:10.1145/3631700.3664890

Rossi, A., Dautenhahn, K., Koay, K. L., and Walters, M. L. (2017). "How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario," in *Social robotics*. Editors A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, et al. (Cham: Springer International Publishing), 42–52. doi:10.1007/978-3-319-70022-9_5

Rudaz, D. (2024). How do robots become "social robots"? An empirical specification of the (Non-)Emergence of robots as social agents. *Sociol. Inst. Polytech. Paris,alaiseau*. Available online at: https://theses.hal.science/tel-04884570v1.

Rudaz, D., and Licoppe, C. (2024). "Playing the Robot's Advocate": bystanders' descriptions of a robot's conduct in public settings. *Discourse Commun.* 18 (4), 869–881. doi:10.1177/17504813241271481

Rudaz, D., Tatarian, K., Stower, R., and 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), 1–31. doi:10.1145/3575806

Ryle, G. (1949). *The concept of mind*. Chicago, IL: University of Chicago Press.

Sacks, H. (1995). "Lectures on conversation: volumes I and II (ed. G Jefferson). 1. publ," in *One paperback volume* (Oxford: Blackwell).

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013). To err is human(-like): effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robotics* 5 (3), 313–323. doi:10.1007/s12369-013-0196-9

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015). Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. *Proc. tenth Annu. ACM/IEEE Int. Conf. human-robot Interact.* 2015, 141–148. doi:10.1145/2696454.2696497

Schegloff, E. A. (1992). Repair after next turn: the last structurally provided defense of intersubjectivity in conversation. *Am. J. Sociol.* 97, 1295–1345. doi:10.1086/229903

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Res. Lang. and Soc. Interact.* 26 (1), 99–128. doi:10.1207/s15327973rlsi2601_5

Schegloff, E. A. (2007). *Sequence organization in interaction: a primer in conversation analysis*. Cambridge: Cambridge University Press. Available online at: http://ebooks.cambridge.org/ref/id/CBO9780511791208 (Accessed January 18, 2022).

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Lan.* 53, 361–382. doi:10.1353/lan.1977.0041

Schütte, N., Mac Namee, B., and Kelleher, J. (2017). Robot perception errors and human resolution strategies in situated human–robot dialogue. *Adv. Robot.* 31 (5), 243–257. doi:10.1080/01691864.2016.1268973

Schutz, A. (1972). "Common-sense and scientific interpretation of human action," in *Collected papers I: the problem of social reality*. Editor M. Natanson (Dordrecht: Springer Netherlands), 3–47. doi:10.1007/978-94-010-2851-6_1

Searle, J. R. (1969). *Speech acts: an essay in the philosophy of language*. Cambridge: Cambridge University Press. Available online at: https://www.cambridge.org/core/books/speech-acts/D2D7B03E472C8A390ED60B86E08640E7.

Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). "No fair!! an interaction with a cheating robot," in *2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (IEEE), 219–226. doi:10.1109/hri.2010.5453193

Shotter, J. (1996). Now I can go on:' Wittgenstein and our embodied embeddedness in the 'Hurly-Burly' of life. *Hum. Stud.* 19 (4), 385–407. doi:10.1007/bf00188850

Sterponi, L., and Fasulo, A. (2010). "How to Go on": intersubjectivity and progressivity in the communication of a child with autism. *Ethos* 38, 116–142. doi:10.1111/j.1548-1352.2009.01084.x

Stivers, T., Rossi, G., and Chalfoun, A. (2023). Ambiguities in action ascription. *Soc. Forces* 101 (3), 1552–1579. doi:10.1093/sf/soac021

Straten, C. L. V., Peter, J., Kühne, R., and Barco, A. (2020). Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. *ACM Trans. Human-Robot Interact.* 9 (2), 1–22. doi:10.1145/3365668

Stukenbrock, A., and Dao, A. N. (2019). "Joint attention in passing: what dual Mobile eye tracking reveals about gaze in coordinating embodied activities at a market," in *Embodied activities in face-to-face and mediated settings: social encounters in time and space*. Editors E. Reber, and C. Gerhardt (Cham: Springer International Publishing), 177–213. doi:10.1007/978-3-319-97325-8_6

Suchman, L. (1987). *Plans and situated actions: the problem of human-machine communication. Plans and situated actions: the problem of human-machine communication*. New York, NY, US: Cambridge University Press.

