# Signaling Human Intentions to Service Robots: Understanding the Use of Social Cues during In-Person Conversations

ANONYMOUS AUTHOR(S)

As social service robots become commonplace, it is essential for them to effectively interpret human signals, such as verbal, gesture, and eye gaze, when people need to focus on their primary tasks to minimize interruptions and distractions. Toward such a socially acceptable Human-Robot Interaction, we conducted a study (N=24) in an AR-simulated context of a coffee chat. Participants elicited social cues to signal intentions to an anthropomorphic, zoomorphic, grounded technical, or aerial technical robot waiter when they were speakers or listeners. Our findings reveal common patterns of social cues over intentions, the effects of robot morphology on social cue position and conversational role on social cue complexity, and users' rationale in choosing social cues. We offer insights into understanding social cues concerning perceptions of robots, cognitive load, and social context. Additionally, we discuss design considerations on approaching, social cue recognition, and response strategies for future service robots.

**ACM Reference Format:**
Anonymous Author(s). 2024. Signaling Human Intentions to Service Robots: Understanding the Use of Social Cues during In-Person Conversations. 1, 1 (December 2024), 31 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Service robots have been widely used in public spaces such as retail, healthcare, and hospitality [37, 39]. Within the human-robot interaction (HRI) community, there have been extensive discussions on how a social robot, as an autonomous and intelligent agent, should behave socially (e.g., [60, 86]). At the same time, it is equally important to understand how humans intuitively interact with and signal their intentions to robots in social settings [26]. Humans have many means to convey their intentions and apply them dynamically according to context. For example, when meeting potential employers over coffee, a person may want to ask a drink-serving robot to move aside to avoid potential interruption. Since they need to focus on the ongoing conversation, they might not have enough bandwidth to open the mobile app and send a command or feel embarrassed to speak aloud [55]. They may prefer simply gesturing to the robot in this situation. In other words, when conventional graphical user interfaces (GUI) and conversational interfaces (CUI) are not efficient, service robots should still be able to understand humans' intentions exhibited in other manners.

In human-human interactions (HHI), people often use different social cues to signal their intentions and emotions to minimize distractions and interruptions to the main tasks. These social cues come in different modalities, such as gaze, gesture, facial expression, body language, and vocal signals [10, 42]. For example, people often nod to signal awareness of a friend passing by and wave hands to turn down a floor cleaning service during a conversation. The same interaction preference also occurs in human-robot communications: users are likely to use natural and intuitive social cues to signal their intentions to robots, because this is part of human social communication [92] and individuals are inclined to treat robots as social actors [67]. Previous HHI research suggested that the choice of social cues can be affected by people's perceived relationship with the other parties at the scene and their occupancy by the main task [10]. Similarly, when interacting with service robots, humans' choices of social signals may be affected by the robot's morphology, social expectations, and norms [28].

Backed by the development of voice recognition and vision-based nonverbal cue detection algorithms, many HRI studies explored how humans interact with robots with social cues such as speech [17], gesture [16, 17, 27], gaze [51], posture [58], and movement of different body parts [65], to name a few. However, most existing works focused on primary interactions between a single user and a robot in lab settings rather than in uninterruptible social scenarios where human-robot interaction is a side task. In addition, some of them only analyzed a selected set of social signal modalities or examined different social cues with a specific robot form (e.g., wearable robot arms [65] or drones [17, 27]). Taken together, there is still a lack of a comprehensive understanding of humans' choice among diverse social cues to express their intentions to different forms of robots in social situations where they are engaging with other people as communicators or respondents.

In this paper, we explore how humans choose and combine different modalities of social cues to communicate with a service robot during an important social encounter. Specifically, we investigate how robot morphologies and roles in the primary activity influence the use of social cues to express intentions to the robot. To achieve this, we conducted an elicitation study where participants interacted with a robot server during a simulated coffee chat with potential employers (played by two actors). Participants were free to use any social cues they deemed intuitive and appropriate to convey 13 designated intentions (referents) representing common human feedback types during social interactions with robots. We selected four representative robots with distinct forms, based on the morphology taxonomy from [68]: anthropomorphic, zoomorphic, grounded technical, and aerial technical. To ensure stability and consistency, we simulated the robots as virtual prototypes [41, 65] using the augmented reality headset and controlled their actions through a Wizard of Oz (WoZ) approach. Each participant ($N$ = 24) completed two elicitation sessions, alternating between two conversation roles: speaker, introducing themselves and their experiences, and listener, observing the conversation but free to interject. Each session featured a different robot morphology (2 out of 4 per participant), with counterbalanced orders to mitigate learning and order effects. Following the elicitation sessions, we conducted retrospective think-aloud interviews to understand participants' mental models and decision-making processes. By integrating quantitative analysis of observed social cues with qualitative insights from interviews, we identified patterns in participants' preferences and rationales. Participants favored intuitive and context-appropriate cues, such as using eye gaze to signal awareness or hand gestures to guide actions. Robot morphology influenced their choices, particularly in gestures and verbal features, shaped by perceptions of the robot's sensory capabilities. These findings underscore participants' goals of minimizing conversational disruptions, ensuring clarity in communication, and maintaining politeness in professional social contexts.

In this paper, we make the following contributions:

- We conducted an elicitation study to explore how humans interact with robot waiters on the side during an important coffee chat with potential employers, and how the robot forms and users' conversation roles affect the choice of social cues to signal intentions.
- Through retrospective think-aloud and follow-up interviews, we analyze the rationales of participants' interactions and discuss factors that may influence their intentions when choosing social cues.
- Based on our findings, we further discuss the design implications for the human-robot interaction system in social settings and provide suggestions for future elicitation research for human-robot interactions.

## 2 Related Work

### 2.1 Service Robots in Social Settings

*2.1.1 Definitions and Scope.* Social settings are systems centered on *social processes* (i.e., interactions between two or more individuals) structured by *resources* and the *organization of resources* [87]. In this study, we refer to social encounters where human-human interaction dominates, such as a coffee chat, a group discussion, or a dinner party, as social processes, with social service robots acting as resources to support the social processes. Social robots, as defined by Yan et al. [96], are "robots which can execute designated tasks, and the necessary condition turning a robot into a social robot is the ability to interact with humans by adhering to certain social cues and rules". They share key features such as sensing and responding to environmental cues, interacting with humans (or other robots), and understanding and following social rules [79]. Yan et al. [96] also emphasize the importance of recognition capabilities and social cues for social service robots.

*2.1.2 Challenges in Designing Intuitive HRI in Social Settings.* Human-robot interaction (HRI) for social robots presents unique challenges that distinguish it from general HRI. The ability of social service robots to understand and respond to users' intentions and preferences has been considered critical to ensure safety, human control, and alignment with human expectations and preferences [81, 83]. Tian and Oviatt [85] present a comprehensive taxonomy of social errors in HRI, illustrating the complexity of social interactions between humans and robots. This complexity extends beyond mere functionality to encompass aspects such as timing, appropriateness, and emotional congruence. Key considerations in social robot interactions include social appropriateness and adaptability [60, 86], emotional and affective aspects [44], temporal dynamics [82], error handling and repair [49], and context-specific behaviors [48]. While prior work has investigated how robots perceive and respond to human behavior and highlighted the importance of properly understanding humans in social HRI, there is a limited understanding of how humans intuitively interact with service robots in social encounters. In the social settings of our study, robots are positioned in a side task, providing services to the participants rather than central actors among humans. We examine how humans naturally communicate with robots in such contexts, aiming to better interpret human signals, inform the design of social service robots, and provide intuitive services to human-centric social encounters.

## 2.2 Social Cues in Human-Robot Interaction

*2.2.1 Human Social Cues.* Social Signal Processing (SSP) analyzes the social behaviors and cues in both Human-Human (HHI) and Human-Computer interactions (HCI) contexts [92]. Vinciarelli et al. [92, 93] provide a comprehensive summary of human nonverbal behavior cues, their functions and social behavior modeling. Multimodal analysis of nonverbal behaviors in social interactions includes applications in modeling multimodal behaviors for face-to-face social interaction [62], automatic categorization of autism spectrum disorder [18], and classifying perceptions of interdependence [24]. The relationship between human social cues and emotions is also widely explored [31, 52]. However, much of the work in SSP has focused on the analysis and processing of human behaviors, with a limited understanding of how humans select and adapt social cues when interacting with robots in dynamic social settings. While social cue processing for robots has advanced, focusing on how robots express their own social intentions through gaze [22, 63], speech [22], gestures [47, 57], and facial expressions [14, 57], less attention has been paid to how humans intuitively communicate their intentions to social service robots, particularly in complex, real-world settings where robots play a peripheral role. Our study addresses this gap by an exploratory study focusing on the variety of social cues humans employ to convey different intentions in the interactions with service robots during social encounters.

*2.2.2 Social Cues Elicitation in Human-Robot Interaction.* Research in HCI has been exploring how humans interact with robots with different social cues in various systems, such as public displays [75], smart rings [30], and chairs [11].

Villarreal-Narvaez et al. [90] systematically reviewed the literature on gesture elicitation studies in HCI. Furthermore, studies on HRI have also examined the use of different social cues to interact with robots. Cauchard et al. [17] explored how humans interact with drones with gesture and speech, and Firestone et al. [27] collected elicitation of gestures for small Unmanned Aerial Systems (sUAS) to understand and model human-drone interactions. Canuto et al. [16] proposed a frustration-based elicitation approach and studied the intuitiveness of human gestures to signal robots with basic commands. However, most of these studies have focused on primary interactions in controlled, lab-based settings between a single user and a robot, and have neglected more complex, uninterrupted social scenarios where HRI occurs as a side task. Furthermore, these studies limit their focus to a narrow range of social signal modalities or specific robot forms (e.g., , drones [17, 27] or wearable robot arms [65]), failing to provide a comprehensive understanding of the variety of social cues employed in dynamic, real-world HRI scenarios. We aim to explore human interactions with multiple forms of service robots when occupied by primary social encounters, to elicit social cues that cover all possible modalities from human bodies, and to understand the choice of social cues to interact with service robots as a peripheral task, expanding the scope of social cue elicitation research in HRI.

## 2.3 Morphology of Social Service Robots

*2.3.1 Classification of Robot Morphology.* Robot morphology is one of the fundamental classification parameters in HRI research [28, 68, 97]. As refined by Onnasch and Roesler [68], robot morphology can be classified as *anthropomorphic* (human-like and androids), *zoomorphic* (animal-like), or *technical* (machine-like). Robot morphology determines a robot's physical embodiment and influences users' perceptions of its functional and communicative capabilities [68]. Eyssel [26] provides an experimental psychological perspective on social robotics, emphasizing how human cognitive and social psychological processes influence perceptions of and interactions with social robots. In our experiment, we apply Onnasch and Roesler [68]'s taxonomy of robot morphology, aiming to investigate how the visual design of a robot influences human perceptions of its embodiment and capabilities and human usages of social cues during interactions with service robots in social settings.

*2.3.2 Morphologies of Social Service Robots.* Social robots have been widely applied in service industries [37, 39, 60] with various morphologies, including *anthropomorphic* [9, 20, 50], *zoomorphic* [5, 23, 38] and *technical* [2, 3, 6, 8] robots. The *technical* morphology of service robots, i.e., product-oriented robots, is most commonly seen among the industrial applications of service robots, such as drones [6, 8, 45], cleaning robots [4], and delivery robots [3, 95]. Comparing with the *technical* morphology, research has been exploring how *anthropomorphic* robots affect users' perceptions and interactions with social robots. Kwak [46] compare the social presence and sociability of a human-oriented robot and a product-oriented robot. Stroessner and Benitez [84] examine the effects of gendered and machine-like features on the social perception of humanoid and non-humanoid robots. Most research of *zoomorphic* social service robots has been focused more on companion and guidance applications [23, 36, 56]. Hauser et al. [33] explore human perceptions of incidental encounters with service robot dogs in a lab setting, suggesting a positive experience for the human. Existing research on three morphologies of social service robots has been limited to academic settings, with few empirical results on the effects of robot morphology on human interactions with robots, especially in real-world social settings. Thus, we add robot morphology as an independent variable in our study to explore the effects of robot morphology on human interactions with social service robots.

## 2.4 Exploratory Prototyping in Human-Robot Interaction

Exploratory prototyping plays a crucial role in HRI research, enabling researchers to rapidly test and assess the feasibility and usability of experimental robot designs [98]. Research by Rojas et al. [76] indicates that virtual and physical robot prototypes are comparable in observable aspects (e.g., color and shape) and emotional responses, as measured by the Self-Assessment Manikin instrument, but differ in social perceptions, such as discomfort and warmth. Beyond image and video prototyping, virtual reality (VR) has been adopted as a simulation tool for social robots, offering immersive testing environments [77, 80]. For instance, Kamide et al. [41] indicates that while VR robots and physical robots may elicit differing subjective impressions, participants' behavior and desired personal space remain consistent between the two prototyping approaches. Additionally, studies have demonstrated the feasibility of using augmented reality (AR) to collect human social cues in real-world interaction contexts [35, 53, 70]. Our primary objective is to gather participants' social cues—specifically, their behaviors when interacting with service robots. Given that virtual robot prototypes effectively replicate key aspects of interaction observed with physical robots [41], the use of augmented reality for prototyping is a practical and effective choice for our study.

