The Effects of Anthropomorphism and Non-verbal Social Behaviour in Virtual Assistants

Dimosthenis Kontogiorgos
diko@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Andre Pereira
atap@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Olle Andersson
oll@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Marco Koivisto
marcoko@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Elena Gonzalez Rabal
elgr@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Ville Vartiainen
villevar@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Joakim Gustafson
jkgu@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

The adoption of virtual assistants is growing at a rapid pace. However, these assistants are not optimised to simulate key social aspects of human conversational environments. Humans are intellectually biased toward social activity when facing anthropomorphic agents or when presented with subtle social cues. In this paper, we test whether humans respond the same way to assistants in guided tasks, when in different forms of embodiment and social behaviour. In a within-subject study (N=30), we asked subjects to engage in dialogue with a smart speaker and a social robot. We observed shifting of interactive behaviour, as shown in behavioural and subjective measures. Our findings indicate that it is not always favourable for agents to be anthropomorphised or to communicate with non-verbal cues. We found a trade-off between task performance and perceived sociability when controlling for anthropomorphism and social behaviour.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**; *Empirical studies in HCI*; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

human-computer interaction, conversational artificial intelligence, empirical studies, smart speakers, social robots

Figure 1: Situated interaction with a human and a human-like social robot engaging in task-oriented dialogue.

ACM Reference Format:
Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The Effects of Anthropomorphism and Non-verbal Social Behaviour in Virtual Assistants. In *ACM Int'l Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308532.3329466

1 INTRODUCTION

With virtual assistants and domestic technology on the rise, several questions remain open on how humans engage with interactive agents when in different forms of anthropomorphism and social behaviour. A wide range of interaction modalities has been designed and researched for virtual assistants, that come in various forms such as smart speakers [2] and social robots [7]. However, by design, social robots provide additional modes of pragmatic communication. Social robots can express their internal state using not only speech but also non-verbal behaviour. By generating multimodal communicative behaviours such as gaze cues, facial expressions

IVA '19, July 2–5, 2019, Paris, France

ACM ISBN 978-1-4503-6672-4/19/07…$15.00
https://doi.org/10.1145/3308532.3329466

and communicative gestures [8, 38], social robots enable different manifestations of interaction similar to how humans interact with each other [47].

Ever since the adoption of virtual assistant devices at home has increased, several reports have discussed the imperious style of language used when addressing these assistants with speech [36, 46]. This effect may be due to the lack of affordance in situational awareness of these devices or the unimodal design of using only voice. A question remains open on whether an anthropomorphic face that displays human inspired non-verbal social behaviours can positively impact the conversational cues towards virtual assistants by conveying higher levels of sociability that facilitate the interaction. The current command-based nature of the interactions may be efficient for factual question answering, but is still far from socially contingent interactions, and constrained to a number of implicit and imperious conversational cues [6, 54].

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 [35, 39]. In addition, patterned social behaviours may facilitate social interaction with users, however, generating and interpreting these cues can induce higher levels of cognitive load [53].

Social robots do employ such behaviours, and provide the possibility of generating non-verbal social behaviours in their interactions with humans [13]. Many of these behavioural elements are subtle social cues (e.g. gaze shifts), that are highly important for situated human conversational environments. One reason why face-to-face interaction is preferred is that a lot of familiar information is encoded in the non-verbal cues that are being exchanged.

In this paper we contribute to this emerging field with an empirical evaluation of the elements of: i) *anthropomorphic design* and ii) *non-verbal social behaviour* in guided tasks with virtual assistants. We study 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 anthropomorphic face or the non-verbal features that contribute to the perceived differences and remove the non-verbal behaviour of the social robot in a third condition. The aim of the study is therefore to investigate the following question:

- What are the effects in human behaviour when simulating anthropomorphism and social behaviour in virtual assistants during task-oriented dialogue?

2 RELATED WORK

2.1 Virtual assistants

Virtual assistants are becoming ubiquitous in society, and they are embedded in various forms and embodiments, from smart watches and phones to voice-based smart speakers such as Amazon Echo and Google Home. There seems to be an interest in literature on how different representations of physical embodiment and anthropomorphic features affect performance and the perception of presence in virtual assistants. Several studies have compared agents in digital screens to social robots [12, 26, 53] and they have shown that anthropomorphic agents that are physically co-located are generally preferred and are perceived to be more socially present than their virtually embodied versions [5, 21, 24, 25, 28, 32, 52], or remote video representations of the same agents [44, 55]. Other studies have shown that social robots' perceived situation awareness is higher [34] and by adding non-verbal cues, the same agent is perceived more socially present [15, 42].