Svennevig, J. (2004). Other-repetition as display of hearing, understanding and emotional stance. *Discourse Stud.* 6, 489–516. doi:10.1177/1461445604046591

Ten Have, P. (2007). *Doing conversation analysis. 1 oliver's yard, 55 city road, London England EC1Y 1SP*. United Kingdom: SAGE Publications, Ltd. Available online at: https://methods.sagepub.com/book/doing-conversation-analysis (Accessed February 18, 2024).

Tian, L., and Oviatt, S. (2021). A taxonomy of social errors in human-robot interaction. *ACM Trans. Human-Robot Interact.* 10 (2), 1–32. doi:10.1145/3439720

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., et al. (2020). "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (Cambridge United Kingdom: ACM), 3–12. doi:10.1145/3319502.3374793

Tuncer, S., Gillet, S., and Leite, I. (2022). Robot-mediated inclusive processes in groups of children: from gaze aversion to mutual smiling gaze. *Front. Robotics AI* 9, 729146. doi:10.3389/frobt.2022.729146

Tuncer, S., Licoppe, C., Luff, P., and Heath, C. (2023). Recipient design in human–robot interaction: the emergent assessment of a robot's competence. *AI and Soc.* 39, 1795–1810. doi:10.1007/s00146-022-01608-7

Uchida, T., Kameo, N., and Ishiguro, H. (2024). Improving the closing sequences of interaction between human and robot through conversation analysis. *Sci. Rep.* 14 (1), 29554. doi:10.1038/s41598-024-81353-7

Venker, C. E., and Kover, S. T. (2015). An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *J. Speech, Lang. Hear. Res.* 58 (6), 1719–1732. doi:10.1044/2015_jslhr-l-14-0304

Verhagen, R. S., Neerincx, M. A., and Tielman, M. L. (2021). "A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable," in *International workshop on explainable, transparent autonomous agents and multi-agent systems, 2021* (Springer), 119–138.

Verhagen, R. S., Neerincx, M. A., and Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Front. Robotics AI* 9, 993997. doi:10.3389/frobt.2022.993997

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: survey of an emerging domain. *Image Vis. Comput.* 27 (12), 1743–1759. doi:10.1016/j.imavis.2008.11.007

vom Lehn, D. (2019). From garfinkel's 'Experiments in Miniature' to the ethnomethodological analysis of interaction. *Hum. Stud.* 42 (2), 305–326. doi:10.1007/s10746-019-09496-5

Wittgenstein, L. (1953). *Philosophical investigations*. Wiley-Blackwell.

Wortham, R. H., and Theodorou, A. (2017). Robot transparency, trust and utility. *Connect. Sci.* 29 (3), 242–248. doi:10.1080/09540091.2017.1313816

Wortham, R. H., Theodorou, A., and Bryson, J. J. (2016). "What does the robot think? Transparency as a fundamental design requirement for intelligent systems," in *IJCAI 2016 ethics for AI workshop*.

Yao, L., Pan, Y., and Jiang, D. (2011). "Effects of automated transcription delay on non-native speakers' comprehension in real-time computer-mediated communication," in *Proceedings of the 13th IFIP TC 13 international conference on human-computer interaction - volume part I* (Berlin, Heidelberg: Springer-Verlag), 207–214.

Zimmerman, D. H., and Pollner, M. (2013). *The everyday world as a phenomenon*' in *People and information: pergamon general psychology series 6* (Elsevier), 33.