Grounding behaviours with conversational interfaces: effects of embodiment and failures

Dimosthenis Kontogiorgos¹ · Andre Pereira¹ · Joakim Gustafson¹

Received: 1 November 2019 / Accepted: 1 February 2021 / Published online: 24 March 2021
© The Author(s) 2021

Abstract

Conversational interfaces that interact with humans need to continuously establish, maintain and repair common ground in task-oriented dialogues. Uncertainty, repairs and acknowledgements are expressed in user behaviour in the continuous efforts of the conversational partners to maintain mutual understanding. Users change their behaviour when interacting with systems in different forms of embodiment, which affects the abilities of these interfaces to observe users' recurrent social signals. Additionally, humans are intellectually biased towards social activity when facing anthropomorphic agents or when presented with subtle social cues. Two studies are presented in this paper examining how humans interact in a referential communication task with wizarded interfaces in different forms of embodiment. In study 1 (N = 30), we test whether humans respond the same way to agents, in different forms of embodiment and social behaviour. In study 2 (N = 44), we replicate the same task and agents but introduce conversational failures disrupting the process of grounding. Findings indicate that it is not always favourable for agents to be anthropomorphised or to communicate with non-verbal cues, as human grounding behaviours change when embodiment and failures are manipulated.

Keywords Conversational artificial intelligence · Grounding · Failures · Embodiment · Social robots · Smart speakers

1 Introduction

With conversational interface technology on the rise, several questions remain open on how humans engage with interactive agents in different forms of embodiment and social behaviour. Controlled using voice, these interfaces have changed the way we interact with technology in profound ways. Through an always present, screen-less and hands-free interface, users are encouraged to engage in embodied interactions and use natural forms of communication [23].

A wide range of interaction modalities has been designed and researched for conversational agents, that come in various forms such as smart speakers [4] and social robots [11]. However, by design, social robots provide additional modes of communication. Social robots interact using not

only speech but also non-verbal behaviour. By generating multimodal communicative behaviours such as gaze cues, facial expressions and gestures [12,59], social robots enable embodied contributions to common ground similar to how humans interact and establish mutual understanding [73]. *Grounding*, the coordinated process in which humans collaboratively establish mutual understanding [17], is also central in interactions between humans and machines, when conversation is the main interface. The medium of communication is an important factor to how common ground is established [17], and consequently robot embodiment and anthropomorphic elements may influence people's grounding behaviours.

In the fields of human-computer interaction and human-robot interaction, anthropomorphism is often leveraged as a way to make machines more 'comfortable' to use. The additional comfort comes from ascribing human features to machines with the aim to simplify the complexity of technology [56,60]. In addition, patterned social behaviours may facilitate social interaction with users, however, generating and interpreting these cues can induce higher levels of cognitive load [79]. Social robots do embody such behaviours, and provide the possibility of generating non-verbal social behaviours in their interactions with humans [26]. Many of

✉ Dimosthenis Kontogiorgos
diko@kth.se

Andre Pereira
atap@kth.se

Joakim Gustafson
jkgu@kth.se

¹ KTH Royal Institute of Technology, Stockholm, Sweden

Fig. 1 Illustrative image of the task and interactive setup used in the studies: a social robot guiding a human how to cook

these behavioural elements are subtle social cues (e.g. joint attention and mutual gaze), that are highly important for establishing common ground in situated human conversational environments. One reason why face-to-face interaction is preferred may be that a lot of familiar information is encoded in the non-verbal cues that are being exchanged (Fig. 1).

However, like any other interface, conversational interfaces are bound to fail in daily interactions with humans. These failures can be critical because they require human intervention and can cause users to lose trust in the agents' assistance and capabilities. The social environment that these interfaces are immersed in can cause one subset of failures that are defined as social failures, which can potentially lead to violations of social norms [37,58]. A lot of research approaches assume perfect interactions, and many critical failure aspects are often overlooked. Systems that interact with humans will inevitably have to deal with failures, user uncertainty and confusion. Nevertheless, in human-human communication, mistakes and imperfections can make humans more likeable and attractive [58]. However, little is known on how agent embodiment affects users' reactions when system failures disrupt the process of mutual understanding.

Additionally, research that has focused on robot failures often involves failures of low severity where little is at stake for participants, and less work has been done on failures where there is robot-induced risk for users. Failure severity can impact disruptions to common ground in human–robot interactions. In this work, we approach both *low severity* failures (little to no consequence of the failure) and *high severity* failures (more severe consequence of the failure).

Paper aims

While robot embodiment has shown to positively impact interactions with humans, it remains to be explored if this effect persists in displays of mutual understanding, and also

when the robot fails. This paper aims to contribute to this emerging field with two empirical evaluations using the elements of: (i) *embodiment and non-verbal behaviour* and (ii) *conversational failures* on changes in human grounding behaviours. In study 1, we examine whether a human-like face (social robot), capable of displaying non-verbal cues, shifts interactive behaviour in comparison to a voice-only assistant (smart speaker). To comprehend the effects of the comparison further, we test whether it is the human-like face or the non-verbal features that contribute to variability in behaviour and remove in a separate manipulation the non-verbal behaviour of the social robot. In study 2, we extend this work by studying the impact of the same factors on people's grounding behaviour with a robot in different task severities and when it fails. These two studies examine the following research questions:

- **RQ 1:** What are the effects in human grounding behaviour when manipulating **robot embodiment** and **social behaviour** during task-oriented dialogue?
- **RQ 2a:** How do different robot embodiments affect people's grounding behaviours after **conversational failures**?
- **RQ 2b:** Does **failure severity** interact with the above manipulations and with people's grounding behaviours?

2 Related work

2.1 Common ground

An essential aspect of human–robot collaboration is coordination in communication; natural language, eye contact, deictic gestures, are significant in embodied language grounding¹ between humans and machines. For effective collaboration, humans and robots need to establish, maintain and repair common ground in situated referential communication [17]. Furthermore, performance in social and collaborative situations depends fundamentally on the ability to detect and react to embodied social signals that underpin human communication [40]. These signals are complex and their coordination is achieved in both verbal and non-verbal forms.

To establish common ground, speakers work together with references to entities in the shared space of attention [20]. This requires a process of synchronisation with embodied contributions to the ground, where the listener needs to under-

¹ In the robotics community the symbol grounding problem [34] is prominent, however in this paper, we consider grounding as series of contributions in conversation [19]. The symbol grounding relates symbolic representations of language to mental phenomena, whereas Clark's contribution model, central in this paper, places grounding in the view that shared beliefs and knowledge coordinate how common ground is established moment by moment among speakers.

stand the utterance at the same time it is spoken and provide feedback or comply with the speaker's requests. Eye-gaze in particular, is a fundamental form of pragmatic feedback, in that the listener attends the speaker [19], and maintaining attention to the task, is the listener's signal of understanding [17,18]. According to Clark [18], as long as listeners' attention is undisturbed, they maintain positive evidence of understanding [25]. If listeners look confused or do not attend as expected, speakers will engage into correcting action. If grounding is not satisfied, conversational partners need to collaboratively resolve any failures that arise.

The impact of the medium for establishing grounding is also important. While face-to-face is the richest form of communication in humans [27], there are potential barriers in collaboratively establishing common ground with robots, in the same way human speakers do [13,36]. We can not assume that robot gaze will elicit the same responses from people in the referential process. Studies have shown that robot gaze is interpreted differently than human gaze [1]. For example, humans tend to look at the robot's face longer when referring to objects comparing to human speakers, indicating a concern on the robot's understanding [85]. Nevertheless, how robot embodiment affects the process of mutual understanding remains largely unexplored.