2.2 Anthropomorphism

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 [35]. 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 [13]. 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 [22]. The features that drive human-likeness have been thoroughly studied, and some prominent features are facial surface looks, body manipulators, facial features and locomotion [43].

In embodied interactions, the social and physical are intertwined by exploiting the sense of familiarity with the everyday world. By emphasising physical presence, embodiment is defined through a physical manifestation in the world and embodied phenomena occur in real time and space [11]. Previous work in interaction design has shown that physical and embodied interaction is vital to consider when building interactive agents that integrate the physical and computational world, as such embodied interaction encourages "thinking through doing" [27].

Anthropomorphic virtual assistants take advantage of design elements afforded in their shape and movements [16]. A very human-like agent will make humans expect a higher degree of *social interaction*, which is essential when designing the physical appearance of an agent. However, it is not just the physical embodiment of the robot that has implications on its perceived *social presence*, but the behaviour and actions of the robot as well [50].

2.3 Non-verbal social behaviour

Socially interactive agents that make use of social behaviour features promise an opportunity to bring social values into computing and help coordination between humans and machines by taking advantage of their social cues and intentions [11]. While conversational interfaces manifest intent using language, social robots communicate intentions with the use of multimodal cues, and additionally encourage users to anticipate joint actions and shared intent in the same physical space [11]. Sharing the same physical space involves situatedness, grounded between social actions and settings in the physical world [51].

Voice as a social cue can create the impression that a machine has multiple distinct entities [40]. Non-verbal behaviour however, is used for communication, signalling and social coordination. The more human-like the agents' responses, the more they are attributed

(a) Smart Speaker (b) Social Robot

Figure 2: The agents used in the study. *Smart Speaker* morphology: Cylinder speaker with led. Output modality: Voice. *Social Robot* morphology: Human-like back-projected face. Output modalities: Voice and head movement.

as social actors [35, 40] and agents that do not use the rich set of social behaviours evoke lesser feelings of social presence.

Social presence has been defined as the degree of salience of interlocutors in the interaction and interpersonal relationships [48]. It is measured in the degree of initial awareness, attentional allocation and affective comprehension [19]. 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 [9, 45].

Our work differs and in part extends the discussed studies as follows. First, differently from [5, 42, 44, 55], we simulate both anthropomorphism and non-verbal behaviour concurrently, and second we apply the comparison in physically present voice-based agents, where we discuss the implications of improved robot's sociability against task performance, differently than [26, 34, 53] which involve smart-home, mobile devices and their perceived usability measures.

3 METHOD

In order to investigate the impact of anthropomorphism and social non-verbal behaviour, we defined three embodied virtual assistants using two embodied conversational agents.

3.1 Experimental conditions

1. The *Smart Speaker (SS)* is an embodied conversational agent (Figure 2a) that interacts with speech. We used a first generation Amazon Echo smart speaker, which was connected via Bluetooth and a TTS service similar to the default Echo TTS was used to send pre-scripted voice commands.

2. The *AnthropoMorphic Robot (AMR)* is an embodied assistant (Figure 2b) in the form of a robotic head with a human-like face, that as the SS uses speech to interact and no other modalities. We used a back-projected robotic head called Furhat [1]. The robot was stationary and did not use any head or eye movement, but statically looked at the user. The robot had a TTS of equivalent quality to Echo, speaking the same pre-scripted utterances. Choosing a robotic head instead of a full-body embodied robot limits the modalities of communication and makes it easier to control for comparison to a smart speaker.

3. The *AnthropoMorphic Social Robot (AMSR)* is the same robotic head as AMR, that also uses voice for interaction and additionally generates a set of social eye-gaze behaviours using head movement. These included task-based functional behaviours such as gazing to objects during a referring expression and a turn-taking gaze mechanism.

3.2 Hypotheses

Towards answering the research question defined above, we posed the following hypotheses:

- ☐ **H1.** As a social robot with a human like design but also non-verbal social behaviour, *the AMSR will be perceived by human users as having higher levels of sociability when compared to the SS and the AMR.*
- ☐ **H2.** While non-vebal behaviour should increase levels of sociability, a human-like design without non-verbal cues should not induce the same differences. *Differences in levels of sociability will not apply between the SS and AMR.*
- ☐ **H3.** Task time should be irrelevant of non-verbal cues or human-like design. *There will not be any difference in task completion time across the AMSR and SS.*
- ☐ **H4.** Due to its social cues, *the AMSR will generally be preferred for the task.*
- ☐ **H5.** Due to human-like coordination of non-verbal behaviour, *the AMSR will be perceived to be more intelligent.*

3.3 Experimental design

Using the three fore-mentioned agents, an exploratory within-subject user study was conducted to analyse the impact of the anthropomorphic and non-verbal behaviour features. To test our hypotheses, we manipulated two independent variables [*embodiment* and *social eye-gaze*], in three conditions [*SS*, *AMR*, *AMSR*], presented in different order to participants using a balanced Latin Square. 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.