## 3 Method

In our study, we simulated a coffee chat scenario to serve as a representative social encounter, incorporating 4 robot morphologies and 13 referents relevant to the interaction context. A pilot study was conducted to refine the study protocol for clarity and practical execution. Following this, we carried out an elicitation study to investigate human social cues when interacting with robots in a social context. Behavioral and interview data were collected during the study, subsequently transcribed and coded for statistical and thematic analyses. All experiments received approval from the Institutional Review Board (IRB) approval from the University Research Ethics Committee.

### 3.1 Simulation of Robots and Public Context Settings

*3.1.1 Social Encounter.* Our study examines human interaction with the robot as a side task, positioning the social encounter, i.e., human-human interaction, as the primary task to be interleaved by robots in our social settings. Conversation has been a central topic in the analysis of human-human interactions [32, 72, 91]. Thus, we choose the conversation as the main task of our experiment settings. To select an appropriate social scenario for our experiment, we have the following requirements: First, it should involve the participation of multiple people. Second, interruption is unwanted and should be minimized in such social scenarios, but all communication channels should not be fully occupied, allowing users to maintain sufficient options when utilizing social signals. Based on the above considerations, five researchers brainstormed appropriate scenarios for our experiment and selected a coffee chat with potential employers. Considering that users' choice of social cues when interrupted by a robot may vary based on their ongoing activity, particularly their conversational role as listener or speaker, we included these roles as an independent variable ConversationRole in our experimental design. The ConversationRole is simulated by letting the participants hold the conversational flow (*speaker*), or mainly listen to the conversation led by the potential employers (*listener*). We recruited actors to play the roles of two potential employers (advisors) for future jobs, internships, or advanced study according to the participant's status and goals. The actors were trained to pay attention to the participants and create more chances for the participants to continue their speech when participants were in the *speaker* session, and to take the dominance of the conversation when participants were in the *listener* session. During the coffee chat, natural turns of conversation and attention were maintained to ensure the authenticity of the social encounter.

*3.1.2 Robot.* We also set the robot morphology as our independent variable, denoted as RobotMorphology. To select the robot morphologies, we initially adopted the taxonomy proposed by [68], which classifies robot morphologies

into three categories: *technical*, *anthropomorphic*, and *zoomorphic*. Given that the height of a robot may influence user interaction [34, 74] and perception [40], and considering the prevalence of drones [6–8] and food delivery carts [1–3] in catering services, we further divided technical robots into two subcategories: aerial robots (drones) and grounded robots (food delivery carts). Thus, the independent variable RobotMorphology had four experiment groups: *aerial technical*, *grounded technical*, *anthropomorphic*, and *zoomorphic* robots. Different morphologies of robots were simulated using Augmented Reality (AR) technology to provide participants with an immersive and interactive experience while maintaining flexibility in testing different robot morphologies. As discussed in Section 2.3.2, humans behave similarly toward virtual and real robots, making AR simulation a suitable choice for prototype experimental settings. This approach provides valuable insights into understanding users' behavior and informing robot design. Virtual simulation also ensures consistency across conditions, reduces logistical complexity, and allows for convenient human social cue collection. To identify representative robots, we first compiled a diverse set of examples for each morphology. Five researchers independently voted for up to three robots per category based on the following criteria: (1) popularity, (2) ability to serve drinks, (3) perceived safety for novice users, and (4) representativeness of the morphology. Based on the voting results, we selected Pepper[1] as the anthropomorphic representative and Spot[2] as the zoomorphic representative. For the technical robots, we created a grounded food delivery cart mesh commonly seen in coffee shops and selected a drone equipped with safety measures from the Unity asset store[3]. Figure 1 shows the four robots used in our experiment.


Fig. 1. The Four Forms of Robots Used in Our Experiment: (a) Anthropomorphic Robot; (b) Zoomorphic Robot; (c) Grounded Technical Robot; (d) Aerial Technical Robot.

*3.1.3 Referents.* We aimed to design effective elicitation referents for human-robot interactions in social encounters by first categorizing the types of necessary interactions between humans and robots. A literature survey was conducted to summarize the situations where robots require human interaction [61, 83], providing a structured foundation for the elicitation referents. Based on this review, five necessary interaction types were identified, as shown in Table 1. In our coffee chat scenario, we brainstormed and carefully designed 13 elicitation referents, covering each of the identified interaction types. These referents were designed to reflect common interactions between human waiters and customers and to address the important tasks of a robot waiter as outlined in [29]. Table 2 details the elicitation referents, their associated interaction situations, and their classification by interaction type. To create a realistic social encounter, we situated our study in a shared public workspace with low-volume background music and ambient noise, conditions typical of coffee chats. This setting mirrors natural environments where human-human social encounters

[1] https://us.softbankrobotics.com/pepper
[2] https://bostondynamics.com/products/spot/
[3] https://assetstore.unity.com/packages/3d/vehicles/air/simple-drone-190684

Table 1. Five Necessary Interaction Types for HRI in Social Settings

| | | |
|---|---|---|
| Robot Active Seeking for Human Input | i. When the robot is not sure | |
| | ii. When the robot asks for evaluation | |
| Robot Passive Receiving Human Input | iii. When the human signals awareness | |
| | When the robot has an error [85] | iv. Performance error |
| | | v. Social error |

Table 2. 13 Referents and Their Types Used in Our Experiment

| **Referents** | **Situation** | **Type** |
|---|---|---|
| *Signal Awareness* | The robot was moving to you. | iii |
| *Signal the Robot to Serve Partner* | The robot stopped near you, carrying a bottle of drink you had just ordered. | v |
| *Do Not Provide Feedback* | The robot sent the drink to your partner, and asked "Please rate my service." | ii |
| *Interrupt* | The robot was saying "Ok, your next cup of drink is expected to arrive at ..." | v |
| *Signal Wrong Drink* | The robot went away then came back, but brought a wrong drink to you. | iv |
| *Signal Emergency Stop* | The robot encountered a malfunction and was rushing towards you. | iv |
| *Indicate Drink Position* | The robot went near you and asked "Where should I place the drink?" | i |
| *Signal to Prevent Drink Spilling* | The robot was sending out the drink, but the drink was spilling. | iv |
| *Provide Bad Feedback* | The robot finished sending the drink and asked "Please rate my service." | ii |
| *Dismiss* | The robot was wandering around nearby, disturbing your conversation. | v |
| *Call Over* | Your partner just finished her drink, and wanted the robot to collect the cup. | iii |
| *Signal the Robot to Collect Cup* | The robot moved to you and asked "How may I help you?" | i |
| *Provide Good Feedback* | The robot collected your partner's cup, and asked "Please rate my service." | ii |

and interactions with service robots frequently occur, ensuring ecological validity. We programmed the robots to response to each interaction situation with appropriate actions and sounds in the AR headset for participants. The robots' actions were controlled using a Wizard of Oz (WoZ) approach, enabling a human operator (wizard) sitting at another table behind the participants to manage the robots' responses in real-time, thereby simulating seamless and context-appropriate interactions. And since our actors were familiar with the robots' routines and actions, they could infer the progress of the virtual robot and act as if they could see the robots during the referent elicitation, which helped to maintain the authenticity of the AR simulation of the scenario.

*3.1.4 Pilot Study.* To validate our design, we conducted a pilot study (N=7) where participants were asked to use social cues to signal the AR-simulated robot in a coffee chat with potential employers under different combinations of conditions: 4 robot morphologies (*aerial technical*, *grounded technical*, *anthropomorphic*, *zoomorphic*) × 2 conversation roles (*speaker*, *listener*) × 2 postural configurations (*sitting around a table*, *standing around a long bar*). Each participant completed four sessions, which included two robot forms, both roles, and both postural configurations. For each referent, participants were required to elicit three times and make each elicitation as distinct from the others as possible.

According to the post-experiment interview, we improved our study design in the following aspects: First, all participants found the requirement of eliciting three times during conversation to be too cognitively demanding. Thus, we reduced the three compulsory social cues to one, and participants were free to provide alternatives either during the elicitation or in the interview. Second, when asked about the differences between sitting and standing, all participants believed that the effects were much smaller than those caused by roles, except for the emergency case (referent *Signal Emergency Stop*). Considering that sitting is more common for a coffee chat scenario and for the participants' comfort, we only kept the sitting setting. Finally, we improved the AR simulation experience according to the participants' feedback, such as more authentic robot sound effects, approaching directions, and interaction distances.

### 3.2 Elicitation Study

*3.2.1 Participants.* We recruited 24 participants (7 females and 17 males) aged between 19 to 33 ($M = 24.08, SD = 2.7$). Most of them are students from various majors at local universities, which made them more relatable as future position seekers in our experiment. We randomly assigned each participant to two robot groups, with 12 participants in each group. The participants' familiarity to the four robot morphologies are shown in Table 3. We also collected data on the participants' dominant hand, with 23 being right-handed and only one left-handed. All participants were compensated at a rate of 12$ per hour.

Table 3. Participants' familiarity to the four robot morphologies used in our experiment.

| | *aerial technical* | *anthropomorphic* | *grounded technical* | *zoomorphic* |
|---|---|---|---|---|
| **Never heard of** | 0 (0.00%) | 1 (8.33%) | 0 (0.00%) | 3 (25.00%) |
| **Heard of but never interacted with** | 3 (25.00%) | 8 (66.67%) | 4 (33.33%) | 8 (66.67%) |
| **Interacted with** | 9 (75.00%) | 3 (25.00%) | 8 (66.67%) | 1 (8.33%) |

*3.2.2 Experiment Settings.* In our elicitation study, the participants were asked to engage in a coffee chat with two potential employers, with a robot waiter serving around. The chat was conducted in shared public workspace where people could occasionally pass by. The robots were simulated in the Augmented Reality (AR) of Quest Pro[4] using Unity[5] and controlled with Wizard of Oz (WoZ). We used Quest Pro's built-in cameras to capture participants' eye gazes, facial expressions, and body poses, and recorded a first-view video from Unity to visualize the eye gaze. We also set up another external camera to capture participants' whole bodies so that we can collect other modalities of their social cues. Two actors played the roles of the potential employers (advisors) for future jobs, internships, or advanced study according to the participants' status and goals. To ensure the authenticity of the AR simulation, the actors relied on the sound effects to determine the ongoing task and the robot's state and acted accordingly to pretend that they could see the robot. Figure 2 shows an illustration of the experiment and screenshots of Quest Pro Recordings from the participant's viewpoint.

*3.2.3 Design.* We proposed a mixed-design study, incorporating two within-subjects variables, ConversationRole (nominal, two levels: *speaker* and *listener*, as described in Section 3.1.1) and Referent (nominal, 13 levels, detailed in Table 2), and one mixed-design variable, RobotMorphology (nominal, four levels: *aerial technical*, *grounded technical*,

[4] https://www.meta.com/quest/quest-pro/
[5] version 2022.3.32f1, https://unity.com/

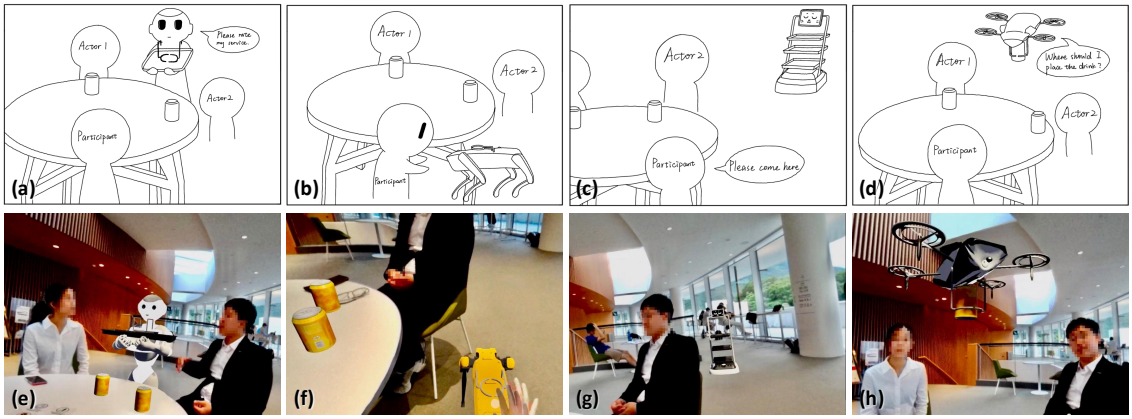


Fig. 2. An illustration of the experiment with four different robots. (a) Pepper, the *anthropomorphic* robot, was asking the participant to provide feedback. (b) The participant was asking Spot, the *zoomorphic* robot, to go away. (c) The participant was asking the *grounded technical* robot to come over. (d) The *aerial technical* robot carried the second cup of drink to the participant and asked him/her where to place it. (e-h) Images captured from the participants' viewpoint during their interaction with each of the above robots.

*anthropomorphic*, *zoomorphic*, as introduced in Section 3.1.2). During the coffee chat, participants engaged in both conversational roles across two sessions: 1) the *speaker*, who was asked to introduce themselves and their experiences to the potential employers; 2) the *listener*, who mainly listened to the conversation between the two potential employers, but could also cut in to ask questions or make comments. Across these two sessions, participants interacted with two out of four robot morphologies (*aerial technical*, *grounded technical*, *anthropomorphic*, *zoomorphic*), which acted as waiters in the study scenario. Participants elicited interactions for all 13 referents described in Section 3.1.3. Each ConversationRole group included 24 participants, while each RobotMorphology group comprised 12 participants. This resulted in six participants for each combination of ConversationRole × RobotMorphology. To reduce possible learning and order effects, the combination of robot and role, together with their order of presentation was counterbalanced.