2.2 Robot embodiment

There seems to be an interest in literature on how different representations of physical embodiment and human-like features affect interaction performance and the perception of agents. Several studies have compared agents in digital screens to social robots [24,44,79] and they have shown that human-like agents that are physically co-located are generally preferred and are perceived to be more socially present than their virtually embodied versions [8,38,41,43,52] or remote video representations of the same agents [65,82]. Other studies have shown that social robots' perceived situation awareness is higher [54] and by adding non-verbal cues, the same agent is perceived as more socially present [32,64].

Anthropomorphising is to ascribe human-like features and characteristics to an otherwise non-human object and has become a common metaphor in the domain of computing [56]. Anthropomorphic features have been used in social robots to augment their functional and behavioural characteristics, and it has been argued that for interactions with humans, social robots need to be structurally and functionally similar to humans [26]. Using anthropomorphic features, agents provide a form of illusion, leading the user to believe that the agent is sophisticated in its actions. It has been shown that anthropomorphic robots with faces are better at establishing agency and at communicating intent [39].

Embodiment also influences people's willingness to comply with robots' requests. People are more likely to comply

with unusual requests from physically present robots than from robots present over live-video [9]. Additionally, in task-oriented interactions, people address a smart-speaker differently than a more anthropomorphic social robot in terms of visual attention. Humans tend to have a preference for social robots with contingent gaze behaviours, which may not always be a conscious choice. This also indicates that people may utilise different social mechanisms (e.g. turn-taking coordination) towards smart-speakers than to social robots [47,50].

However, it is not just the physical embodiment of the robot that has implications on its perceived intentions, but the behaviour and actions of the robot as well [78]. Research in HCI has advocated human-likeness and human-like coordination of verbal and non-verbal cues as the only way to convey human-like intelligence [14,15]. While most conversational interfaces communicate intent using language, social robots use verbal and non-verbal cues, and additionally encourage users to anticipate shared actions in the same space of attention [23]. Non-verbal behaviour is therefore used for communication, signalling and social coordination. The more human-like the agents' responses, the more they are attributed as social actors [56,62] and agents that do not use the rich set of social behaviours may evoke lesser feelings of mutual understanding. Research has shown that when artificial agents take advantage of human-like coordination of non-verbal behaviour, they are perceived to be more collaborative and intelligent [16,67].

2.3 Robot failures

As interactions with conversational agents are becoming increasingly common, it is more likely that people will encounter failures with these systems. It is therefore important to investigate how people's behaviours are affected when system failures cause misunderstandings [10,57]. Researchers have however reported mixed results in the effects of robot failure on people's behaviour and perception of the robot.

While faulty robots are perceived as less trustworthy and reliable, they do not always influence people's willingness to comply with robot requests [68,71]. Correia et al. [22] has found a decrease in trustworthiness when robots fail, however, the effect is mitigated if the robot attributes the failure to a technical problem. Mitigation strategies depend on several factors, such as the nature of the task [51], failure timing [53] and failure severity [61].

The effects of robot failures on robot perceptions are nevertheless not consistent. Robot failures can also positively affect user behavioural responses. Robots that exhibit erroneous behaviours in games engage users more [74]. Moreover, while erroneous robots are perceived to be less intelligent, competent and reliable, users perceive the inter-

actions as easier and more enjoyable [58,66]. Similarly, incongruent multimodal behaviour is rated as more human-like and likeable [70], indicating a preference for ‘imperfect robots’.

Research in HRI has also investigated how robot failures impact user behaviours, including patterns in eye-gaze, head movements, and speech - social signals that exhibit either established grounding sequences or implicit behavioural responses to failures [6,31,35,76,80]. Behavioural signals have also been examined at unexpected responses from human–robot interactions in the wild [5,30,75], with the use of social signals from low-level sensor input, to high-level features that represent affect, attention and engagement. Research has also showed that users tend to enact different behavioural responses to failures from human-like robots in contrast to smart-speaker embodiments [49].

Finally, less work has been done on failure in high severity situations, as this is difficult to convincingly simulate in a laboratory environment. In that direction, Morales et al. [61] studied people’s behaviour to robot failures that involve personal risk. Providing a human-like face showed to influence people’s willingness to help the robot. Additionally, people seem to trust less a robot that makes failures with severe consequences [69] and the consequence of the failure may affect how people attribute blame in severe failures [81].

3 Present paradigm

To examine the questions on the role of *embodiment* and *failures* in the process of grounding, we use a paradigm where messages are exchanged between conversational partners in a task-oriented setting. We defined a referential communication task of an instructional nature, where the speaker makes continuous task requests (by naming objects) that the listener needs to accomplish. We use the term speaker to indicate the conversational partner that initiates message requests (in this paradigm the conversational interface), and listener the recipient of the intended messages (the user).

We also make the assumption that in task-oriented dialogue, task actions² convey contributions to common ground. When a speaker makes a request (‘can you pass me the salt?’), a contingent compliance to that request is expected, likely with an acknowledgement of receiving the message (‘sure’, *-passes the salt*). Uncertainty or hesitations that either interrupt attention or cause delays in the task lead to problems in grounding. In such cases of miscommunication, the speaker

² In what Austin calls perlocutionary effects [7], rational and cooperative speakers will perform a requested action, assuming they have understood correctly an instruction utterance, and thereby satisfy the speaker’s communicative goal [29]. Complying to robot requests, user actions also represent turns, even when they are carried out in silence [28].

Fig. 2 The experimental paradigm used in the studies. For each new ingredient added in a cooking recipe, users had to ask the robot for instructions on how to proceed. Recipe ingredients were laid out on a table in front of the subjects. Distracting objects were used, and the recipes were unknown to participants

will need to repair and reformulate the message to help the recipient accomplish the intended task (‘it’s on your right’) [48]. Given each speaker request, listener’s actions are conditionally relevant, and expected to contribute to common ground, the mutual belief that the listener has understood what the speaker meant [19,72].

In the task, the speaker (robot) guides the listener (user) to complete cooking instructions (Fig. 2). The instructions are not trivial, therefore the user is dependent on the robot that has knowledge of the task. In this paradigm, we keep the task the same in both studies 1 and 2 and manipulate the robot embodiment and non-verbal behaviour (Study 1), as well as instructions that are either with or without failures (Study 2). To represent acceptance of a robot request (instruction), we make use of user acknowledgements, and as for understanding, we expect users to successfully comply to robot requests, represented in user motion. Other speech and eye-gaze features are also examined to represent turn-taking behaviours. In the rest of the paper we compare these two studies in which embodiment and failures are manipulated, and examine grounding behaviours in the directions of the research questions presented in Sect. 1.

4 Study 1: Robot embodiment

In order to investigate the impact of robot human-likeness and social non-verbal behaviour in grounding, we defined three embodied personal assistants (using two embodied conversational agents).

4.1 Experimental design

1. We utilised a *Smart Speaker [SS]* embodied conversational interface that only interacts with speech. A first generation

USER: [FINISHED-ACTION] So, what's next?
 ROBOT: [INSTRUCTION] Next, take three pieces of lettuce and put it in the spring roll.
 USER: [HESITATION] ...
 USER: [CLARIFICATION-Q] Ehm, where is the lettuce?
 ROBOT: [CLARIFICATION-A] The lettuce is the green thing in the middle.
 USER: [LOOKS-AT-INGREDIENT]
 USER: [STARTED ACTION] Uh, yes! Okay!
 USER: [ACTION-IN-PROGRESS]
 USER: [FINISHED-ACTION] Okay, what's next?

Fig. 3 An illustration of human–robot task-oriented dialogue

Amazon Echo was used, which was connected via Bluetooth and a TTS service similar to the default Echo TTS generated pre-scripted voice commands. Morphology: Cylinder speaker. Output modality: Voice.