3.4 Task

We asked subjects to cook 3 variations of fresh spring rolls without providing the recipes; they had to get the recipes by interacting with the agent, a common scenario in usage of commercially available virtual assistants. Different varieties of ingredients and amounts were used. The experiment setup also included ingredients not used in any of the recipes, encouraging participants to interact with the virtual assistants to find out the correct ingredients for each recipe. The task was the same in each condition, but different recipes were used.

To ensure participants would engage with the agents, they were told that if they followed the recipe with the correct ingredients and

amounts, they would take the food with them after the experiment. 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*". AMSR however, also gazed at the referent ingredients (0.5s 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 (Figure 3).

Figure 3: Subjects were given a variety of ingredients to cook fresh spring rolls recipes with the help of an agent.

3.5 Dialogue policy

All agents followed the same dialogue policy and interaction protocol, which was defined upon a set of *dialogue acts* within the action space of the interaction [30]. Given a human action or utterance, an appropriate response was selected from the dialogue policy. A sample dialogue:

USER: [FINISHED ACTION] So, what's next?

AGENT: [INSTRUCTION] Next, take three pieces of lettuce and put it in the spring roll.

USER: [CLARIFICATION-Q] Oh, where is the lettuce?

AGENT: [CLARIFICATION-A] The lettuce is the green thing in the middle.

USER: [STARTED ACTION] Uh, yes! Okay!

To dismiss potential problems in speech recognition and language understanding, we used a human wizard (WoZ) to control the behaviours of the agents in timings and dialogue act selection. The human wizard selected the appropriate agent response, as triggered by user speech or actions. The WoZ application and dialogue acts were the same across all conditions. For every dialogue act, a set of predefined utterances was available, that the system would choose at random to speak, given the current dialogue act in the task. The WoZ therefore indicated only the current dialogue act in conversation, and not what to say.

3.6 Gaze for facilitating turn-taking

Gaze has been shown to be important for regulating conversational turn-taking, as people look towards the listener at the end of their utterances to indicate they have finished their turn [23, 41]. Employing such a behaviour in agents, leads to human-like conversational turn-taking where each listener waits for the speaker's end of utterance before taking a turn [3, 49]. In order to facilitate natural turn-taking mechanisms from the agent, we defined a heuristic gaze model on timings for *turn-taking gaze* and *referential gaze* to objects.

The AMSR agent engaged in *mutual gaze* and *joint attention* with the subjects during the interactions. Before an utterance, the agent 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. The agent always gazed at referent objects right before they were mentioned (defined as 0.5s before) [17, 37].

3.7 Experimental procedure

Participation in the study was individual and the experiment was divided in 3 phases. 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 the agent, and after that filled a questionnaire on the perceived social presence and intelligence of the agent. They repeated that phase 3 times with a new agent every time (counter-balanced). Last, participants filled an exit questionnaire with their preference of the agent, and their opinions on the agents in open-ended questions. During the agent trials, participants were alone in the room, and the WoZ was monitoring their actions using a ceiling camera with a live feed of the room (Figure 1). Participants were not told that the agents were controlled by a human wizard. On average, each cooking trial was 3.7 minutes long.

Participants were not motivated to finish the task as quickly as possible, 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 and the agent were situated on the side of the table. The ingredients were fixed in place and the order of them remained consistent throughout the full experiment. The agents were situated on the sides of the table, with only the agent relevant to the task visible. The distance between the agent and the subject was set to about 1.6m, defined as an appropriate social distance by Hall [18].

3.8 Participants

Participants were instructed that they are able to stop the experiment at any time and were compensated with a cinema ticket and the food they cooked during the study. 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 smart speaker and a robot before, while 6 with none. Overall, their experience with digital technology was 4.8 from 1-7 (stdev=1.6).

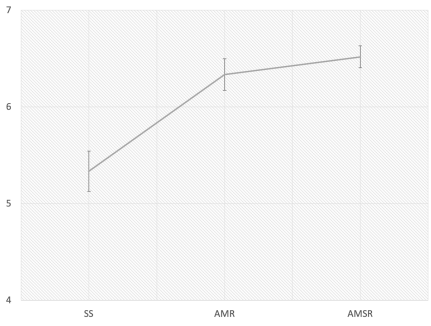

Figure 4: Co-presence.