*3.2.4 Procedure.* Upon arrival, participants first filled out a demographic questionnaire and signed a consent form. Then, they were briefed on the requirements and the whole flow of the experiment: Participants were instructed to utilize any social cues they deemed natural and appropriate for conveying a set of referents to the robot while maintaining their attention on the conversation as much as possible; Additionally, participants were asked to envision future robot waiters and assume that these robots are capable of identifying any modality of their social cues. For each referent, participants were encouraged to elicit as many ways as possible, but this was not forced to ensure that they were not overloaded and to guarantee the naturalness of the elicited cues. After confirming that the participants had a clear understanding of the task, participants were instructed to put on the Quest Pro headset, and calibrate the eye-tracking function within the headset. The participants were given preliminary exposure to all the referents through a robot demonstration before commencing the formal experiment to make them more familiar with the referents and avoid omissions or misinterpretations. The participants then elicited social cues under their assigned conditions for the two sessions. After the experiment, the participants were presented with the synchronized videos from all four sources (three videos for eye gaze, facial expression and body poses separately, and one video from the external camera for the

full body) and were asked to identify the social cues they used for each referent using retrospective think-aloud. Finally, we conducted a semi-structured interview to learn 1) their rationales for choosing or not choosing specific modalities and social cues and 2) how their ways of interaction might be similar or different under different robot forms and roles. The whole process took approximately 2 hours, with around 30 minutes for filling in the questionnaire and briefing, 10 minutes for warm-up, 10 minutes for each session, and 40 minutes for retrospective think-aloud and interview. Breaks were guaranteed between each stage.

### 3.3 Data Analysis

*3.3.1 Social Cue Coding.* We adopted an iterative approach to develop our codebook for social cues. Considering that most participants were only able to intuitively propose one social cue for each referent, we only coded the first elicited cue. Our social cue coding process involved the following steps:

- **Step 1. Initial Coding**: We first went through all the videos to list the observed frequent patterns to form our initial codebook. The codes were roughly classified by different body parts, e.g., arm and hand, eye, upper body, etc.
- **Step 2. Discussion and Iterative Development of the Codebook**: We used an iterative process to refine our codebook. We divided the whole dataset (24 participants × 2 sessions = 48 sessions) into 6 batches, each comprising 8 sessions. Two coders first used the initial codebook to code one batch of data and then discussed how to resolve conflicts with the intervention of a third researcher. Codes may be added, removed, or reorganized during this process, and the revised codebook is used to code the next batch. After two batches, there are no more updates to the codebook. After the third batch, we used Perreault & Leigh's approach ($I_r$) [71] to compute the inter-rater reliability (IRR) between the two coders, and got an IRR of 0.896, exceeding the pre-set requirement of 0.7, thus the iterative development of the codebook finished.
- **Step 3. Formal Coding**: The remaining data was divided up and coded independently by the two coders. The data used in the iterative development step was also re-coded using the finalized codebook.

Upon finalizing the codebook, we obtained 85 **codes**, categorized by different **modality (body parts)**, including Gesture, Verbal, Eye, Head, etc. For each modality, there are **articulation codes** which decompose the **social cues** elicited by participants into basic elements. Articulations are coded as explicit or implicit, where explicit signals refer to those explicitly reported by participants during the post-interaction retrospective think-aloud process, and implicit signals are those observed but not verbally acknowledged by participants. Due to the nuances of Gesture and Verbal, we further developed **feature codes** describing their detailed characteristics. A complete description of our codebook is as follows. Some of the most frequently used codes are illustrated in Figure 3.

- Gesture
  - Articulation codes for Gesture are unique gesture names, including waving gestures (e.g., *dismissive wave*, *beckoning wave*, *wave*), pointing gestures (e.g., *palm point across the table*, *palm point to the place for the drink*, *finger point to partner*) and others (e.g., *palm to stop the robot*, *hold the drink*, *thumb up*). Note that multiple gestures may be used in one social cue, so each code is treated as a binary variable.
  - Feature codes of Gesture further describes characteristics that apply to any gesture. This includes *handedness* (near-side hand / far-side hand), *number-of-hands* (single hand / both hands), *hand height* (lower than table level / not lower than table but lower than head level / head level / above head level), *repetition* (repetition / no repetition).

- Verbal
  - Articulation codes for Verbal include the *exact content* and *speech act* [78] of the content (declarative / interrogative / imperative / exclamative / short interjections).
  - Feature codes for Verbal include *volume* (decrease / no change / increase, compared with the chatting volumn), *unclear reference* (exist / not exist) and politeness. To measure politeness, we adopted the features in `politeness` R package [6] which is built upon previous research on computational linguistics [15, 21, 94], and manually selected 9 features that are applicable in HRI scenario as our codes: *Apology*, *Could You*, *First-Person Plural*, *Gratitude*, *Hello*, *Please*, *Positive Emotion*, *Reasoning*, *Subjectivity*, and *Subjectivity*.
- Eye. Articulation codes for Eye is defined based on the gaze target: *turn to the robot*, *turn to the drink* (further divided into *turn to the wrong drink*, *turn to the correct drink*), *turn to across the table* and *turn to the place for the drink*.
- Head. Articulation codes for Head is divided into two parts: 1) head motions that are caused by eye gaze, including *turn to robot*, *turn to drink* (further divided into *turn to wrong drink*, *turn to correct drink*), *turn to across the table*, *turn to the place for the drink*. 2) Head motions that are irrelevant to gaze, including *shake*, *nod*, *jaw point to across the table* and *turn to the side*.
- UpperBody. Articulation codes for UpperBody include *lean away from the robot*, *lean towards the robot*, *lean back*.
- LegAndFoot. Articulation codes for LegAndFoot include *stamp*, *foot draw circle*, *foot draw line*, *kick away*.
- Facial. Articulation codes for Facial include *purse the lips*, *frown*, *raise eyebrow* and *smile*.

For subsequent analysis, we define a **unique social cue** as a unique set of articulations from all modalities in our codes.

*3.3.2 Statistics.* Our study was a mixed design with two within-subjects variables (ConversationRole and Referent) and one mixed-design variable (RobotMorphology). The dependent variables are the modalities (Modality) and the feature codes (Gesture and Verbal) of human social cues. We analyzed the Referent variable with the statistical distributions and the agreement rates, and fitted Cumulative Link Mixed Models to assess the effects of ConversationRole and RobotMorphology on the Modality, Gesture and Verbal features of human social cues.

*Agreement Rate (AR).* To understand participants' consensus on each Referent representing their intentions, we calculated agreement rates for all unique cues. As introduced in [65, 89], the agreement rate for each referent $r$ is calculated with the following function:

$$AR(r) = \frac{|P|}{|P|-1} \sum_{P_i \subseteq P} \left( \frac{|P_i|}{P} \right)^2 - \frac{1}{|P|-1}, \tag{1}$$

where $P$ is the set of all social cues elicited for the referent $r$, and $P_i$ is the $i^{th}$ subset of identical codes in $P$. The margins for interpretation are $\leq 0.1$ for low agreement, $0.1 < AR \leq 0.3$ for medium agreement, $0.3 < AR \leq 0.5$ for high agreement, and $AR > 0.5$ for very high agreement [89].

*Statistical Analysis.* We employed Cumulative Link Mixed Models fitted with the adaptive Gauss-Hermite quadrature approximation to assess the effects of RobotMorphology and ConversationRole [19] on the Modality, Gesture and Verbal features of human social cues. Following the practice of [11], we treated participants and Referent as a

[6] https://cran.r-project.org/web/packages/politeness/vignettes/politeness.html#politeness-features

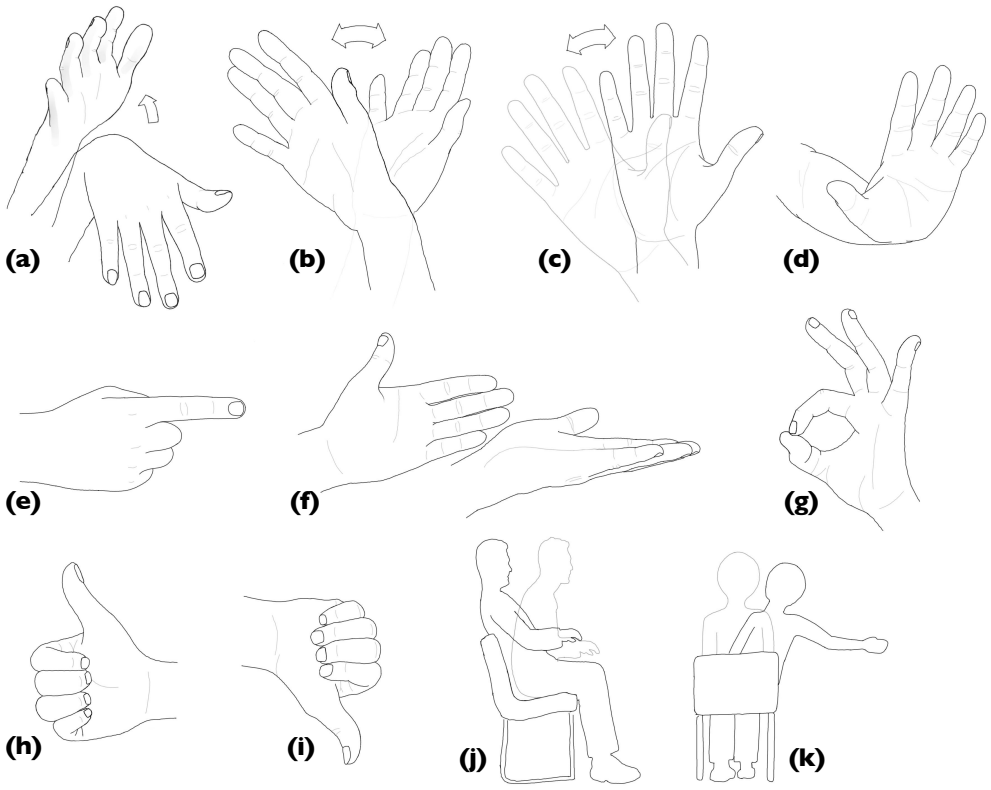


Fig. 3. Some Common Codes in Gesture Modality: (a) *dismissive wave*; (b) *beckoning wave*; (c) *wave*; (d) *palm to stop the robot*; (e) *finger point (to different targets)*; (f) *palm point (to different targets)*; (g) *ok*; (h) *thumb up*; (i) *thumb down*; and UpperBody Modality: (j) *lean back*; (k) *lean towards the robot*.

random effect since we regard the referents as samples from the common human-robot interactions in the coffee chat scenario. The baseline for the ConversationRole variable was set to *listener*. Since the location coefficients may vary in p-values for different baseline chosen for the RobotMorphology variable, we expanded RobotMorphology to four dummy variables *aerial technical*, *grounded technical*, *anthropomorphic*, and *zoomorphic* with levels 0, 1 to compare the effects of each robot morphology. The following cumulative link mixed model was fitted to the dependent variable $Y$ (Modality, Gesture or Verbal features of human social cues):

$$\begin{aligned}\text{logit}(P(Y_i \leq j)) = \theta_j &+ \beta_1(\text{ConversationRole}_i) \\ &+ \beta_2(\textit{aerial technical}_i) + \beta_3(\textit{grounded technical}_i) + \beta_4(\textit{anthropomorphic}_i) + \beta_5(\textit{zoomorphic}_i) \\ &+ \gamma_1(\text{ConversationRole}_i \times \textit{aerial technical}_i) + \gamma_2(\text{ConversationRole}_i \times \textit{grounded technical}_i) \\ &+ \gamma_3(\text{ConversationRole}_i \times \textit{anthropomorphic}_i) + \gamma_4(\text{ConversationRole}_i \times \textit{zoomorphic}_i) \\ &+ u(\text{Referent}_i) + v(\text{participant}_i),\end{aligned} \tag{2}$$

where $i = 1, \ldots, 78$, and $j$ represents the ordinal level of the dependent variable $Y$. To analyze the statistical significance of each model term, we applied forward selection with likelihood-ratio chi-squared tests [19] for each pair of models with progressive complexities.

### 3.4 Interview Data Analysis

All the video-recorded interviews are transcribed into text. Three researchers first independently read and watched the videos of the retrospective think-aloud and the interview process and familiarized themselves with 12 out of 24 experiments. Alongside the analysis, each researcher also referred to the corresponding videos where particular social cues were used. The whole research group then identified major themes focusing on participants' rationales in choosing their social cues through rounds of discussions. Due to the qualitative nature of the data, we did not conduct inter-rater reliability check [59]. Instead, we ensured the reliability of the analysis through both independent coding and cross-checking among the research team.

## 4 Result

### 4.1 Overall Statistics of Social Cues

We analyzed 624 observed (explicit + implicit) social cues (= 24 participants ×2 sessions ×13 referents) from participants' first choice to signal the service robot. We coded a total of 3,387 modality articulations, 3,318 gestural features, and 823 verbal features. Figure 4 presents the distributions of Modality by each referent. Overall, the most frequently used modalities are Eye (94.71%), Gesture (85.42%), and Head (84.78%). Verbal (33.97%) and UpperBody (10.58%) follow, while LegAndFoot (1.12%) as well as Facial (0.64%) are not depicted in the figure due to their relatively low frequency of use. The most frequently used explicit modalities are Gesture (83.81%) and Verbal (33.97%), while the most frequently used implicit modalities are Eye (86.38%) and Head (76.44%). When checking by referents, we can observe that *Signal Awareness* usually involves frequent usage of explicit Eye (31/48) and explicit Head (27/48), with few usages of Verbal (2/48). *Signal Emergency Stop* usually involves significantly more usage of UpperBody (17/48) modality. Besides, *Signal Wrong Drink* shows more usage of Verbal (28/48), possibly because this referent inherently carries more complex semantics.