2. We also used a *Robot without gaze behaviours* [ROBOT (NG)] as an embodied assistant in the form of a human-like robotic head. As SS, it uses speech to interact and no other modalities. A Furhat robot [3] was used, which was stationary and did not utilise any head or eye movements, and statically looked at the user. The robot had equivalent quality TTS to SS, speaking the same pre-scripted utterances. Morphology: Human-like back-projected face. Output modality: Voice.

3. We finally used a *Robot with gaze behaviour* [ROBOT], the same Furhat with pre-designed social gaze mechanisms, that also used voice for interaction. These included task-based functional behaviours such as gazing to objects during a referring expression and a turn-taking gaze mechanism. Morphology: Human-like back-projected face. Output modalities: Voice and head movement.

Using the three aforementioned agents, an exploratory within-subject user study was conducted to analyse the impact of human-likeness and non-verbal behaviour features. We manipulated two independent variables [*embodiment* and *social eye-gaze*], in three conditions [SS, ROBOT (NG), ROBOT], presented in different order to participants using a Latin Square. The following hypotheses were posed, towards the investigation of research question 1:

- **H1.** Similarly to how humans interact in face-to-face communication, when compared to the SS and the ROBOT (NG), the ROBOT will shift people's grounding behaviours by increased measures of attention and verbal behaviour.
- **H2.** While non-verbal behaviour should shift grounding behaviours, a human-like design without non-verbal cues should not induce the same differences. Differences in grounding behaviours should not apply between the SS and ROBOT (NG).
- **H3.** Complying to the task and task time should not be dependent on non-verbal cues or human-like design, as all agents utter the same unambiguous instructions.

Fig. 4 The WoZ operating room was situated in an adjacent room to the interactions and not visible to subjects

4.2 Task and supported dialogue

In order to avoid any misunderstandings on the task and the subjects' role, we began the interactions with a *control trial* with a human instructor. We then asked subjects to cook 3 variations of fresh spring rolls without providing the recipes; they had to get the recipes by interacting with the agents. Different varieties of ingredients and amounts were used (Fig. 2). The experiment setup also included ingredients not used in any of the recipes, encouraging participants to interact with the agents to find out the correct ingredients for each recipe. The task was the same in each condition, but different recipes were used. We had a total of 20 ingredients and a recipe typically included 7 ingredients to prepare.

All agents used a combination of nouns, adjectives and spatial indexicals as linguistic indicators to identify ingredients on the table, “The *cucumber* is the *green* thing on the *right*” (Fig. 3). The ROBOT with gaze however, also gazed at the referent ingredients (0.5 s prior to the reference). The agent's role in the task was therefore to instruct and the subject's role was to assemble the ingredients together.

Participants were led to believe that the robot was autonomous. However, to dismiss potential problems in speech recognition and language understanding, we used a human wizard (WoZ) to control the behaviours of the agents (Fig. 4). The human wizard selected the appropriate agent response, as triggered by user speech. The WoZ application and dialogue policies were the same across conditions and wizards were not able to deviate from the interaction protocol, but only use pre-defined dialogue options. For every dialogue act, a set of predefined utterances was available, that the system would choose at random to generate, given the current dialogue act in the task. The WoZ therefore indicated the current dialogue act in conversation, and not what to say.

Human wizards had the following dialogue options in response to user dialogue acts: (a) [next instruction] the user has finished the current step of the task or has requested the next ingredient, (b) [clarification answers] if users asked for clarification, the agents would provide additional task-based

information by replying to ‘what/where’ is an ingredient, ‘how much’ of an ingredient should be taken, confirmations with ‘yes/no’ answers, and (c) [repeat] the previous instruction. Users were not aware of dialogue options, but would interact with the agents to find out. In rare cases, if a user deviated from the interaction protocol (e.g. ‘what’s the meaning of life?’), the robot uttered ‘I am sorry, I do not understand’ and moved on to the next instruction. Finally, when users selected wrong ingredients, the robots indicated an [incorrect] action.

In order to facilitate a natural turn-taking mechanism, we defined a heuristic gaze model for the gaze ROBOT with pre-determined timings for *turn-taking gaze* and *referential gaze* to objects which is important in directing interlocutors’ attention [2,33,55]. The gaze ROBOT therefore engaged in *mutual gaze* and *joint attention* with the subjects during the interactions. Before an utterance, the robot made a gaze shift to the subject to establish attention, followed by deictic gaze to a referent object indicating it is keeping the floor, and at the end of the utterance a gaze shift back at the participant to establish the end of the turn [40,63,77].

4.3 Participants and procedure

We recruited 30 participants (18 female and 12 male) with ages in range 19–42 and mean 24.2. 17 had interacted with a robot before and 20 with a smart speaker. 13 had interacted with both a robot and a smart speaker before, while 6 with none of the two. Overall, their experience with technology was 4.8 from 1 to 7 (stdev = 1.6). Participants signed a consent form and were instructed that they are able to stop the experiment at any time. They were compensated with a cinema ticket and the food they cooked during the study.

Participation in the study was individual. First, participants filled a demographics questionnaire and then cooked the first recipe with a human instructor. Then, they cooked a recipe with the help of one of the agents. They repeated that phase 3 times with a new agent every time (counter-balanced), and at the end of the study they filled an end-questionnaire. During the trials, participants were alone in the room, and the WoZ was monitoring their actions using a ceiling camera with a live feed. Participants were not told that the agents were controlled by a human wizard, until the end of the study.

Participants were not asked to finish the task with any time pressure, to ensure space for socially interacting with the agents. The human instructor was kept the same for all subjects, and followed the same behaviour and dialogue policy as the agents. The subjects stood in front of a table, with a cutting board and ingredients prepared and laid out in front of them (Fig. 2) and the agent was situated on the side of the table. The ingredients were fixed in place and the order of them remained consistent throughout the experiment.

4.4 Measures

In order to evaluate grounding behaviours with the spoken dialogue agents, we used task-based behavioural measures such as gaze, task time and conversational features. As we manipulated the agents’ embodiment and attentional capabilities, we expected to find differences in conditions on attentional and conversational cues.

We extracted the following behavioural measures that represented user behaviour to robot requests: *Proportional gaze to the agent*: We measured subjects’ gaze using their head pose direction (automatically annotated from a motion capture system [46]), which should indicate subjects’ attention. *Number of conversational turns*: The number of turns the agent responded to human turns (extracted from agent logs). *Clarification questions*: We used the number of times the agent answered clarification questions (extracted from agent logs), to indicate different levels of understanding agent instructions. *Interaction time*: We measured the *task time* from the beginning to the end of the interaction (extracted from agent logs) to count the amount of time subjects engaged with the agents. *Acknowledgements*: we manually annotated user acknowledgements (‘sure’, ‘okay’) right after each agent instruction to compare how often subjects accept a robot message before carrying on to the task. While an acknowledgement represents message acceptance, it does not indicate understanding [17]. As such, we also extracted *head movement* using the motion capture system, to represent user motion³: when users are working on the task there should be more movement⁴ (accumulated in meters) within robot utterances, while confusion and misunderstanding, combined with scanning of the visual scene, can cause lack of movement [31,80] (Fig. 5) or engagement [83].

4.5 Results

We detected subjects’ head pose over time and extracted gaze duration to the agent and the task during agent instructions. Proportional gaze to the agent is reported. Each phase is first normalised per subject to reduce subject variability and then, each interval mean is used for comparison.⁵ A repeated mea-

³ At the end of each action most users also verbally expressed they have completed the current step of the task. 27 out of 30 participants, in 72.3% of all cases subjects gave verbal feedback, without significant differences among embodiment ($p = .886$).