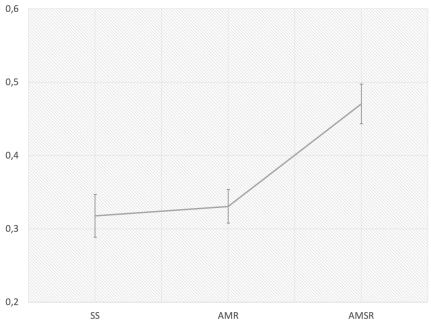

Figure 5: Gaze to agent.

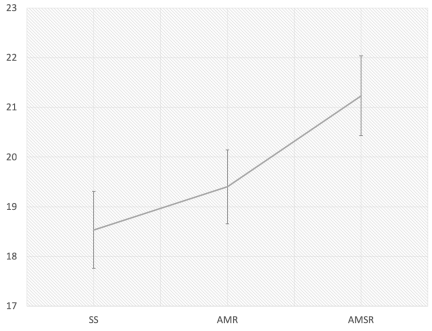

Figure 6: Conversational turns.

Sociability (fig 4-6): Self-reported *co-presence* (left), Proportional *gaze to agent* during agent instructions (middle), and Number of *conversational turns* (right) across agents. Error bars indicate standard error of the mean (n=30).

3.9 Quantitative measures

In order to evaluate the spoken dialogue agents, we used objective and subjective measures along two main themes: *sociability* and *interaction time*. We used task-based behavioural measures such as task time, as well as conversational features. As subjective measures, we used a validated questionnaire (in Likert 1-7) by Harms and Biocca [19], that divides social presence in different dimensions including co-presence, attentional allocation, message understanding and behavioural interdependence. As we manipulated the agents' embodiment and attentional capabilities, we expected to find differences in conditions on attentional and conversational cues. Last, a perceived intelligence question was included.

Sociability. We used the following behavioural measures: *Proportional gaze to the agent:* We measured subjects' gaze using their head pose direction and shifts (automatically annotated from a motion capture system [29, 31]). *Number of conversational turns:* The number of turns the agent responded to human turns/actions (extracted from agent logs). *Clarification questions:* We use the number of times the agent answers clarification questions ["lettuce, right?"] (extracted from agent logs). *Co-presence and attentional allocation:* Subjective measures used to indicate how subjects perceive the presence of the agent (self-reported).

Interaction time. We measured the *task time* from the first agent action to the last, when greeting the subject (extracted from agent logs) to count the amount of time subjects engaged with the agents.

3.10 Qualitative measures

The post-experimental questionnaire included questions asking participants to choose their preferred agent for the task and questions to elaborate on the preference. Participants were also asked to point the differences of the three agents to understand if they are aware of what is tested in the study (therefore less sensitive to our manipulation). Finally, two open-ended questions were included about their experience from the tasks with the agents and their thoughts on what is the intention of the study.

4 RESULTS

We present findings along two themes: *sociability* and *interaction time* using behavioural and subjective measures. We first report the results from quantitative measures, and finally, notable insights from the qualitative data.

4.1 Sociability

A repeated measures ANOVA showed that Co-Presence, Wilks' Lambda = .328, $F(2,28) = 28.63$, $p < .001$) and Attentional Allocation, Wilks' Lambda = .506, $F(2,28) = 13.62$, $p < .001$), were significantly different between the conditions. Pairwise comparisons with Bonferroni corrections, and p value adjusted for multiple comparison, showed that SS (5.3) was perceived as less co-present (Figure 4) than both AMSR (6.5, $p<.001$) and AMR (6.3, $p<.001$). Attentional allocation was also significantly lower in SS (5.4) when compared to AMSR (6.3, $p<.001$) and to AMR (6.0, $p<.001$). No other statistical differences were found.

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. A repeated measures ANOVA to test the effect of gaze showed a significant main effect, Wilks' Lambda = .436, $F(2,28) = 18.07$, $p < .001$). Post-hoc tests with a Bonferroni correction, and p value adjusted for multiple comparisons, revealed that gaze towards AMSR (.47) is statistically greater than gaze to SS (.31, $p<.001$) and AMR (0.33, $p<.001$). No other statistical differences were found in pairwise comparisons (Figure 5).