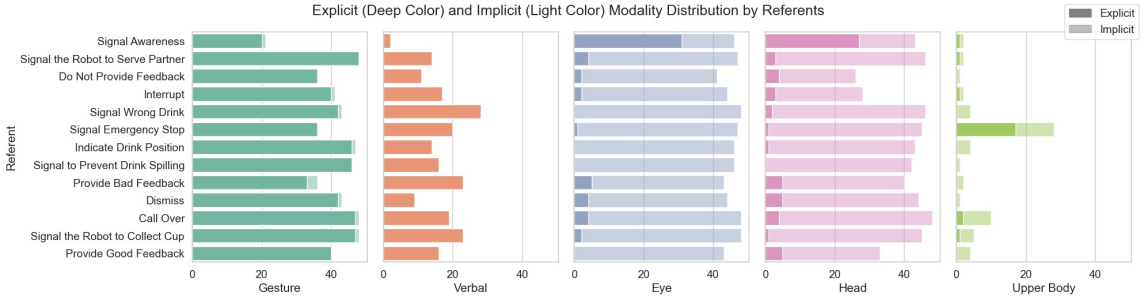


Fig. 4. The distribution of each modality (Gesture, Verbal, Eye, Head, UpperBody) by each referent. In each subplot, deep color refers to explicit usage of this modality; while light color refers to implicit usage.

We counted the number of unique social cues elicited by participants and listed the calculated agreement rates ($AR(r)$) by each referent $r$ for both observed cues and explicit cues, and as shown in Table 4. The agreement rates range

from 0.02 (low agreement, $AR \leq 0.1$) to 0.23 (medium agreement, $0.1 < AR \leq 0.3$) for observed cues and from 0.03 (low agreement, $AR \leq 0.1$) to 0.30 (medium agreement, $0.1 < AR \leq 0.3$) for explicit cues. The rather low agreement rates suggest the potential effects of other two IVs and the complexity of expressing the corresponding intentions through social cues, given the diverse combinations of articulations. Despite the low agreement rates, we still observed patterns of social cue articulations from the top 3 cues for each referent in Table 4. Referent *Signal Awareness* frequently includes head gaze and beckoning waves (Figure 3 (b)). Referent *Signal the Robot to Serve Partner* emphasizes pointing gestures towards the table's opposite side, most pointing gestures were pointing with an open palm (Figure 3 (f)) to show politeness. Referents *Do Not Provide Feedback* and *Interrupt* often feature *dismissive wave* (Figure 3 (a)) or *wave* (Figure 3 (c)) gestures to express "no" and dismissive intention to the robot waiter. Referent *Signal Wrong Drink* involves *wave* to say "wrong" with additional gestural or verbal explanations. For referent *Signal Emergency Stop*, participants use *palm to stop the robot* (Figure 3 (d), often with verbal imperatives. Referent *Indicate Drink Position* highlights pointing and tapping for drink placement. Referent *Signal to Prevent Drink Spilling* shows participants often *hold the drink*, sometimes with verbal cues, showing that participants would like to correct the drink themselves rather than asking the robot to do so. Participants use *thumb down* (Figure 3 (i)) gestures and declarative words to express negative sentiments for referent *Provide Bad Feedback*. The gesture *dismissive wave* is common in *Dismiss*, and the referent *Call Over* has more *beckoning wave*, which are aligned with the referent meanings. Referent *Signal the Robot to Collect Cup* relies on pointing to the opposite side, and *Provide Good Feedback* combines *thumb up* (Figure 3 (h)) gestures with verbal declarative.

Table 4. Top observed and self-reported cues by referents, where $AR(r)$ stands for the agreement rate of the referent.

| **Referents** | **Observed (Explicit + Implicit)** | | **Self-Reported (Explicit)** | |
|---|---|---|---|---|
| | $AR(r)$ | Top 3 Cues | $AR(r)$ | Top 3 Cues |
| *Signal Awareness* | 0.19 | Head gaze turn to robot (33%) | 0.18 | Head gaze turn to robot (35%) |
| | | Head gaze turn to robot + beckoning wave (27%) | | Beckoning wave (21%) |
| | | Head gaze turn to robot + nod (8%) | | Head gaze turn to robot + beckoning wave (10%) |
| *Signal the Robot to Serve Partner* | 0.06 | Head gaze turn to robot and the correct drink + palm point to the opposite side of the table (23%) | 0.09 | Palm point to the opposite side of the table (27%) |
| | | Head gaze turn to robot and the correct drink + verbal (imperative) + palm point to the opposite side of the table (6%) | | Palm point from robot to the opposite side of the table (12%) |
| | | Head gaze turn to robot and the correct drink + palm point from robot to the opposite side of the table (6%) | | Finger point to the opposite side of the table (10%) |
| *Do Not Provide Feedback* | 0.03 | Glance at robot + wave hand (10%) | 0.07 | Dismissive wave (17%) |

Table 4. Top cues by referents, where $AR(r)$ stands for the agreement rate of the referent. (continued)

| Referents | Observed (Explicit + Implicit) | | Self-Reported (Explicit) | |
|---|---|---|---|---|
| | $AR(r)$ | Top 3 Cues | $AR(r)$ | Top 3 Cues |
| | | Head gaze turn to robot + dismissive wave (10%) | | Wave hand (15%) |
| | | Glance at robot (8%) | | (Do nothing) (10%) |
| *Interrupt* | 0.02 | Head gaze turn to robot + dismissive wave (13%) | 0.05 | Dismissive wave (15%) |
| | | Head gaze turn to robot + show palm to stop the robot (8%) | | Show palm to stop the robot (13%) |
| | | Head gaze turn to robot + wave hand + verbal (declarative) (6%) | | Wave hand + verbal (declarative) (8%) |
| *Signal Wrong Drink* | 0.02 | Head gaze turn to robot and the wrong drink + wave hand + dismissive wave (10%) | 0.03 | Wave hand (10%) |
| | | Head gaze turn to robot and wrong drink + verbal (declarative) (8%) | | Wave hand + dismissive wave (10%) |
| | | Head gaze turn to robot and wrong drink + wave hand + verbal (declarative) (6%) | | Wave hand + verbal (declarative) (8%) |
| *Signal Emergency Stop* | 0.04 | Head gaze turn to robot + show palm to stop the robot (19%) | 0.07 | Show palm to stop the robot (21%) |
| | | Head gaze turn to robot + lean away from the robot (8%) | | Show palm to stop the robot + verbal (imperative) (15%) |
| | | Head gaze turn to robot + show palm to stop the robot + verbal (imperative) (6%) | | Show palm to stop the robot + verbal (declarative) (6%) |
| *Indicate Drink Position* | 0.03 | Head gaze turn to robot and the correct drink + palm point to the place for the drink + tap the table (15%) | 0.11 | Palm point to the place for the drink + tap the table (23%) |
| | | Head gaze turn to robot and the correct drink + head gaze turn to and palm point to the place for the drink (6%) | | Palm point to the place for the drink (23%) |

Table 4. Top cues by referents, where $AR(r)$ stands for the agreement rate of the referent. (continued)

| Referents | Observed (Explicit + Implicit) | | Self-Reported (Explicit) | |
|---|---|---|---|---|
| | $AR(r)$ | Top 3 Cues | $AR(r)$ | Top 3 Cues |
| | | Head gaze turn to the robot and the correct drink + palm point to the place for the drink (6%) | | Finger point to the place for the drink + tap the table + verbal (imperative) (6%) |
| *Signal to Prevent Drink Spilling* | 0.17 | Head gaze turn to robot and the correct drink + hold the drink with hand (42%) | 0.30 | Hold the drink with hand (54%) |
| | | Head gaze turn to robot and the correct drink + hold the drink with hand + verbal (imperative) (6%) | | Hold the drink with hand + verbal (short exclamation)(6%) |
| | | Head gaze turn to robot and the correct drink + hold the drink with hand + verbal (declarative) (6%) | | Hold the drink with hand + verbal (imperative) (6%) |
| *Provide Bad Feedback* | 0.04 | Head gaze turn to robot + thumb down (15%) | 0.07 | verbal (declarative) (19%) |
| | | Head gaze turn to robot + verbal (declarative) (8%) | | Thumb down (17%) |
| | | Head gaze turn to robot + wave hand (6%) | | Wave hand + dismissive wave (6%) |
| *Dismiss* | 0.23 | Head gaze turn to robot + dismissive wave (48%) | 0.20 | Dismissive wave (44%) |
| | | Head gaze turn to robot + show palm to stop the robot (6%) | | Show palm to stop the robot (8%) |
| | | (Do nothing) (4%) | | Head gaze turn to robot + dismissive wave (6%) |
| *Call Over* | 0.16 | Head gaze turn to robot + beckoning wave (40%) | 0.19 | Beckoning wave (42%) |
| | | Head gaze turn to robot + beckoning wave + verbal (imperative) (8%) | | Beckoning wave + verbal (imperative) (13%) |
| | | Head gaze turn to robot + beckoning wave + lean back (4%) | | Beckoning wave + verbal (declarative) (6%) |
| *Signal the Robot to Collect Cup* | 0.02 | Head gaze turn to robot + palm point to the opposite side of the table + verbal (imperative) (10%) | 0.04 | Palm point to the opposite side of the table (15%) |

Table 4. Top cues by referents, where $AR(r)$ stands for the agreement rate of the referent. (continued)

| Referents | Observed (Explicit + Implicit) | | Self-Reported (Explicit) | |
|---|---|---|---|---|
| | $AR(r)$ | Top 3 Cues | $AR(r)$ | Top 3 Cues |
| | | Head gaze turn to robot + palm point to the opposite side of the table (8%) | | Palm point to the opposite side of the table + verbal (imperative) (13%) |
| | | Head gaze turn to robot + head gaze turn to the opposite side of the table + palm point to the opposite side of the table (8%) | | Finger point to the opposite side of the table (8%) |
| *Provide Good Feedback* | 0.10 | Head gaze turn to robot + thumb up (23%) | 0.26 | Thumb up (46%) |
| | | Head gaze turn to robot + thumb up + verbal (declarative) (17%) | | Thumb up + verbal (declarative) (21%) |
| | | Glance at robot + thumb up (15%) | | verbal (declarative) (10%) |

### 4.2 Effects of RobotMorphology and ConversationRole On Social Cues

Due to the large number of dependent variables from social cue codes, we only present significant results from the statistical analysis of the effects on independent variables RobotMorphology or ConversationRole on different modalities, modality articulations and main modality features in the following sections.

*4.2.1 Modality.* No significant effect of ConversationRole or RobotMorphology on the use of Gesture or Head was found in the mixed model analysis, and since the labels are rare for LegAndFoot and Facial, we did not include them in the analysis. The effects of ConversationRole and RobotMorphology on the use of Verbal, Eye and UpperBody are summarized in Table 5 and presented below.

*Verbal.* During the coffee chat, participants chose to talk to the robot waiter among 25.96% of their social cues in the *listener* session, and participants chose to use verbal among 41.99% of their social cues in the *speaker* session. The LR test for the mixed models shows a significant effect on ConversationRole ($LR = 40.091, p < 0.001$) and *aerial technical* ($LR = 4.445, p < 0.05$) respectively, and the interaction effect of ConversationRole and *aerial technical* is significant ($LR = 9.316, p < 0.01$). The coefficient of *speaker* ($1.137, p < 0.01$) shows that in the *speaker* session, participants are more likely to include verbal in their social cues. The coefficient of *speaker* × *aerial technical* shows a significant interaction effect ($2.855, p < 0.01$). As *listener*, participants are less likely to use verbal cues when interacting with *aerial technical* robots (16.7%) than non-*aerial technical* robots (29.1%). As *speaker*, participants are more likely to use verbal cues when interacting with *aerial technical* robots (56.4%) than non-*aerial technical* robots (37.2%).

*Eye.* The proportion of observed social cues that include eye gaze of each RobotMorphology is 96.15% for *aerial technical*, 98.08% for *anthropomorphic*, 96.79% for *grounded technical*, and 87.82% for *zoomorphic*. Among the four RobotMorphologys, the LR test for the mixed models shows a significant effect on the use of eye gaze ($LR = 18.212, p <$

Table 5. Regression coefficients for predicting Modality features using Cumulative Link Mixed Models with forward progressive selection using likelihood-ratio Chi-squared tests. For each factor, the baseline is indicated in parentheses. In the table, $***$: $p < 0.001$; $**$: $p < 0.01$; $*$: $p < 0.05$. "LR" stands for likelihood-ratio for model selection. Non-significant interaction terms are not shown.

| | | | Modality | | |
|---|---|---|---|---|---|
| | | | Verbal | Eye | UpperBody |
| ConversationRole (*listener*) | | LR | **40.091**$^{***}$ | 0.554 | 2.693 |
| | | *speaker* | **1.137**$^{**}$ | - | - |
| RobotMorphology | *aerial technical* (non-*aerial technical*) | LR | **4.445**$^{*}$ | 1.261 | 0.742 |
| | | *aerial technical* | **-0.708** | - | - |
| | *anthropomorphic* (non-*anthropomorphic*) | LR | 2.998 | 1.369 | 0.268 |
| | | *anthropomorphic* | - | - | - |
| | *grounded technical* (non-*grounded technical*) | LR | 0.463 | 0.002 | 0.166 |
| | | *grounded technical* | - | - | - |
| | *zoomorphic* (non-*zoomorphic*) | LR | 1.280 | **18.212**$^{***}$ | **34.596**$^{***}$ |
| | | *zoomorphic* | - | **-1.808**$^{***}$ | **2.170**$^{***}$ |
| ConversationRole × RobotMorphology | | LR | **9.316**$^{**}$ | - | - |
| | | *speaker* × *aerial technical* | **2.866**$^{**}$ | - | - |

0.001). Participants are less likely to glance at the robot waiter when it is *zoomorphic* ($-1.808, p < 0.001$), possibly due to the heights of this robot is not in the normal height of human sights.