⁴ We would like to clarify the difference between the use of head movement and head direction, that represent here two separate behaviours. Little to no motion of head with a lot of variance in head direction, should indicate hesitation and scanning of the visual scene. While the opposite, variance in motion and steady (with little variance) head direction should indicate a determined course of action.

⁵ Statistical tests were conducted using IBM SPSS. An alpha level of .05 was used for all statistical analyses (automatically corrected for multiple comparisons). Same applies further in Study 2.

Fig. 5 We extracted head movement to represent user motion. More movement should indicate subjects are working on the task according to robot requests, while lack of movement should represent hesitation (as shown in figure when asked to pick up lettuce)

Fig. 6 Proportional gaze to agent during agent instructions. Error bars indicate standard error of the mean ($n = 30$)

sures ANOVA to test the effect of *gaze* showed a significant main effect, $F(2,28) = 18.07$, $p < .001$). Post-hoc tests with a Bonferroni correction, and p -value adjusted for multiple comparisons, revealed that gaze towards ROBOT (.47) is statistically greater than gaze to SS (.31, $p < .001$) and ROBOT (NG) (0.33, $p < .001$). No other statistical differences were found in pairwise comparisons (Fig. 6).

A repeated measures ANOVA on the *number of conversational turns* showed a significant main effect: $F(2,28) = 5.23$, $p = .012$). Post-hoc tests with a Bonferroni correction revealed that conversational turns with ROBOT (21.23) are statistically greater to ROBOT (NG) (19.40, $p = .036$) and to SS (18.53, $p = .033$). No statistical differences were found between the other two conditions (Fig. 7).

When compared across conditions, repeated measures ANOVA tests revealed significant differences among the three conditions on the *number of clarification questions* $F(2,28) = 4.83$, $p = .016$). Post-hoc pairwise tests with Bonferroni correction were carried out for the three pairs of groups. The results indicated a significant difference between SS (2.5) and ROBOT (4.0, $p = .019$) (Fig. 8). There were no other statistical differences.

We trivially found that *task time* was correlated with the number of *conversational turns* ($r = .654$, $p < .001$), and the number of *clarifying questions* ($r = .566$, $p < .001$).

Fig. 7 Number of conversational turns across agents. Error bars indicate standard error of the mean ($n = 30$)

Fig. 8 Clarification questions across agents. Error bars indicate standard error of the mean ($n = 30$)

Fig. 9 Total interaction time. Error bars indicate standard error of the mean ($n = 30$)

We tested for comparison the sequence of the task, and no statistical difference was found, meaning that the task sequence did not affect task performance. However, when compared across conditions, a repeated measures ANOVA showed a significant effect in interaction time $F(2,28) = 4.94$, $p = .014$. Post-hoc tests with a Bonferroni correction revealed that interaction time with ROBOT (232.93 s) is statistically greater to ROBOT (NG) (217.26s, $p = .023$) and to SS (212.66 s, $p = .041$). No other statistical differences were found (Fig. 9).

A comparison across conditions in user *acknowledgements* with repeated measures ANOVA tests revealed significant differences among the three conditions, $F(2,28) = 3.41$, $p = .043$. Post-hoc pairwise tests with Bonferroni correction were carried out for the three pairs of groups. The results indicated a significant difference between SS (.50), ROBOT (NG) (1.06) and ROBOT (1.00) (Fig. 10).

Finally, subjects' *head movement* between agent utterances showed a significant main effect, $F(2,28) = 4.42$, $p = .019$. Accumulated head movement with ROBOT (NG) (2.24 m) was lower than SS (2.46 m) and ROBOT (2.46 m) (Fig. 11).

5 Study 2: Robot failures

5.1 Experimental design

In study 1 we saw large differences between SS and the gaze ROBOT, however ROBOT (NG) was similar to SS in most of the behavioural measures and similar to the gaze ROBOT in measures such as acknowledgements. We assumed that when a human-like face is presented, human-like coordination of non-verbal cues is expected too, as seen in our findings, and therefore removed the ROBOT (NG) condition in study 2 to limit the number of trials across subjects.

We also added more steps in the interaction in order to introduce robot instructions that include failures and induce situations of misunderstanding. This means that each interaction in study 2 takes longer as more robot requests (instructions) are implemented in comparison to study 1. The robot instructions implemented were nevertheless the same in studies 1 and 2. We otherwise kept the task the same, and also the devices (using the same TTS), gaze behaviour and human trial in the beginning of every interaction. Subjects that took part in study 2, had not taken part in study 1 and were therefore new to the task.

With the addition of the variable of conversational failures, we attempted to replicate results from study 1, and in light of general findings, we discuss human grounding behaviours under different experimental conditions. We expected that under the same conversational and attentional measures, subjects should display different grounding behaviours when

Fig. 10 Average number of acknowledgements by users after robot instructions. Both human-like embodiments showed to significantly increase user acknowledgement behaviour

Fig. 11 Accumulated head movement in meters across embodiment between agent instructions. Error bars indicate standard error of the mean

there are misunderstandings and disruptions of common ground, yet they should maintain similar behaviours (to Study 1) when no failures occur.

Conversational failures. We used a set of failures, informed by taxonomies of failures in previous studies in HRI [37]. These induced failures in the interactions represented typical robot malfunctions that have been reported in human–robot interactions, and they are either task-oriented (giving incorrect guidance), or failures that violate social protocols of interaction (not responding) [37]. All failures had the consequence of delaying users in completing the task:

Disengagement. The system simulates ‘losing’ user engagement and restarts the interaction. It utters the welcome message when a new user has entered the task and fifteen seconds after the failure has occurred it becomes responsive and continues the guidance.

Incomplete instruction. In this failure the robot times speech improperly by producing an incomplete instruction, and after a short delay continues its utterance.

No response. The robot simulates lack of user speech input by not responding for 20 seconds.

USER: [FINISHED-ACTION] Okay I'm ready, what's next?
 ROBOT: [INSTRUCTION] Next, take three pieces of cucumber
 and put it in the spring roll.
 USER: [ACTION-IN-PROGRESS]
 USER: [FINISHED-ACTION] Mhm, what's next?
 ROBOT: [INSTRUCTION] It is time to put some cucumber on
 your spring roll.
 USER: [LOOKS-AT-INGREDIENT]
 USER: [LOOKS-AT-ROBOT] But I just.. again?
 USER: [LOOKS-AT-TIMER]
 USER: [ACTION-IN-PROGRESS]
 USER: [FINISHED-ACTION] Okay I did it, what's next?

Fig. 12 An illustration of a human–robot dialogue in the task when a [REPEATING] type of failure occurs

Repeating. The robot repeats a previous statement by asking the user to perform (again) the previous instruction (example in Fig. 12).

Incorrect guidance. The robot produces an erroneous instruction by asking the user to pick a non-existing object (ingredient).

Both agents were designed to not display any awareness that they have failed or apply any error recovery strategies. When users asked for clarifications trying to resolve failures the agents would simulate not understanding and prompt the user to continue to the next instruction. This would ensure to leave certain parts of the task ‘ungrounded’ until the end of the interaction, yet subjects would still be able to proceed to the next steps. To obtain as similar circumstances as possible, the order of failure stimuli was predetermined per interaction in 2 sequences that were counter-balanced per embodiment.

Failure severity. Another factor that can affect human–robot collaboration in guided tasks is time pressure. In this study, we introduced time pressure with a timer on a computer screen right next to the task. We expected that under time pressure, the same failures would have a higher severity on the task and would influence users’ behaviour with the robots that refrain them from an anticipated reward. Only half of the participants in this study were introduced to time pressure in the task, which is therefore introduced as a between design factor. Participants were rewarded with a cinema ticket for their participation, however participants under time pressure were told that they would receive one extra cinema ticket if they would finish the task on the top 20% fastest of all previous interactions. Subjects were debriefed at the end of the study that this was part of the experiment manipulation. In sum, participants that experienced failures under time pressure are in the ‘high severity’ condition, while participants that had no time pressure experienced failures of ‘low severity’.