*UpperBody.* Participants rarely move their bodies during the coffee chat. Among all observed social cues, only 8.33% movement of upper bodies for *aerial technical*, 6.41% for *anthropomorphic*, 4.49% for *grounded technical*, and 23.08% for *zoomorphic*. The effects of *zoomorphic* on the use of upper body movement are significant ($LR = 34.596, p < 0.001$). Participants move their upper bodies more when interacting with the *zoomorphic* robot waiter ($2.170, p < 0.001$), possibly for bending over to interact with it (19.23%).

*4.2.2 Gesture Feature.* Among all coded Gesture features, the ConversationRole has significant effects on *repetition* and *number-of-hands*, while *aerial technical* and *zoomorphic* both have a significant effect on *hand height*. None of the ConversationRole or four types of RobotMorphology has significant effects on other Gesture features. The results are summarized in Table 6.

*Gesture - repetition.* The proportion of Gesture with repetition in the *listener* session is 46.79%, and the proportion of Gesture with repetition in the *speaker* session is 51.28%. The proportion of Gesture without repetition in the *listener* session is 39.42%, and the proportion of Gesture without repetition in the *speaker* session is 33.01%. The ConversationRole has a significant effect on the use of Gesture with repetition ($LR = 4.829, p < 0.05$). When the participants are in the *speaker* session, they are less likely to use Gesture with repetition ($-0.391, p < 0.01$), while they are more likely to use repetitive Gesture as *listener*.

*Gesture - number-of-hands.* The distribution of Gesture with different *number-of-hands* used in the *listener* session is 75% for *single-handed* and 11.54% for *two-handed*. The distribution of Gesture with different *number-of-hands* used

Table 6. Regression coefficients for predicting Gesture features using Cumulative Link Mixed Models with forward progressive selection using likelihood-ratio Chi-squared tests. For each factor, the baseline is indicated in parentheses. In the table, $***$: $p < 0.001$; $**$: $p < 0.01$; $*$: $p < 0.05$. "LR" stands for likelihood-ratio for model selection. Non-significant interaction terms are not shown.

| | | | **Gesture Feature** | | |
|---|---|---|---|---|---|
| | | | *repetition* | number-of-hands | *hand height* |
| ConversationRole (*listener*) | | LR | **4.829*** | **5.478*** | - |
| | | *speaker* | **-0.391**** | **-0.539*** | - |
| RobotMorphology | *aerial technical* (non-*aerial technical*) | LR | 0.530 | 0.012 | **44.537***** |
| | | *aerial technical* | - | - | **1.644***** |
| | *anthropomorphic* (non-*anthropomorphic*) | LR | 0.214 | 0.725 | 0.876 |
| | | *anthropomorphic* | - | - | - |
| | *grounded technical* (non-*grounded technical*) | LR | 1.832 | 0.113 | 0.876 |
| | | *grounded technical* | - | - | - |
| | *zoomorphic* (non-*zoomorphic*) | LR | 2.509 | 1.719 | **36.178***** |
| | | *zoomorphic* | - | - | **-1.449***** |

in the *speaker* session is 77.24% for *single-handed* and 7.05% for *two-handed*. The ConversationRole has a significant effect on the *number-of-hands* ($LR = 5.478, p < 0.05$). Participants are less likely to use both hands in the *speaker* session ($-0.539, p < 0.05$).

*Gesture - hand height.* The distribution of different *hand height* for different RobotMorphology is as follows: *aerial technical* - 41.67% for between table and head, 32.05% for head level, and 13.46% for above head. *anthropomorphic* - 3.21% for below table, 64.74% for between table and head, 15.38% for head level, and 2.56% for above head. *grounded technical* - 0.64% for below table, 73.72% for between table and head, 9.62% for head level, and 0.64% for above head. *zoomorphic* - 30.77% for below table, 50.64% for between table and head, 1.92% for head level, and 0.64% for above head. The LR test shows a significant effect on *aerial technical* ($LR = 44.537, p < 0.001$) and *zoomorphic* ($LR = 36.178, p < 0.001$). The height of participants' hands are lower when interacting with *zoomorphic* ($-1.449, p < 0.001$) robots and higher when interacting with *aerial technical* robots ($LR = 1.644, p < 0.001$), which is consistent with the height of these robots.

*4.2.3 Verbal Feature.* Among all coded Verbal features, ConversationRole has significant effects on *unclear reference*, politeness - *Gratitude* and politeness - *Reasoning*. For *volume*, the effects of ConversationRole and *zoomorphic* are significant. A summary of the effects of RobotMorphology and ConversationRole on verbal features is presented in Table 7.

*Verbal - unclear reference.* An *unclear reference* is a deictic expression (e.g., this / that / here / there) whose real-world meaning is dependent on the context. Among all the social cues, the proportion of *unclear reference* in the *listener* session is 7.37%, and the proportion of *unclear reference* in the *speaker* session is 12.50%. The LR test for the mixed models shows a significant effect of ConversationRole on the use of pronouns ($LR = 7.612, p < 0.01$). When the participants are act as *speaker*, they are more likely to referencing something unclearly to robot in their social cues ($0.985, p < 0.01$).

Table 7. Regression coefficients for predicting Verbal features using Cumulative Link Mixed Models with forward progressive selection using likelihood-ratio Chi-squared tests. For each factor, the baseline is indicated in parentheses. In the table, $***$: $p < 0.001$; $**$: $p < 0.01$; $*$: $p < 0.05$. "LR" stands for likelihood-ratio for model selection. Non-significant interaction terms are not shown.

| | | | **Verbal Feature** | | | |
|---|---|---|---|---|---|---|
| | | | unclear reference | politeness - *Gratitude* | politeness - *Reasoning* | *volume* |
| ConversationRole (*listener*) | | LR | **7.612**$^{**}$ | **8.302**$^{**}$ | **7.661**$^{**}$ | **52.485**$^{***}$ |
| | | *speaker* | **0.985**$^{**}$ | **1.487**$^{**}$ | **1.508**$^{*}$ | **1.622**$^{***}$ |
| Robot Morphology | *aerial technical* (non-*aerial technical*) | LR | 0.040 | 0.631 | 1.394 | 1.193 |
| | | *aerial technical* | - | - | - | - |
| | *anthropomorphic* (non-*anthropomorphic*) | LR | 1.428 | 0.351 | 2.924 | <0.001 |
| | | *anthropomorphic* | - | - | - | - |
| | *grounded technical* (non-*grounded technical*) | LR | 0.279 | **3.985**$^{*}$ | 0.936 | 1.163 |
| | | *grounded technical* | - | **-1.659** | - | - |
| | *zoomorphic* (non-*zoomorphic*) | LR | 3.009 | 0.045 | 0.091 | **8.294**$^{**}$ |
| | | *zoomorphic* | - | - | - | **-0.855**$^{**}$ |

*Verbal - politeness words.* For politeness word *Gratitude*, the result of LR test shows a significant effect on ConversationRole ($LR = 8.302, p < 0.01$) and *grounded technical* ($LR = 3.985, p < 0.05$). 2.88% of the social cues include *Gratitude* words in their speech content when the participants are the *listener*, and 7.05% of the social cues include *Gratitude* words when the participants are the *speaker*. The *speaker* role uses significantly more *Gratitude* words ($1.487, p < 0.01$). *Reasoning* is another way of showing politeness, and the ConversationRole also has a significant effect on the use of *Reasoning* words ($LR = 7.661, p < 0.01$). 2.88% of the social cues include *Reasoning* words in their speech content when the participants are the *listener*, and 6.41% of the social cues include *Reasoning* words when the participants are the *speaker*. As *speaker*, participants are significantly more likely to use *Reasoning* words ($1.508, p < 0.05$) in their social cues.

*Verbal - volume.* Both ConversationRole ($LR = 52.485, p < 0.001$) and *zoomorphic* ($LR = 8.294, p < 0.01$) have significant effects on the verbal *volume* of the social cues. The coefficient of *speaker* ($1.622, p < 0.001$) shows that participants are more likely to use higher volume in the *speaker* session. The coefficient of *zoomorphic* ($-0.855, p < 0.01$) shows that participants are more likely to use lower volume when talking to the *zoomorphic* robot.

### 4.3 Interview

*4.3.1 Considerations on Modality and Articulation.* Generally, participants would choose minimum articulations for their social cues to guarantee efficiency and minimize cognitive load according to the interview results. As most explicit social cues include Gesture or Verbal modalities, we would mainly focus on the rationales for choosing these two modalities in the following sections.

*Gesture.* For explicit cues mentioned in the interview, participants chose to use the Gesture modality the most when signaling the robot waiter during the conversation (Section 4.1). The rationale is that they are more familiar with the Gesture, and it is natural and intuitive for them to use common gestures (U04, U09, U15, U22). Many participants reckoned that the robot waiter should recognize commonsense gestures like *dismissive wave* for "Go away", *beckoning*

*wave* for "Come", and *thumbs up or down* for good or bad feedback (U07, U21, U22). For more complex situations, some of them assumed that the robot waiter should understand the context and the intention behind the gestures (U04, U05, U22), such as whether the *pointing to a drink* gestures mean that the robot should offer the drink or collect the empty cup (U13, U19, U21, U23, U24), or the *wave* gesture to express "No" or "Wrong".

*Verbal.* Participants mentioned that they would use verbal cues when they want to express their intentions clearly and efficiently (U04, U07, U12). Some of them mentioned that they think gestures cannot properly express some of their intentions, so they would use verbal cues to clarify (U15, U19, U22), for example, they would *Signal Wrong Drink* with the *wave* gesture and the verbal cue "Fanta" to the robot waiter with the wrong order. Meanwhile, U20 mentioned that they were not used to speaking to the robot in public, thus, they did not use verbal cues in the whole experiment.

*Other Modalities.* We also asked participants about their eye gaze if we noticed that their gaze turned to the referent-related objects or the robot waiter during the conversation. Most participants said their gaze would unconsciously follow their social cues (U01 U11, U22), some of them mentioned that they were attracted by the robot waiter's movement even if they did not intend to signal the robot waiter (U06, U09), and some said they kept looking at the robot waiter to make sure it understands their signal and does not do anything unexpected (U11, U18). The Head modality mainly goes with gaze, or some participants would nod to *Signal Awareness* or *Provide Good Feedback* to the robot (U03). As for other modalities, participants mentioned that they would not use Facial to signal the robot waiter since most thought it was not socially appropriate (U15), and only a few participants mentioned smiling, raising their eyebrows, or blinking to signal awareness or confirmations to the robot waiter (U17, U20). And they would not use LegAndFoot to signal the robot waiter since they were sitting and did not have the habit of using their legs or feet to signal others (U20, U22). The only participant using LegAndFoot mentioned that they had an experience participating in a project that collected their foot signals, and it reminded them to use their legs to signal the robot waiter (U15).

*4.3.2 Conversation Role.* Participants generally felt more cognitively demanding when they were the *speaker* in the conversation, so they would choose the most simple social cue or even ignore the robot to avoid distractions when they were in the middle of presenting something to the potential employers (U19, U24). Some participants mentioned that they would use more verbal cues when they were the *speaker* in the chat interrupted by the robot waiter (U03, U05, U07, U08, U10). Since they were the ones who managed the conversation flow, when they could pause the topic, they would quickly clarify their intentions to the robot waiter with simple commands. Compared to gesturing, some of them felt speaking was a more straightforward way to instruct the robot and minimize distractions since they were speaking and did not worry about interrupting the potential employers in the conversation (U07, U08). As for the *listener* session, many participants mentioned that they would avoid speaking to the robot waiter to avoid interrupting the conversation, and they would use more complex gestures to signal the robot waiter since they felt that they could leave the conversation for a little while to settle down the robot interruptions when they were not talking in the coffee chat (U07, U08, U19, U21). Nevertheless, a few participants mentioned that they would still use verbal cues when they were the listener in the conversation (U02, U06). They would turn down their volume to quickly instruct the robot waiter in the *listener* session, while they said that they did not have the cognitive load to use verbal cues to signal the robot in the *speaker* session. Overall, some participants mentioned that they thought it was more appropriate for the robot to find the right people who were not speaking to interact with if the waiter had to interrupt the conversation (U09, U14, U22).