To examine the relative effects of the two independent variables of *embodiment* (smart-speaker and social robot) and *failure severity* (low and high) a 2×2 mixed-design was used. Specifically, robot embodiment was conducted within subjects and failure severity was manipulated between subjects. All participants interacted with both robots (SS and ROBOT), counter-balanced in order. Participants were ran-

domly assigned to either the low or high severity of failure condition but stratified by gender.

To examine RQ2a and RQ2b we posed these hypotheses:

- **H4.** Similarly to Study 1, the ROBOT will shift people’s grounding behaviours by increased measures of attention and verbal behaviour, in addition to when failures occur, as subjects will attempt to resolve failures.
- **H5.** Subjects will shift their grounding behaviours by decreased attention and verbal behaviours when in time pressure (high failure severity).
- **H6.** Complying to the task and task time will be dependent to failures and failure severity.

5.2 Task and supported dialogue

The task and dialogue, wizard protocols, and the ROBOT’s gaze behaviour remained the same as in study 1. One difference was on the wizard decision to proceed to the next step of the interaction. As mentioned in study 1, we noticed that almost all participants verbally requested from the agents to proceed to the next instruction once they finished the requested action. The wizard in study 1 would proceed to the next instruction once a user action was complete (with or without verbal clarification). In contrary, we instructed the wizard in study 2 to only proceed when subjects have explicitly and verbally requested for the next instruction (‘what is next?’), giving the impression of an autonomous system.

5.3 Participants and procedure

44 participants (26 reported female and 18 reported male) were recruited via mailing lists and were rewarded with a cinema ticket for participation. The average age was 26.6 in the range 22–37. Participants signed a consent form before participation and were instructed that we study the impact of smart technologies and robot communication in instructions. 34 had interacted with a smart speaker before and 31 with a robot. The procedure of the experiment was the same as in Study 1 with the difference that Study 2 participants interacted with 2 agents instead of 3. A trial with a human instructor was also introduced before interactions with the agents. The experiment took place in the same room as in Study 1 and with the same conditions and equipment.

5.4 Measures

Using the ELAN annotation software [84], we manually segmented parts of each interaction into failure and no-failure. In these time segments, we extracted temporal measures from users’ behavioural data, similar to the measures reported in Study 1. *Gaze:* Using motion-capture, we collected participants’ gaze annotated by their visual angle and measured

Fig. 13 Proportional gaze to agent during instructions. Error bars indicate standard error of the mean (N = 44)

Fig. 14 Number of conversational turns. Error bars indicate standard error of the mean (n = 44).

proportional amount of gaze towards the robots *during* instructions. *Number of conversational turns*: As in study 1, we extracted the number of turns the agent responded to human turns (extracted from agent logs). *Clarification questions*: The number of times the agent answered clarification questions (extracted from agent logs). *Interaction time*: The *task time* to count the total interaction time with each agent (extracted from agent logs). As a manipulation check, we expected that high severity participants would be faster with time pressure and an anticipated reward at stake. *Acknowledgements*: we manually annotated user acknowledgements right after each agent instruction to compare how often subjects accept a robot message before carrying on to the task. *Head movement*: Finally, we also extracted head movement to represent user motion (accumulated in meters).

5.5 Results

In this section, we present results from non-verbal coordination and verbal performance from users' behavioural measures. Due to sensor errors, data is missing from two subjects (one in each severity condition). Note that in this study we utilise a mixed design with the use of two within design factors (*embodiment* and *failures*) and a between design factor (*severity*), in contrast to Study 1 where only one within design factor was measured (*embodiment*).

Fig. 15 Number of clarification questions. Error bars indicate standard error of the mean (n = 44).

Fig. 16 Interaction time. Error bars indicate standard error of the mean (n = 44)

Three-way ANOVAs on *gaze to agent* showed significant main effects in the factors of *embodiment* and *failure*. Bonferroni corrected pairwise tests revealed that proportional gaze to agent is higher with the ROBOT ($p < .001$), but also higher when failures occur ($p = .002$). These results replicate gaze data shown in Study 1, but also indicate that robot failures affect human gaze grounding behaviours. Failure severity did not affect proportional gaze ($p = .886$). An overview of the gaze data is presented in Table 1 and Fig. 13.

A repeated measures two-way ANOVA on *conversational turns* showed no significant main effect: $F(1,39) = .086$, $p = .771$, $\eta^2 = .002$ across embodiment and neither on failure severity: $F(1,39) = .398$, $p = .532$, $\eta^2 = .010$ (Fig. 14).

Similarly, a repeated measures two-way ANOVA on the number of *clarification questions* showed no significant main effect: $F(1,39) = .023$, $p = .881$, $\eta^2 = .001$ across embodiment. No significant effects were found failure severity either: $F(1,39) = 3.384$, $p = .073$, $\eta^2 = .080$ (Fig. 15).

Applying a two-way ANOVA on *embodiment* and *severity*, we observed that the manipulation of time pressure caused a significant effect on the *interaction time*: $F(1,39) = 12.720$; $p = .001$; $\eta^2 = .246$. Bonferroni corrected pairwise tests revealed that participants spent less time in the high severity condition (352.0 s, SD = 13.4) than in the low sever-

Table 1 Results from a three-way mixed ANOVA on the users' gaze data with 2 within factors (embodiment and failure) and 1 between factor (severity)

Measure	[Embodiment]	[Failure]	[Embodiment * severity]	[Failure * severity]	[Embodiment * failure]
df	(1,40)	(1,40)	(1,40)	(1,40)	(1,40)
F	40.677	10.985	1.049	.296	2.046
<i>p</i> value	< .001	.002	.312	.590	.160
η^2	.504	.215	.026	.007	.049

Significant differences indicated in bold

Fig. 17 Number of acknowledgements after agent instructions. In both embodiments when robots failed, subjects showed a significant decrease in acknowledgement behaviour **Fig. 18** Accumulated head movement across embodiment between agent instructions. Error bars indicate standard error of the mean

ity condition (419.2 s, SD = 13.1) ($p = .001$); participants were indeed rushing to finish faster the task, indicating our failure severity manipulation caused by time pressure was successful. No significant effect was found across embodiment ($p = .166$) (Fig. 16).

A comparison across conditions in user *acknowledgements* with three-way mixed ANOVA tests revealed significant differences among the factor of failure, $F(1,42) = 7.409$, $p = .009$. Post-hoc pairwise tests with Bonferroni correction indicated that subjects uttered more acknowledgements when no failures occurred (1.4, STDERR = .24), while they hesitated to utter acknowledgements when failures occurred (0.9, STDERR = .12) ($p = .009$) (Fig. 17). No significant differences were found among embodiment ($p = .304$) or among failure severity ($p = .521$).

Finally, a three-way mixed ANOVA showed a significant main effect on subjects' *head movement* on the factor of failures, $F(1,40) = 14.934$, $p < .001$. When robots failed, subjects hesitated to take actions and moved less (1.22 m,

STDERR = .06) in contrast to no failures (1.4 m, STDERR = .05) ($p < .001$). An interaction effect was also observed among failure and failure severity, $F(1,40) = 24.540$, $p < .001$ (Fig. 18). No effects of embodiment ($p = .776$) were significant.