*4.3.3 Robot Perception.* The *aerial technical* and *grounded technical* robot waiters were perceived as machines or tools (U08, U12), *zoomorphic* was perceived as more playful to attract customers (U12, U17), and *anthropomorphic* was perceived more like a human waiter (U07, U20). U18 said they would be more strict with the *anthropomorphic* robot since they thought if the robot looked like humans and did not perform the tasks as well as humans, then it would disappoint them. Given the different perceptions, most participants did not think that the appearance of the robot waiter would affect their choice of social cues, but only the physical properties such as heights or more specifically, the sensor positions would affect their social cues (U07, U08, U10, U19, U21, U22, U24). For example, they perceived that the cameras of the robot waiters were in the front or on the "face", so they would gesture at a higher position to *aerial technical* and *anthropomorphic* robots, and blend over to gesture at a lower position to the *zoomorphic* robot (U21, U22, U24). Similarly, some of them felt that the robot waiter could not "hear" them clearly if their perceived microphones of the robot were too far away, so they either chose not to use verbal cues unless the robot was closer (U11) or leaned their bodies to get closer to the robot waiter to speak to it (U13, U22, U23). Moreover, the participants' perceived capability and intelligence of the robots would also affect their behaviors in the interaction. Many participants complained about the interruptions caused by the robot waiter, which reduced their willingness to interact with the robot waiter (U01, U11). Some thought the robot waiter was not responding to their social cues, so they may repeat their cues several times or do the job themselves instead of signaling the robot to do it (U13). Some participants mentioned that they felt the robot waiter was not intelligent enough to understand their social cues, so they would use more explicit cues to signal the robot waiter (U13). Given an emergency error on purpose in our referents (*Signal Emergency Stop*), only one participant (U11) mention feelings of unsafe when interacting with the *aerial technical* robot during the experiment, due to the experience with an unsteady drone. That participant kept a safe distance during the interaction with the *aerial technical* robot. This suggests that although some mentioned that they may change their behavior if the robot waiters become more dangerous (U11, U20), most participants may feel safe when interacting with the robot waiters during the experiment.

*4.3.4 Social Context.* Most participants would consider the politeness and the social appropriateness of their social cues. For the Gesture modality, they would not use the *dismissive wave* or *finger pointing* to the direction of the two potential employers to avoid misunderstanding (U19, U20, U22). Some of them would hide their gestures below the table, or near their body to signal the robot inconspicuously (U01 U12, U19). For the Verbal modality, some participants would cover their mouth and whisper to the robot to avoid interrupting the speakers (U05, U19, U22), and one (U16) hoped that the robot waiter could read their lip so that they could signal the robot without speaking out loud. Some of them said that they would use polite words to instruct the robot waiter either because they politely treat the robot waiter the same as human waiters (U12), or because they want to show their etiquette in front of two potential employers (U12), while some took the robot waiter as a machine or a tool and thus use simple and direct words to instruct the robot waiter (U06). Despite the simulated social scenario in our experiment, participants generally reflected that the chosen public space was appropriate for a coffee chat (U04, U05, U07, U08, U09, U14, U15, U17, U21, U22, U23), while some (U08,U13,U18) mentioned that the environment is quieter than coffee chats they experienced and some (U11,U12) said they would adjust the volume and subtlety of their social cues according to the background music or noise level of the environment.

*4.3.5 Other Rationale and Comment.* Participants mentioned personal reasons for choosing some social cues, for example, some were more introverted, so they would choose inconspicuous cues to signal the robot waiter and avoid speaking (U07, U12), and some may prefer ignoring the robot waiter in most cases to avoid interruptions and distractions

(U02, U03). Some participants kept using similar gestures for different referents based on their habits, for example, U08 waved his hand in all dimensions throughout both sessions, and U22 used the same *palm to stop the robot* gesture for a large portion of referents. Several participants said they preferred to rate on a touch screen rather than use social cues to signal the robot waiter (U02), and some hoped there was a button on the table so that they could press to call over the robot waiter or specify the service type (U10).

## 5 Discussion

### 5.1 Considerations on Human Social Cues for HRI in Social Contexts

Our work provides empirical insights showing how human perceptions of robots, the primary task, and social context affect human social cues in human-robot side interactions. We discuss the implications and generalizations of our findings from the perspectives of humans, robots, and the social context.

*5.1.1 Choice of Social Cues and Human Intention.* Social cues are usually the most efficient way of communication when interacting with a service robot as a side task, according to the qualitative results in Section 4.3.1. Our participants' top explicit choices of Gesture and Verbal modalities are consistent with the common channels of human conversational interaction [73]. We identified some frequently used articulations of explicit cues, such as pointing, waving, and imperative sentences. We found people may use similar gestures for different intentions and expect the robot to understand the context to disambiguate the intentions. At the same time, there were also variances in their choice of social cues according to personal habits and preferences. These findings suggest that the design of social service robots should consider the common patterns of human social cues to understand user intentions, as well as individual differences in human social behaviors to provide personalized services.

*5.1.2 Perception on Robot.* According to the interview analysis in Section 4.3.3, the *anthropomorphic*, *zoomorphic*, or *technical* appearance of robots may not have direct impacts on participants' choice of social cues in out social settings. However, the four different morphologies may affect participants' perceptions of these service robots, leading to the change of interaction behaviors. The reported influences of perceptions mainly lies in the following dimensions: perceived physical capability, perceived safety and perceived intelligence, which are consistent with [12] that robot morphology and behavior are two main factors that affect human perceptions (i.e., anthropomorphism, animacy, likability, perceived intelligence, and perceived safety) of robots.

*Perceived Physical Capability.* The interview results show that participants' assumptions on the robots' physical embodiments and capabilities may affect their behaviors and social cues. Their assumptions usually follows common *anthropomorphic* patterns. For the *anthropomorphic* and *grounded technical* robots, such as they perceived the robots' cameras and microphones are located at the "head" of the robot. And for *zoomorphic* and *aerial technical* robots, participants assume the robots' cameras and microphones are located in the front of the robot, which are more likely to be perceived as the "face" of the robot. These findings align with [68] that the morphology shapes a user's expectations on the functioning of robots, and they can be further connected to the choice of social cues that participants adjust their social cues based on the perceived robot's physical capability to make the interactions effective. Our quantitative results verify that the robot morphology has significant effects on *hand height*, *volume*, and *upper body motion* (Section 4.2). Participants' adjustments of social cues based on the perceived physical embodiment highlight the importance of the design of the sensor placement and the communication of robots' sensing capabilities. To collect valid and high-quality human social cues for better recognition and understanding, the position of the robot sensors should either align with

the common human assumptions or embody clearly on the robot's appearance so that the users know where to signal the robot.

*Perceived Safety.* According to the participants' familiarity to four morphologies (Section 3.2.1) and the interview (Section 4.3.3), most participants have heard of the four morphologies of social service robots and many of them have interacted with them in canteen, hotels, or other places for entertainment before. Participants' general positive knowledge or experience on these social service robots can partially explain why only one participant with an unsafe experience with a drone mention an unsafe feeling with the *aerial technical* robot during the experiment. The virtual prototyping of the robots in the experiment may also contribute to the participants' perceived safety, as they may not feel threatened by the robots in the virtual environment. Nevertheless, connecting with participants' social cues, some participants (8%) lean away from the robot to keep a safe distance for the referent *Signal Emergency Stop* (Table 4) and the significant increase in the use of UpperBody for the *aerial technical* robot waiter suggest that participants use their body signal and proxemics to keep themselves safe. These findings complement the design of Mandal and Baraka [54] that robots can utilize the human proxemics as a corrective feedback signal. To better understand how each robot morphology affects human perceived safety and the choice of social cues, future research can consider explicitly asking participants about their perceptions on each robot morphology's safety given their experience with the robots and examine the relationship between the perceived safety and the choice of social cues.

*Perceived Intelligence.* The interview results show that participants' perceptions of the intelligence of the robots may influence human's willingness to interact with the robot, i.e., they may show more impatient behaviors in their explicit or implicit social cues. From appearance only, the *anthropomorphic* robots are expected to be more intelligent than *zoomorphic*, *grounded technical* and *aerial technical* robots according to the interview (Section 4.3.3). However, if their behaviors are not aligned with the participants' expectations on their intelligence, they may become more impatient when interacting with them. The variation of perceived intelligence is aligned with the findings from Tusseyeva et al. [88]'s survey. Therefore, it is important to design the robot behaviors to match the participants' expectations on their intelligence to maintain the participants' willingness to interact with the robots.

*5.1.3 Primary Task and Social Context.* The primary task and social context also affect human social cues in human-robot side interactions. People may adjust their social cues based on the primary task they are engaged in, as well as the social context they are in. According to our qualitative results (Section 4.3.4), three main factors play important roles in the choice of social cues: their cognitive load, social attention received, and the social appropriateness of the cues. The significantly more choice of Verbal cues in the *speaker* condition (Section 4.2.1), as well as the significantly more complex Gesture (*repetition* in Section 4.2.2, and *handedness* in Section 4.2.2) in the *listener* condition, may be due to the higher cognitive load in the *speaker* condition according to the interview. Thus, highly mentally demanding tasks may lead to simpler and more direct social cues to avoid distractions. At the same time, people usually choose more socially appropriate cues especially when they receive more social attention, echoing the results of significantly more politeness words chosen in the *speaker* session (Section 4.2.3). Therefore, it is important for social service robots to understand the interaction context including the surrounding environment, the primary task, and the social context, in order to provide appropriate services and receive detailed instructions and feedback from the users. In situations that require immediate human responses, social service robots need to be equipped with the ability to understand users' intentions, given the minimum set of social cues.

## 5.2 Design Implications for Social Service Robot Interactions

*5.2.1 Robot Approaching Strategy.* People's general unwillingness to interact with the robot waiter in the *speaker* session and complaints about the interruptions of the robot waiter suggest that the approaching strategy is important for the robot to initiate interactions with users (Section 4.3.3). The robot should find the right time to serve the users, especially when they are engaged in their main tasks. In terms of the conversation in our work, the pauses between conversation turns may be good timing for the robot to approach the users according to the previous HRI research [66, 69]. Nevertheless, in cases that require immediate human feedback, finding the right person to interact with when serving a group of people is also important to minimize the interruption. The cognitive load of the users discussed in Section 5.1.3 suggests that it is better for the robot to approach the person who is less involved in the primary task, where detecting human's cognitive load is a potential challenge and future design considerations.

*5.2.2 Human Social Cue Processing.* The general patterns of explicit and implicit human social cues (Section 4.1) suggest that the robot should be able to process all modalities of cues to better understand human intentions. The robot should be able to deduce the human willingness to interact, the instructions they signaled, and the feedback they provided. The main challenge would be learning the commonsense semantics of human social cues and distinguishing similar social cues in different contexts (Section 4.3.1). One potential consideration from our findings is that the robot should be able to learn from the implicit cues, such as the gaze direction or implicit facial expressions, to disambiguate the human intentions, emphasize the important parts, and figure out the scope of the human feedback. These implicit cues for robots are functioning as expressing emotion and sending relational messages according to [92], which are important parts of human social communications and necessary to be considered in the robot design in human social cue processing.

*5.2.3 Robot Response.* Upon receiving human social cues and understanding human intentions, the robot should be able to respond appropriately to deal with human needs. According to the interview, a slow response will increase the participants' impatience and decrease their willingness to interact with the robot (Section 4.3.3), which aligns with [43]. The robot should be able to respond in a timely and socially appropriate manner to maintain human engagement and interaction, and the response should be intuitive and easy to understand and at the same time guarantee the minimum interruptions to humans. Moreover, the robot should be able to respond to human feedback and instructions and take appropriate actions to fulfill human needs, which highly depends on the recognition and disambiguation of human social cues. Strategies to repair potential failures are also important considerations. On the one hand, when the intention recognition fails, the service robot may need to initiate additional interactions to confirm with humans. On the other hand, the interruptions caused by the social service robot service annoy the users. How the service robot repairs the relationship with the users to regain their trust [25] after undesirable situations is also a key consideration for the robot response strategy.

## 5.3 Limitation and Future Work

Our work has several limitations. First, we used an augmented reality (AR) headset to simulate the robot waiter in the given social context. Previous study [41] suggest the potential of using AR simulations in experiments that require high controllability and the comparable personal spaces and human behaviors towards virtual and real robot prototypes. Thus, using AR to simulate robots or other experiment settings is a good choice for prototype experiment settings to provide insights into understanding users' behavior and implicating the designs. Still, real tests on robots or products

are needed to ground the results in real-world applications. Second, our study had a relatively small sample size. As a result, the independent and identically distributed sample size for each unique condition is small and thus insufficient for the statistical test of the independence between conditions in the same independent variable. Instead, we relied on the mixed effect analysis of each variable (given a large enough overall sample size with repeated measures) and qualitative analysis of the interview to understand how different factors may affect human decisions and rationales for choosing social cues to signal the robot in the scope of this paper. Third, the participants were mainly students from local universities, which may limit the generalizability of the results to other populations, such as older adults who are less active in different body parts [13, 64] or people from a different culture. Fourth, due to the complexity of the social context, we only focused on one scenario of social interaction (i.e., coffee chat), which may not cover all referents and the corresponding social cues in other scenarios involving service robots. In addition, according to the suggestions of our pilot study participants, we chose the sitting pose for their comfort, which may limit the user of cues involving lower limb motion, whole body movement, or change in proximity.