6 Discussion

6.1 Common ground

In two experiments with human subjects interacting with conversational interfaces, we found variability in human grounding behaviours to robot instructions when embodiment and robot failures were manipulated. In each instruction, the agents posed instructions which would remain ungrounded until subjects have complied to the agent's request. Behavioural responses were measured with a variety of multimodal features. Whether subjects accepted the agent message (represented in acknowledgements), asked for clarification, or comply to the request (represented in movement), seemed to be dependent on the agent embodiment or the failure of the agent to provide a reliable and well grounded instruction.

Utilising a referential communication task meant that discourse between humans and the agents would be based on referent objects and how to establish their referential identities. This constrained nature of the task allowed us to keep consistency in robot and human behaviour across conditions and across studies. How subjects attributed their attention to the agents, or acknowledge they have received their messages also seems to be dependent on their intrinsic motivation to complete the task, manipulated with time pressure in Study 2. Sometimes a message may be accepted with an acknowledgement (with back-channel responses), or with continuous attention [19]. In task-oriented dialogues however, successful completion of the task is strong evidence of understanding, even with the absence of these social signals [20,21].

Further in this section, we address these topics based on the research questions formulated in the beginning of this article. In particular we discuss RQ1 in Sect. 6.2 on how robot embodiment affected mutual understanding and questions

RQ2a and RQ2b in Sect. 6.3 on the effects of conversational failures in grounding behaviours.

6.2 Robot embodiment

The agents we compared, represent different levels of embodiment in conversational agents. Dialogue with the gaze ROBOT condition in Study 1 was longer in conversational turns in comparison to the less anthropomorphic⁶ SS. It is interesting to mention however, that most participants were more familiar with smart speakers than they were with social robots, which could indicate a novelty effect while interacting with the agent. Social robots are at time of writing still emerging platforms and not as common and commercially available, as smart speakers are. In both studies, behavioural data indicate that users do change their behaviour with a human-like robot, as shown in their increased proportional gaze towards the robot and conversational styles. Intuitively, participants are unaware of their increased measures of attention, yet they still exhibit reactive communication traits typically seen in human-human grounded communication.

In Study 1, we expected to find no differences in grounding behaviour between SS and ROBOT (NG) [H2]. Our assumption was that anthropomorphic face features, without non-verbal behaviours would not be enough to create more socially contingent interactions than SS: it is a combination of the two features that facilitate the notion of mutual understanding with users. In line with our initial expectations, we did not observe any statistical differences in eye gaze, conversational turns and interaction time when comparing SS with ROBOT (NG). However, we had some conflicting results that contradicts this initial expectation. Favouring both ROBOT embodiments, a statistically significant difference in the average number of acknowledgements was found, indicating that an anthropomorphic agent stimulates face-to-face grounding behaviour to a greater extent when compared to a less anthropomorphic one, even without non-verbal behaviours.

Looking at the accumulated head movement, the ROBOT (NG) agent stimulated significantly less head movements in participants when compared to both the ROBOT and SS [H3], indicating that a lack of gaze behaviours in combination with an anthropomorphic body can actually be counterproductive to stimulate non-verbal grounding behaviours as well. It appears that a human-like design is not enough to fully establish common ground with humans; human-like coordination may be expected as well [27], when anthropomorphic designs are manifested. Our assumption is that the gaze ROBOT

has *joint attention* afforded as an embodied phenomenon in its actions, giving the impression it is aware on the situatedness of the task. Conversational interfaces come without an instruction manual, as Kiesler suggests [45], with little time for learning what the agent can do. Its appearance and behaviour will create expectations over its capabilities and intentions [H1].

Eye-gaze here is therefore attributed as a social function where it regulates turn-taking, closer to how humans do when they interact with each other by showing the speaker they are still attending. Smart speakers, while embodied, do not facilitate the same grounding mechanisms as social robots with non-verbal behaviour, likely due to the lack of eye-gaze and other non-verbal behaviours. While social robots resemble human face-to-face conversations, smart speakers approach human conversational dynamics with reduced channels of communication, similar to computer-mediated communication [13,21,36], or conversations over the phone.

Despite the grounding benefits seen in human-like agents, there is controversy if they add interaction value in all task-oriented dialogues. In some tasks, users may prefer guidance without social and non-verbal signals. This may explain expressed feelings of preference over the lack of social cues from smart speakers, observed in participants' reports:

"I preferred the [ROBOT] as it instructed me as a human does.. But I think the [SS] is best when you just want things done, and have minimum interaction.."

"The [SS] is the least intrusive I would say, if just cooking, I would prefer this one.. Social robots may be good for someone who seeks interaction or for children."

6.3 Robot failures

In Study 2, we compared two robot embodiments in different types of failure and failure severity situations to examine grounding behaviours with conversational agents. We did find differences in users' gaze reactions to failures; participants looked at the ROBOT longer during instructions, as in Study 1, however when failures occurred, gaze to ROBOT was further increased in comparison to SS [H4]. Intuitively, the turn-taking gaze mechanism might be invoking in subjects an attempt to establish grounding via attention in cases of failure. It was also apparent that subjects looked at the agent in Study 1 when they required more information. In Study 2 they did so as well, but they also gazed towards the agent when there was a failure that needed to be resolved.

User acknowledgements followed the reverse trend in failures, as participants acknowledged the agents' instructions to a higher extent, when no failures occurred. No significant interaction effect with failure severity was found either with gaze or user acknowledgements [H5]. It is important to note that acknowledgements are also subject dependent. Some

⁶ Any conversational interface is more or less anthropomorphic when interacting with speech. While anthropomorphism is not a binary degree of human-likeness, we maintain in this paper the notion of human-likeness mainly in physical characteristics, and assert that the social robot in these studies is more anthropomorphic than the smart speaker.

subjects tend to use verbalised acknowledgement mechanisms while others only display understanding through task actions. This was more apparent as acknowledgements also existed when failures occurred. Subjects gave acknowledgements even when they were not able to satisfy the requested action (i.e. in the process of identifying a non-existing ingredient). Additionally, we did not see significant differences in Study 2 on conversational turns or the number of clarification questions among embodiment. This may have been skewed in the attempts to establish mutual understanding after continuous conversational failures that did not exist in Study 1.

Moreover, in low failure severity interactions, we found significantly less head movement when failures occurred, as participants might have gotten confused by the agent's behaviour in both embodiment conditions and therefore focusing less in the task. In high severity, participants were not affected by failures as much and relentlessly continued attempting to complete the task equally in both embodiment conditions, even if instructions still remained ungrounded. Overall, high failure severity participants had less movement in their actions, performing more fast and precise actions to finish the task as quickly as possible. Low severity participants however did show hesitation in their movement, indicating they spend more time to resolve misunderstanding before moving on to the task [H6].

It is possible the system in high severity failures might have distracted the user from the task by displaying additional social behaviours [42], as more attention needed to be given to the system. In such cases, users may want to get the task done as quickly as possible, and potentially get frustrated when having to speak longer than necessary:

“The [ROBOT] had a human face and was a bit more distracting than the smart speaker. It was easier to focus on the task with [SS].”

“The [ROBOT] was much more distracting than just listening to the instructions.”

7 Conclusion

In this paper, we discussed how grounding behaviours are shaped when empirically controlling for embodiment and failure parameters on guided tasks with conversational agents. This is particularly important to applications in which socially interactive agents engage in a variety of tasks, and depending on the nature of the task, agents may benefit from more anthropomorphic embodiments in the process of grounding. Failures in interactions with humans are inevitable, and there is already a research focus on how to avoid misunderstandings by improving systems' sensory equipment as well as language understanding capabilities. We can see however in our findings that other parameters

such as the agent's physical appearance also contribute in what behavioural responses the system should attend to in its continuous efforts to maintain mutual understanding. Future robots should inevitably be human-centred but not always human-like.

Tying these findings to the agents' differences in embodiment is of course one possible interpretation. To understand which of the variables contributed to the general behavioural differences with the robot, we concluded that while an anthropomorphic physical embodiment increases subjects' gaze and speech features, agent non-verbal behaviours are also expected when human-like embodiments are manifested. We also saw that subjects' social behaviours are not coming solely from the chosen dialogue and speech synthesis, but rather from simulating visual attention (joint attention + mutual gaze) with a more anthropomorphic embodiment.

It is also important to mention that while in Study 2 conversational failures appeared by design, misunderstandings also happened when no failure was designed to take place, and similarly in Study 1. Misunderstandings are interactional phenomena, therefore uncertainty, clarification requests and repairs will occur even in perfectly executed and non-ambiguous instructions. In the studies presented we attempted to resolve such misunderstandings by designing a simple task-based clarification request mechanism, yet a lot of user uncertainty may have been nevertheless unresolved.

Future research should be conducted in different HRI and task-oriented settings, to investigate variability in the nature of the task and failures and its relation to social engagement between humans and agents. In sum, situation aware social robots hold a good interaction paradigm for enabling improved social interactions with users. Focus should be given on how these findings can be best applied to designing robots for guided tasks that will inevitably have to deal with failures and uncertainty in interactions with humans.

Acknowledgements This research was funded by the Swedish Foundation for Strategic Research (GMT14-0082). We would like to thank Iolanda Leite, Sanne van Waveren, Olle Wallberg, Olle Andersson, Marco Koivisto, Elena Gonzalez Raval and Ville Vartiainen for their contributions, and the 74 participants that cooked 252 spring rolls. We would also like to thank the anonymous reviewers for contributing to the current version of the paper.

Funding Open access funding provided by Royal Institute of Technology.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Admoni H (2016) Nonverbal communication in socially assistive human–robot interaction. PhD thesis. Ph.D. Dissertation, Yale University
- Admoni H, Scassellati B (2017) Social eye gaze in human–robot interaction: a review. *J Hum Robot Interact* 6(1):25–63
- Al Moubayed S, Beskow J, Skantze G, Granström B (2012) Furhat: a back-projected human-like robot head for multiparty human–machine interaction. In *Cognitive behavioural systems*. Springer, pp 114–130
- Alam MR, Reaz MBI, Ali MAM (2012) A review of smart homes—past, present, and future. *IEEE Trans Syst Man Cybern* 42:1190–1203
- Andrist S, Bohus D, Kamar E, Horvitz E (2017) What went wrong and why? Diagnosing situated interaction failures in the wild. In: *International conference on social robotics*. Springer, pp 293–303
- Aneja D, McDuff D, Czerwinski M (2020) Conversational error analysis in human–agent interaction. In: *Proceedings of the 20th ACM international conference on intelligent virtual agents*, pp 1–8
- Austin JL (1975) *How to do things with words*, vol 88. Oxford University Press, Oxford
- Bainbridge WA, Hart J, Kim ES, Scassellati B (2008) The effect of presence on human–robot interaction. In: *RO-MAN 2008—the 17th IEEE international symposium on robot and human interactive communication*. IEEE, pp 701–706
- Bainbridge WA, Hart JW, Kim ES, Scassellati B (2011) The benefits of interactions with physically present robots over video-displayed agents. *Int J Soc Robot* 3(1):41–52
- Bohus D, Rudnicky A (2005) Sorry and i didn't catch that! An investigation of non-understanding errors and recovery strategies. In: *Proceedings of the 6th SIGdial workshop on discourse and dialogue*, pp 128–143
- Breazeal C, Dautenhahn K, Kanda T (2016) *Social robotics*. Springer handbook of robotics. Springer, Berlin, pp 1935–1972
- Breazeal C, Fitzpatrick P (2000) That certain look: social amplification of animate vision. In *AAAI*
- Cahn JE, Brennan SE (1999) A psychological model of grounding and repair in dialog. In: *Proceedings of fall 1999 AAAI symposium on psychological models of communication in collaborative systems*
- Cassell J, Bickmore T, Billingham M, Campbell L, Chang K, Vilhjálmsdóttir H, Yan H (1999) Embodiment in conversational interfaces: Rea. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp 520–527
- Cassell J, Sullivan J, Churchill E, Prevost S (2000) *Embodied conversational agents*. MIT Press, Cambridge
- Clark HH (2005) Coordinating with each other in a material world. *Discourse Stud* 7(4–5):507–525
- Clark HH, Brennan SE et al (1991) Grounding in communication. *Perspect Soc Shar Cognit* 13:127–149
- Clark HH, Krych MA (2004) Speaking while monitoring addressees for understanding. *J Mem Lang* 50(1):62–81
- Clark HH, Schaefer EF (1989) Contributing to discourse. *Cognit Sci* 13(2):259–294
- Clark HH, Wilkes-Gibbs D (1986) Referring as a collaborative process. *Cognition* 22(1):1–39
- Cohen PR (1984) The pragmatics of referring and the modality of communication. *Comput Linguist* 10(2):97–146
- Correia F, Guerra C, Mascarenhas S, Melo FS, Paiva A (2018) Exploring the impact of fault justification in human–robot trust. In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. International foundation for autonomous agents and multiagent systems, pp 507–513
- Dourish P (2004) *Where the action is: the foundations of embodied interaction*. MIT Press, Cambridge
- Druga S, Williams R, Breazeal C, Resnick M (2017) Hey Google is it OK if I eat you? Initial explorations in child–agent interaction. In: *Conference on interaction design and children*
- Eberhard KM, Spivey-Knowlton MJ, Sedivy JC, Tanenhaus MK (1995) Eye movements as a window into real-time spoken language comprehension in natural contexts. *J Psycholinguist Res* 24(6):409–436
- Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robot Auton Syst* 42(3–4):143–166
- Foster ME (2019) Face-to-face conversation: why embodiment matters for conversational user interfaces. In: *Proceedings of the 1st international conference on conversational user interfaces*. ACM, p 13
- Galati A (2011) Assessing common ground in conversation: the effect of linguistic and physical co-presence on early planning. Ph.D. Dissertation. The Graduate School, Stony Brook University, Stony Brook, NY
- Garoufi K (2013) Interactive generation of effective discourse in situated context: a planning-based approach. Ph.D. Dissertation. Universität Potsdam
- Gehle, R., Pitsch, K., Dankert, T., & Wrede, S. (2015). Effects of a robot's unexpected reactions in robot-to-group interactions. Presented at the IEMCA 2015, Kolding, Denmark
- Giuliani M, Mirnig N, Stollnberger G, Stadler S, Buchner R, Tschelligi M (2015) Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Front Psychol* 6(2015):931
- Goble H, Edwards C (2018) A robot that communicates with vocal fillers has... Uhhh... greater social presence. *Commun Res Rep* 35:256–260
- Hanna JE, Brennan SE (2007) Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *J Mem Lang* 57(4):596–615
- Harnad S (1990) The symbol grounding problem. *Physica D* 42(1–3):335–346
- Hayes CJ, Moosaei M, Riek LD (2016) Exploring implicit human responses to robot mistakes in a learning from demonstration task. In: *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, pp 246–252
- Hildreth PM, Kimble C, Wright P (1998) Computer mediated communications and communities of practice. In: *Proceedings of Ethicomp* (Vol. 98, pp. 275–286)
- Honig S, Oron-Gilad T (2018) Understanding and resolving failures in human–robot interaction: literature review and model development. *Front Psychol* 9(2018):861
- Jung Y, Lee KM (2004) Effects of physical embodiment on social presence of social robots. In: *Proceedings of PRESENCE*
- Kalegina A, Schroeder G, Allchin A, Berlin K, Cakmak M (2018). Characterizing the design space of rendered robot faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 96–104)
- Kendon A (1967) Some functions of gaze-direction in social interaction. *Acta Psychol* 26(1967):22–63
- Kennedy J, Baxter P, Belpaeme T (2015) Comparing robot embodiments in a guided discovery learning interaction with children. *Int J Soc Robot* 7:293–308
- Kennedy J, Baxter P, Belpaeme T (2015) The robot who tried too hard: social behaviour of a robot tutor can negatively affect child