Future work may recruit a larger number of participants from more diverse backgrounds to explore the influence of various personal and cultural factors in human-robot communication. Further exploration may consider adding standing pose, other social arrangements, and richer service scenarios to acquire a more comprehensive understanding of human interactions with robots in the social context. Processing and learning from the collected social cue data is another potential direction for future work, aiming to equip service robots with the ability the to understand and tackle users' preferences and needs in real contexts.

## 6 Conclusion

This work aims to gain empirical insights into how the robot morphology and human leading task roles have an impact on the human choice of social cues to express their intentions in social encounters. We conducted an elicitation study with 24 participants in a simulated coffee chat scenario with potential employers, where participants interacted with four different robots (*aerial technical*, *anthropomorphic*, *grounded technical*, and *zoomorphic*) in two different roles (*speaker* and *listener*). Our study collected 624 for observed social cues, including detailed coded articulations and features for each social cue. The statistics and quantitative results reveal patterns in human social cues' modalities, articulations, and features. Additionally, qualitative analysis of the interview backs up the quantitative results. It provides a deep insight into the participants' rationale in choosing the social cues and comments on the robots' appearance, behavior, and social context. The conversational role significantly affects the adoption of verbal cues, verbal features, and gesture features, mainly due to different cognitive loads and social norms brought by the two roles. The robot morphology significantly affects the adoption of different modalities, gesture features and verbal features, mainly due to the robot's appearance (e.g., height and similarity to humans). From these findings, we identify implications for understanding human social cues and inform robot design. For example, service robots with distinct morphologies should account for intuitive adjustments users make based on robot appearance, such as designing sensor placements that align with users' expectations. Additionally, service robots operating in professional or social contexts should adopt response strategies that minimize disruption, such as detecting subtle gazes or hand gestures and responding with polite, context-sensitive feedback. These design considerations provide actionable insights for service robots, particularly in their approach strategies, interpretation of human social cues, and interaction responses, to better integrate into dynamic human-centric environments.

## References


[1] Smart Delivery Robot-Pudu Robotics [n. d.]. *BellaBot-Pudu Robotics*. Smart Delivery Robot-Pudu Robotics. https://www.pudurobotics.com/products/bellabot

[2] Smart Delivery Robot-Pudu Robotics [n. d.]. *PuduBot2-Pudu Robotics*. Smart Delivery Robot-Pudu Robotics. https://www.pudurobotics.com/products/pudubot2

[3] 365Robot Pte Ltd [n. d.]. *Robot Waiter | Restaurant Food Delivery Robot*. 365Robot Pte Ltd. https://www.365robot.sg/robot-waiter/

[4] [n. d.]. *Roomba® Robot Vacuum Cleaners | iRobot®*. https://www.irobot.com/en_US/roomba.html

[5] [n. d.]. Tourists Amazed as China Uses Robot Dog to Transport Waste from Mountain Hotspot. https://www.newsflare.com/video/688892/tourists-amazed-as-china-uses-robot-dog-to-transport-waste-from-mountain-hotspot

[6] [n. d.]. Waiter Drone | Video | Facebook. https://www.facebook.com/watch/?v=1368908014016431

[7] 2013-06-18. Drones as Waiters: Sushi Takes Off. https://www.youtube.com/watch?v=ciNphCpmnSo

[8] Tech in Asia 2015-02-10T08:36:28. *Singapore Restaurant Shows off Autonomous Drone Waiters*. Tech in Asia. https://www.techinasia.com/singapore-restaurant-autonomous-drone-waiters

[9] 2019-01-25T12:51:44. Robots Serve up Food and Fun in Budapest Cafe. *Reuters* (2019-01-25T12:51:44). https://www.reuters.com/article/technology/robots-serve-up-food-and-fun-in-budapest-cafe-idUSKCN1PJ1EK/

[10] Nalini Ambady, Frank J. Bernieri, and Jennifer A. Richeson. 2000. Toward a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of the Behavioral Stream. In *Advances in Experimental Social Psychology*. Vol. 32. Academic Press, 201–271. https://doi.org/10.1016/S0065-2601(00)80006-4

[11] Alexandru-Tudor Andrei, Laura-Bianca Bilius, and Radu-Daniel Vatavu. 2024. Take a Seat, Make a Gesture: Charting User Preferences for On-Chair and From-Chair Gesture Input. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3642028

[12] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[13] Mriganka Biswas, Marta Romeo, Angelo Cangelosi, and Ray B. Jones. 2020. Are Older People Any Different from Younger People in the Way They Want to Interact with Robots? Scenario Based Survey. *Journal on Multimodal User Interfaces* 14, 1 (March 2020), 61–72. https://doi.org/10.1007/s12193-019-00306-x

[14] Joost Broekens and Mohamed Chetouani. 2021. Towards Transparent Robot Learning Through TDRL-Based Emotional Expressions. *IEEE Transactions on Affective Computing* 12, 2 (2021), 352–362. https://doi.org/10.1109/TAFFC.2019.2893348

[15] Penelope Brown, Stephen C. Levinson, and John J. Gumperz. 1987. *Politeness: Some Universals in Language Usage* (1 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

[16] Clebeson Canuto, Eduardo O Freire, Lucas Molina, Elyson A N Carvalho, and Sidney N Givigi. 2022. Intuitiveness Level: Frustration-Based Methodology for Human–Robot Interaction Gesture Elicitation. 10 (2022).

[17] Jessica R. Cauchard, Jane L. E, Kevin Y. Zhai, and James A. Landay. 2015. Drone & Me: An Exploration into Natural Human-Drone Interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Osaka Japan, 361–365. https://doi.org/10.1145/2750858.2805823

[18] Chin-Po Chen, Xian-Hong Tseng, Susan Shur-Fen Gau, and Chi-Chun Lee. 2017. Computing Multimodal Dyadic Behaviors During Spontaneous Diagnosis Interviews Toward Automatic Categorization of Autism Spectrum Disorder. *Interspeech 2017* (2017). https://doi.org/10.21437/interspeech.2017-563

[19] Rune Haubo Bojesen Christensen. 2024. Ordinal: Regression Models for Ordinal Data. https://cran.r-project.org/web/packages/ordinal/index.html

[20] Sophie Curtis. [n. d.]. *Pizza Hut Hires ROBOT Waiters to Take Orders and Process Payments*. The Mirror. http://www.mirror.co.uk/tech/who-needs-waiters-pizza-hut-8045172

[21] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. https://doi.org/10.48550/ARXIV.1306.6078

[22] Isabella Glans Diethelm, Sara Skov Hansen, Frederikke Birkeholm Leth, Kerstin Fischer, and Oskar Palinko. 2021-03-08. Effects of Gaze and Speech in Human-Robot Medical Interactions. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA) *(HRI '21 Companion)*. Association for Computing Machinery, 349–353. https://doi.org/10.1145/3434074.3447190

[23] Robin Dos Santos, Kai Eßmann, Niklas Fartmann, Heiko Meier, Eric Sandberg, Alexander Schulze, Sven Luzar, and Gernot Bauer. [n. d.]. Zoo Visitors' Initial Assessment of an Animaloid Robot as a Zoo Exhibit. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany, 2023-04-19). ACM, 1–7. https://doi.org/10.1145/3544549.3585762

[24] Bernd Dudzik, Simon Columbus, Tiffany Matej Hrkalovic, Daniel Balliet, and Hayley Hung. 2021. Recognizing Perceived Interdependence in Face-to-Face Negotiations through Multimodal Analysis of Nonverbal Behavior. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/3462244.3479935

[25] Connor Esterwood and Lionel P. Robert. 2022. A Literature Review of Trust Repair in HRI. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1641–1646. https://doi.org/10.1109/RO-MAN53752.2022.9900667

[26] Friederike Eyssel. 2017. An experimental psychological perspective on social robotics. *Robotics and Autonomous Systems* 87 (2017), 363–371.

[27] Justin W. Firestone, Rubi Quiñones, and Brittany A. Duncan. 2019. Learning from Users: An Elicitation Study and Taxonomy for Communicating Small Unmanned Aerial System States Through Gestures. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 163–171. https://doi.org/10.1109/HRI.2019.8673010

[28] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems* 42, 3 (March 2003), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

[29] Juan Miguel Garcia-Haro, Edwin Daniel Oña, Juan Hernandez-Vicen, Santiago Martinez, and Carlos Balaguer. 2021-01. Service Robots in Catering Applications: A Review and Future Challenges. *Electronics* 10, 1 (2021-01), 47. Issue 1. https://doi.org/10.3390/electronics10010047

[30] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 623–635. https://doi.org/10.1145/3196709.3196741

[31] Donald Glowinski, Sélim Yahia Coll, Naëm Baron, Maëva Sanchez, Simon Schaerlaeken, and Didier Grandjean. 2017. Body, Space, and Emotion: A Perceptual Study. *Human Technology* 13, 1 (2017), 32–57. https://doi.org/10.17011/ht/urn.201705272517

[32] Charles Goodwin and John Heritage. [n. d.]. Conversation Analysis. 19 ([n. d.]), 283–307. jstor:2155967 https://www.jstor.org/stable/2155967

[33] Elliott Hauser, Yao-Cheng Chan, Parth Chonkar, Geethika Hemkumar, Huihai Wang, Daksh Dua, Shikhar Gupta, Efren Mendoza Enriquez, Tiffany Kao, Justin Hart, Reuth Mirsky, Joydeep Biswas, Junfeng Jiao, and Peter Stone. [n. d.]. "What's That Robot Doing Here?": Perceptions Of Incidental Encounters With Autonomous Quadruped Robots. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (Edinburgh United Kingdom, 2023-07-11). ACM, 1–15. https://doi.org/10.1145/3597512.3599707

[34] Yutaka Hiroi and Akinori Ito. 2016. Influence of the Height of a Robot on Comfortableness of Verbal Interaction. *IAENG International Journal of Computer Science* 43, 4 (2016), 447–455. https://www.academia.edu/download/87236949/IJCS_43_4_06.pdf

[35] Zhiming Hu, Jiahui Xu, Syn Schmitt, and Andreas Bulling. [n. d.]. Pose2Gaze: Eye-Body Coordination During Daily Activities for Gaze Prediction From Full-Body Poses. ([n. d.]), 1–12. https://doi.org/10.1109/TVCG.2024.3412190

[36] Hochul Hwang, Hee-Tae Jung, Nicholas A Giudice, Joydeep Biswas, Sunghoon Ivan Lee, and Donghyun Kim. [n. d.]. Towards Robotic Companions: Understanding Handler-Guide Dog Interactions for Informed Guide Dog Robot Design. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024-05-11) *(CHI '24)*. Association for Computing Machinery, 1–20. https://doi.org/10.1145/3613904.3642181

[37] International Federation of Robotics 2021. *Service Robots.* International Federation of Robotics. https://ifr.org/service-robots

[38] Sofie Jackson. [n. d.]. *Robot Dog Waiter Brings a Round of Beers to Stunned Bargoers.* Daily Star. https://www.dailystar.co.uk/news/world-news/robot-dog-waiter-brings-round-22951867

[39] Keith Jones and Liz Schmidlin. 2011. Human-Robot Interaction: Toward Usable Personal Service Robots. *Reviews of Human Factors and Ergonomics* 7 (Sept. 2011), 100–148. https://doi.org/10.1177/1557234X11410388

[40] Jana Jost, Thomas Kirks, Stuart Chapman, and Gerhard Rinkenauer. 2019. Examining the Effects of Height, Velocity and Emotional Representation of a Social Transport Robot and Human Factors in Human-Robot Collaboration. In *Human-Computer Interaction – INTERACT 2019*, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 517–526. https://doi.org/10.1007/978-3-030-29384-0_31

[41] Hiroko Kamide, Yasushi Mae, Tomohito Takubo, Kenichi Ohara, and Tatsuo Arai. 2014. Direct Comparison of Psychological Evaluation between Virtual and Real Humanoids: Personal Space and Subjective Impressions. *International Journal of Human-Computer Studies* 72, 5 (May 2014), 451–459. https://doi.org/10.1016/j.ijhcs.2014.01.004

[42] Knut K. W. Kampe, Chris D. Frith, and Uta Frith. 2003. "Hey John": Signals Conveying Communicative Intention toward the Self Activate Brain Regions Associated with "Mentalizing," Regardless of Modality. *Journal of Neuroscience* 23, 12 (June 2003), 5258–5263. https://doi.org/10.1523/JNEUROSCI.23-12-05258.2003

[43] Dahyun Kang, Changjoo Nam, and Sonya S. Kwak. 2024. Robot Feedback Design for Response Delay. *International Journal of Social Robotics* 16, 2 (Feb. 2024), 341–361. https://doi.org/10.1007/s12369-023-01068-z

[44] Rachel Kirby, Jodi Forlizzi, and Reid Simmons. 2010. Affective social robots. *Robotics and Autonomous Systems* 58, 3 (2010), 322–332.

[45] Hwayeon Kong, Frank Biocca, Taeyang Lee, Kihyuk Park, and Jeonghoon Rhee. [n. d.]. Effects of Human Connection through Social Drones and Perceived Safety. 2018, 1 ([n. d.]), 9280581. https://doi.org/10.1155/2018/9280581

[46] Sonya S Kwak. 2014-05-31. The Impact of the Robot Appearance Types on Social Interaction with a Robot and Service Evaluation of a Robot. *Archives of Design Research* (2014-05-31). https://doi.org/10.15187/adr.2014.05.110.2.81

[47] Minae Kwon, Sandy H. Huang, and Anca D. Dragan. 2018-02-26. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago IL USA). ACM, 87–95. https://doi.org/10.1145/3171221.3171276

[48] Wing-ting Law, Ki-sing Li, Kam-wah Fan, Tiande Mo, and Chi-kin Poon. 2022. Friendly elevator co-rider: An hri approach for robot-elevator interaction. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 865–869.