- learning. In: 2015 10th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp 67–74
43. Kidd CD, Breazeal C (2004) Effect of a robot on user perceptions. In: IROS
 44. Kidd CD, Breazeal C (2008) Robots at home: understanding long-term human-robot interaction. In: IROS
 45. Kiesler S (2005) Fostering common ground in human-robot interaction. In: ROMAN 2005. IEEE international workshop on robot and human interactive communication, 2005. IEEE, pp 729–734
 46. Kontogiorgos D, Avramova V, Alexandersson S, Jonell P, Oertel C, Beskow J, Skantze G, Gustafsson J (2018) A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In: LREC
 47. Kontogiorgos D, Pereira A, Andersson O, Koivisto M, Gonzalez RE, Vartiainen V, Gustafson J (2019) The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In: International conference on intelligent virtual agents. ACM
 48. Kontogiorgos D, Pereira A, Gustafson J (2019) Estimating uncertainty in task oriented dialogue. In: ACM international conference in multimodal interaction
 49. Kontogiorgos D, Pereira A, Sahindal B, van Waveren S, Gustafson J (2020) Behavioural responses to robot conversational failures. In: Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction, pp 53–62
 50. Kontogiorgos D, Skantze G, Abelho PAT, Gustafson J (2019) The effects of embodiment and social eye-gaze in conversational agents. In: 41st annual meeting of the cognitive science (CogSci), Montreal July 24th–Saturday July 27th, 2019
 51. Kontogiorgos D, van Waveren S, Wallberg O, Pereira A, Leite I, Gustafson J (2020) Embodiment effects in interactions with failing robots. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–14
 52. Lee KM, Jung Y, Kim J, Kim SR (2006) Are physically embodied social agents better than disembodied social agents? The effects of physical embodiment, tactile interaction, and people's loneliness in human-robot interaction. *Int J Hum Comput Stud* 64:962–973
 53. Lucas GM, Boberg J, Traum D, Artstein R, Gratch J, Gainer A, Johnson E, Leuski A, Nakano M (2018) Getting to know each other: the role of social dialogue in recovery from errors in social robots. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. ACM, pp 344–351
 54. Luria M, Hoffman G, Zuckerman O (2017) Comparing social robot, screen and voice interfaces for smart-home control. In: Proceedings of the 2017 CHI conference on human factors in computing systems. ACM, pp 580–628
 55. Macdonald RG, Tatler BW (2015) Referent expressions and gaze: reference type influences real-world gaze cue utilization. *J Exp Psychol Hum Percept Perform* 41(2):565
 56. Marakas GM, Johnson RD, Palmer JW (2000) A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *Int J Hum Comput Stud* 5:719–750
 57. Marge M, Rudnicky AI (2019) Miscommunication detection and recovery in situated human-robot dialogue. *ACM Trans Interact Intell Syst* 9(1):1–40
 58. Mirnig N, Stollnberger G, Miksch M, Stadler S, Giuliani M, Tscheligi M (2017) To err is robot: how humans assess and act toward an erroneous social robot. *Front Robot AI* 4(2017):21
 59. Mizoguchi H, Sato T, Takagi K, Nakao M, Hatamura Y (1997) Realization of expressive mobile robot. In: Robotics and automation
 60. Moon Y, Nass C (1996) How “real” are computer personalities? Psychological responses to personality types in human-computer interaction. *Commun Res* 23:651–674
 61. Morales CG, Carter EJ, Tan XZ, Steinfeld A (2019) Interaction needs and opportunities for failing robots. In: Proceedings of the 2019 on designing interactive systems conference. ACM, pp 659–670
 62. Nass C, Steuer J (1993) Voices, boxes, and sources of messages: computers and social actors. *Hum Commun Res* 19(4):504–527
 63. Novick DG, Hansen B, Ward K (1996) Coordinating turn-taking with gaze. In: ICSLP 96
 64. Pereira A, Prada R, Paiva A (2014) Improving social presence in human-agent interaction. In: SIGCHI conference on human factors in computing systems
 65. Powers A, Kiesler S, Fussell S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: International conference on human-robot interaction
 66. Ragni M, Rudenko A, Kuhnert B, Arras KO (2016) Errare humanum est: erroneous robots in human-robot interaction. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 501–506
 67. Richardson DC, Dale R, Kirkham NZ (2007) The art of conversation is coordination. *Psychol Sci* 18:407–413
 68. Robinette P, Li W, Allen R, Howard AM, Wagner AR (2016) Overtrust of robots in emergency evacuation scenarios. In: The eleventh ACM/IEEE international conference on human robot interaction. IEEE Press, pp 101–108
 69. Rossi A, Dautenhahn K, Koay KL, Walters ML (2017) How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In: International conference on social robotics. Springer, pp 42–52
 70. Salem M, Eyssel F, Rohlfing K, Kopp S, Joubin F (2013) To err is human (-like): effects of robot gesture on perceived anthropomorphism and likability. *Int J Soc Robot* 5(3):313–323
 71. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K (2015) Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction. ACM, pp 141–148
 72. Schegloff EA (2007) Sequence organization in interaction: a primer in conversation analysis I, vol 1. Cambridge University Press, Cambridge
 73. Shibata T, Tashima T, Tanie K (1999) Emergence of emotional behavior through physical interaction between human and robot. In: Robotics and automation
 74. Short E, Hart J, Vu M, Scassellati B (2010) No fair!! An interaction with a cheating robot. In: 2010 5th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp 219–226
 75. Short ES, Chang ML, Thomaz A (2018) Detecting contingency for HRI in open-world environments. In: Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction, pp 425–433
 76. Skantze G (2007) Error handling in spoken dialogue systems. Computer Science and Communication Department of Speech, Music and Hearing
 77. Skantze G, Hjalmarsson A, Oertel C (2014) Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Commun* 65:50–66
 78. Straub I (2016) ‘It looks like a human!’ The interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI Soc* 31:553–571
 79. Torta E, Oberzaucher J, Werner F, Cuijpers RH, Juola JF (2013) Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. *J Hum Robot Interact* 1(2):76–99
 80. Trung P, Giuliani M, Miksch M, Stollnberger G, Stadler S, Mirnig N, Tscheligi M (2017) Head and shoulders: automatic error detection in human-robot interaction. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 181–188

81. van Waveren S, Carter EJ, Leite I (2019) Take one for the team: the effects of error severity in collaborative tasks with social robots. In: Proceedings of the 19th ACM international conference on intelligent virtual agents. ACM, pp 151–158
82. Wainer J, Feil-Seifer DJ, Shell DA, Mataric MJ (2006) The role of physical embodiment in human–robot interaction. In: ROMAN 2006
83. Witchel H, Westling C, Tee J, Healy A, Needham R, Chockalingam N (2014) What does not happen: quantifying embodied engagement using NIMI and self-adaptors. *Particip J Audience Recept Stud* 11(1):304–331
84. Wittenburg P, Brugman H, Russel A, Klassmann A, Sloetjes H (2006) ELAN: a professional framework for multimodality research. In: 5th international conference on language resources and evaluation (LREC 2006), pp 1556–1559
85. Yu C, Schermerhorn P, Scheutz M (2012) Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Trans Interact Intell Syst* 1(2):13

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.