[49] Christine P Lee, Pragathi Praveena, and Bilge Mutlu. 2024. REX: Designing User-centered Repair and Explanations to Address Robot Failures. *arXiv preprint arXiv:2405.16710* (2024).

[50] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. [n. d.]. Ripple Effects of an Embedded Social Agent: A Field Study of a Social Robot in the Workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin Texas USA, 2012-05-05). ACM, 695–704. https://doi.org/10.1145/2207676.2207776

[51] Songpo Li and Xiaoli Zhang. 2017. Implicit Intention Communication in Human–Robot Interaction Through Visual Behavior Studies. *IEEE Transactions on Human-Machine Systems* 47, 4 (Aug. 2017), 437–448. https://doi.org/10.1109/THMS.2017.2647882

[52] Yuhan Luo, Junnan Yu, Minhui Liang, Yichen Wan, Kening Zhu, and Shannon Sie Santosa. 2024. Emotion Embodied: Unveiling the Expressive Potential of Single-Hand Gestures. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. https://doi.org/10.1145/3613904.3642255

[53] Karthik Mahadevan, Mauricio Sousa, Anthony Tang, and Tovi Grossman. 2021-05-06. "Grip-that-there": An Investigation of Explicit and Implicit Task Allocation Techniques for Human-Robot Collaboration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan). ACM, 1–14. https://doi.org/10.1145/3411764.3445355

[54] Adwitiya Mandal and Kim Baraka. [n. d.]. Using Proxemics as a Corrective Feedback Signal during Robot Navigation. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2024-03-11) *(HRI '24)*. Association for Computing Machinery, 732–736. https://doi.org/10.1145/3610978.3640746

[55] Swathi Mannem, William Macke, Peter Stone, and Reuth Mirsky. 2023. Exploring the Cost of Interruptions in Human-Robot Teaming. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. 1–8. https://doi.org/10.1109/Humanoids57100.2023.10375236

[56] Emanuela Marchetti, Sophie Grimme, Eva Hornecker, Avgi Kollakidou, and Philipp Graf. [n. d.]. Pet-Robot or Appliance? Care Home Residents with Dementia Respond to a Zoomorphic Floor Washing Robot. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA, 2022-04-29). ACM, 1–21. https://doi.org/10.1145/3491102.3517463

[57] Marco Matarese, Alessandra Sciutti, Francesco Rea, and Silvia Rossi. 2021. Toward Robots' Behavioral Transparency of Temporal Difference Reinforcement Learning With a Human Teacher. *IEEE Transactions on Human-Machine Systems* 51, 6 (2021), 578–589. https://doi.org/10.1109/THMS.2021.3116119

[58] Derek McColl, Chuan Jiang, and Goldie Nejat. 2017. Classifying a Person's Degree of Accessibility From Natural Body Language During Social Human–Robot Interactions. *IEEE Transactions on Cybernetics* 47, 2 (Feb. 2017), 524–538. https://doi.org/10.1109/TCYB.2016.2520367

[59] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174

[60] Emily McQuillin, Nikhil Churamani, and Hatice Gunes. [n. d.]. Learning Socially Appropriate Robo-waiter Behaviours through Real-time User Feedback. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2022-03). 541–550. https://doi.org/10.1109/HRI53351.2022.9889395

[61] Shaunak A. Mehta and Dylan P. Losey. 2024. Unified Learning from Demonstrations, Corrections, and Preferences during Physical Human–Robot Interaction. *J. Hum.-Robot Interact.* 13, 3 (2024), 39:1–39:25. https://doi.org/10.1145/3623384

[62] Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2015. Learning Multimodal Behavioral Models for Face-to-Face Social Interaction. *Journal on Multimodal User Interfaces* 9, 3 (Sept. 2015), 195–210. https://doi.org/10.1007/s12193-015-0190-7

[63] Ajung Moon, Daniel M. Troniak, Brian Gleeson, Matthew K.X.J. Pan, Minhua Zheng, Benjamin A. Blumer, Karon MacLean, and Elizabeth A. Croft. 2014. Meet Me Where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI '14)*. Association for Computing Machinery, New York, NY, USA, 334–341. https://doi.org/10.1145/2559636.2559656

[64] Lucas Morillo-Mendez, Martien G. S. Schrooten, Amy Loutfi, and Oscar Martinez Mozos. 2024. Age-Related Differences in the Perception of Robotic Referential Gaze in Human-Robot Interaction. *International Journal of Social Robotics* 16, 6 (June 2024), 1069–1081. https://doi.org/10.1007/s12369-022-00926-6

[65] Marie Muehlhaus, Marion Koelle, Artin Saberpour, and Jürgen Steimle. 2023. I Need a Third Arm! Eliciting Body-based Interactions with a Wearable Robotic Arm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581184

[66] Katashi Nagao and Akikazu Takeuchi. 1994. Social Interaction: Multimodal Conversation with Social Agents. (1994).

[67] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[68] Linda Onnasch and Eileen Roesler. 2021. A Taxonomy to Structure and Analyze Human–Robot Interaction. *International Journal of Social Robotics* 13, 4 (July 2021), 833–849. https://doi.org/10.1007/s12369-020-00666-5

[69] Oskar Palinko, Kohei Ogawa, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2018. How Should a Robot Interrupt a Conversation Between Multiple Humans. In *Social Robotics*, Shuzhi Sam Ge, John-John Cabibihan, Miguel A. Salichs, Elizabeth Broadbent, Hongsheng He, Alan R. Wagner, and Álvaro Castro-González (Eds.). Springer International Publishing, Cham, 149–159. https://doi.org/10.1007/978-3-030-05204-1_15

[70] Xueni Pan and Antonia F. de C. Hamilton. 2018. Why and How to Use Virtual Reality to Study Human Social Interaction: The Challenges of Exploring a New Research Landscape. 109, 3 (2018), 395–417. https://doi.org/10.1111/bjop.12290

[71] William D. Perreault and Laurence E. Leigh. 1989. Reliability of Nominal Data Based on Qualitative Judgments. *Journal of Marketing Research* 26, 2 (May 1989), 135–148. https://doi.org/10.1177/002224378902600201

[72] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2016. Using Mobile Phones in Pub Talk. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1649–1661. https://doi.org/10.1145/2818048.2820014

[73] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal Human Discourse: Gesture and Speech. *ACM Trans. Comput.-Hum. Interact.* 9, 3 (Sept. 2002), 171–193. https://doi.org/10.1145/568513.568514

[74] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. The Influence of Height in Robot-Mediated Communication. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13)*. IEEE Press, Tokyo, Japan, 1–8.

[75] Isabel Benavente Rodriguez and Nicolai Marquardt. 2017. Gesture Elicitation Study on How to Opt-in & Opt-out from Interactions with Public Displays. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17)*. Association for Computing Machinery, New York, NY, USA, 32–41. https://doi.org/10.1145/3132272.3134118

[76] Juan-Carlos Rojas, Jaime Alvarez, Arantza Garcia-Mora, and Paulina Méndez. 2024. Comparison of Robot Assessment by Using Physical and Virtual Prototypes: Assessment of Appearance Characteristics, Emotional Response and Social Perception. In *Design, User Experience, and Usability* (Cham), Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares (Eds.). Springer Nature Switzerland, 127–145. https://doi.org/10.1007/978-3-031-61353-1_9

[77] Ofir Sadka, Jonathan Giron, Doron Friedman, Oren Zuckerman, and Hadas Erel. [n. d.]. Virtual-Reality as a Simulation Tool for Non-humanoid Social Robots. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020-04-25) *(CHI EA '20)*. Association for Computing Machinery, 1–9. https://doi.org/10.1145/3334480.3382893

[78] Jerrold M. Sadock. 1988. Speech Act Distinctions in Grammar. In *Linguistics: The Cambridge Survey* (1 ed.), Frederick J. Newmeyer (Ed.). Cambridge University Press, 183–197. https://doi.org/10.1017/CBO9780511621055.011

[79] Mauro Sarrica, Sonia Brondi, and Leopoldina Fortunati. 2019-04-12T00:00:00Z. How Many Facets Does a "Social Robot" Have? A Review of Scientific and Popular Definitions Online. *Information Technology & People* 33, 1 (2019-04-12T00:00:00Z), 1–21. https://doi.org/10.1108/ITP-04-2018-0203

[80] Azadeh Shariati, Mojtaba Shahab, Ali Meghdari, Ali Amoozandeh Nobaveh, Raman Rafatnejad, and Behrad Mozafari. [n. d.]. Virtual Reality Social Robot Platform: A Case Study on Arash Social Robot. In *Social Robotics* (Cham, 2018), Shuzhi Sam Ge, John-John Cabibihan, Miguel A. Salichs, Elizabeth Broadbent, Hongsheng He, Alan R. Wagner, and Álvaro Castro-González (Eds.). Springer International Publishing, 551–560. https://doi.org/10.1007/978-3-030-05204-1_54

[81] Abdel-Nasser Sharkawy and Panagiotis N. Koustoumpardis. 2022. Human–Robot Interaction: A Review and Analysis on Variable Admittance Control, Safety, and Perspectives. *Machines* 10, 7 (July 2022), 591. https://doi.org/10.3390/machines10070591

[82] Samantha Stedtler, Valentina Fantasia, Trond A. Tjøstheim, Birger Johansson, Ingar Brinck, and Christian Balkenius. 2024-03-11. Is There Really an Effect of Time Delays on Perceived Fluency and Social Attributes between Humans and Social Robots? A Pilot Study. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA) *(HRI '24)*. Association for Computing Machinery, 1013–1017. https://doi.org/10.1145/3610978.3640667

[83] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong-ah (Kate) Jeong, Shin'ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. 2019. Seven HCI Grand Challenges. *International Journal of Human–Computer Interaction* 35, 14 (Aug. 2019), 1229–1269. https://doi.org/10.1080/10447318.2019.1619259

[84] Steven J. Stroessner and Jonathan Benitez. 2019-04-01. The Social Perception of Humanoid and Non-Humanoid Robots: Effects of Gendered and Machinelike Features. *International Journal of Social Robotics* 11, 2 (2019-04-01), 305–315. https://doi.org/10.1007/s12369-018-0502-7

[85] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.

[86] Jonas Tjomsland, Sinan Kalkan, and Hatice Gunes. 2022. Mind your manners! a dataset and a continual learning approach for assessing social appropriateness of robot actions. *Frontiers in Robotics and AI* 9 (2022), 669420.

[87] Vivian Tseng and Edward Seidman. 2007-06-01. A Systems Framework for Understanding Social Settings. *American Journal of Community Psychology* 39, 3 (2007-06-01), 217–228. https://doi.org/10.1007/s10464-007-9101-8

[88] Inara Tusseyeva, Anara Sandygulova, and Matteo Rubagotti. [n. d.]. Perceived Intelligence in Human-Robot Interaction: A Review. 12 ([n. d.]), 151348–151359. https://doi.org/10.1109/ACCESS.2024.3478751

[89] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1325–1334. https://doi.org/10.1145/2702123.2702223

[90] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, Eindhoven Netherlands, 855–872. https://doi.org/10.1145/3357236.3395511

[91] Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F. Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, Dirk Heylen, Rene Kaiser, Maria Koutsombogera, Alexandros Potamianos, Steve Renals, Giuseppe Riccardi, and Albert Ali Salah. [n. d.]. Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. 7, 4 ([n. d.]), 397–413. https://doi.org/10.1007/s12559-015-9326-z

[92] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social Signals, Their Function, and Automatic Analysis: A Survey. In *Proceedings of the 10th International Conference on Multimodal Interfaces*. ACM, Chania Crete Greece, 61–68. https://doi.org/10.1145/1452392.1452405

[93] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 69–87. https://doi.org/10.1109/T-AFFC.2011.27

[94] Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect. *Proceedings of the National Academy of Sciences* 114, 25 (June 2017), 6521–6526. https://doi.org/10.1073/pnas.1702413114

[95] Ash Yaw Sang Wan, Yi De Soong, Edwin Foo, Wai Leong Eugene Wong, and Wai Shing Michael Lau. [n. d.]. Waiter Robots Conveying Drinks. 8, 3 ([n. d.]), 44. https://doi.org/10.3390/technologies8030044

[96] Haibin Yan, Marcelo H. Ang, and Aun Neow Poo. 2014. A Survey on Perception Methods for Human–Robot Interaction in Social Robots. *International Journal of Social Robotics* 6, 1 (2014), 85–119. https://doi.org/10.1007/s12369-013-0199-6

[97] H.A. Yanco and J. Drury. 2004-10. Classifying Human-Robot Interaction: An Upyeard Taxonomy. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, Vol. 3. 2841–2846 vol.3. https://doi.org/10.1109/ICSMC.2004.1400763

[98] J.D. Zamfirescu-Pereira, David Sirkin, David Goedicke, Ray LC, Natalie Friedman, Ilan Mandel, Nikolas Martelaro, and Wendy Ju. [n. d.]. Fake It to Make It: Exploratory Prototyping in HRI. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, 2021-03-08) *(HRI '21 Companion)*. Association for Computing Machinery, 19–28. https://doi.org/10.1145/3434074.3446909