A repeated measures ANOVA on the number of conversational turns showed a significant main effect Wilks' Lambda = .728, $F(2,28) = 5.23$, $p = .012$). Post-hoc tests with a Bonferroni correction, and p value adjusted for multiple comparisons, revealed that conversational turns with AMSR (21.23) are statistically greater to AMR (19.40, $p=.036$) and to SS (18.53, $p=.033$). No statistical differences were found between the other two conditions (Figure 6).

When compared across conditions, repeated measures ANOVA tests revealed significant differences among the three conditions on the number of clarification questions Wilks' Lambda = .743, $F(2,28)$

= 4.83, p = .016). Post-hoc pairwise tests with Bonferroni correction, and p value adjusted for multiple comparisons, were carried out for the three pairs of groups. The results indicated a significant difference between SS (2.5) and AMSR (4.0, p=.019). There were no other statistical differences.

No differences were found in perceived intelligence. Also, behavioural interdependence was not dependent on the conditions but related to having interacted with robots before (r=-.273, p=.009). Experience with technology (r=-.228, p=.031) and previous exposure to robots (r=-.506, p=.000) seem to affect perceived intelligence. Previous exposure to smart speakers did not appear to be correlated to perceived intelligence or any social presence dimensions.

4.2 Interaction time

We trivially found that *task time* is correlated with the number of *conversational turns* (r=.654, p<.001), and the number of *clarifying questions* (r=.566, p<.001). 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 Wilks' Lambda = .739, F(2,28) = 4.94, p = .014. Post-hoc tests with a Bonferroni correction, and p value adjusted for multiple comparisons, revealed that interaction time with AMSR (232.93) is statistically greater to the AMR condition (217.26, p=.023) and to SS (212.66, p=.041). No other statistical differences were found (figure 8).

Finally, message understanding was not significantly different between conditions (F=.29, p=.737) but it was negatively correlated with task time (r=-.304, p=.004), the number of conversational turns (r=-.347, p=.001), and the number of clarifying questions (r=-.353, p=.001).

4.3 Qualitative evaluation

Participants mentioned that they connected more with AMSR, however, they also described that lack of social cues may be beneficial, depending on the task:

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

> *"[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"*

Perceived differences between the agents. Out of 30 subjects, 18 replied this question. While differences in the embodiment were very obvious between [SS] and [AMR/AMSR], 66% of the participants did not notice a difference between [AMR] and [AMSR]. We found they identified that there was head movement from the social robots, but were not aware that only one of them [AMSR] employed that behaviour.

Preferred agent for the task. 69% of the participants preferred AMSR, while 24% AMR and 7% SS (χ^2 = 17.862, p < .001). Looking at the subjects who did not notice a difference between AMR and AMSR, 2/3 chose AMSR as the preferred robot for the task. However, from 1/3 of the participants who identified the difference in gaze, therefore less sensitive to our manipulation, all preferred AMSR.

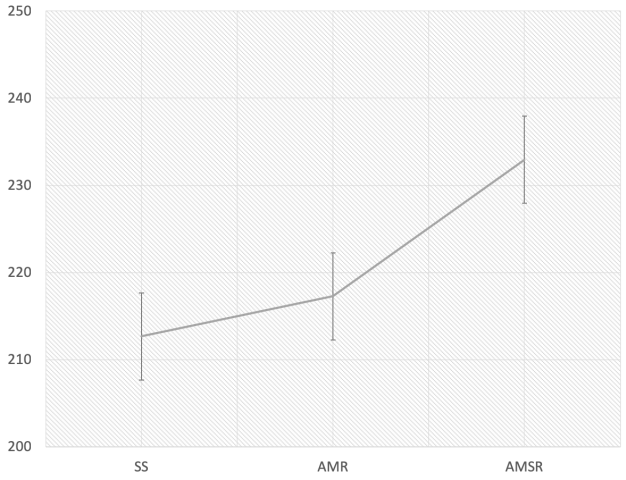

Figure 8: *Interaction time* (in seconds) per condition. Error bars indicate standard error of the mean (n=30).

5 DISCUSSION

In an experiment with human subjects, we found a lack of positive effects in task performance on interactions with an anthropomorphic social agent. Nonetheless, our findings show higher rates of sociability and engagement in conversation with the social robot.

The agents we compared, represent different levels of embodiment in virtual assistants. The most preferred agent for the task had an anthropomorphic embodiment and a set of scripted non-verbal behaviours. While AMSR was preferred, dialogue was longer in conversational turns, which explains why task time was increased by 10% on average in comparison to the less anthropomorphic in embodiment SS. Participants looked at AMSR longer during instructions and started following up on the agent's instruction close to the end of its turn. Intuitively, a turn-taking gaze mechanism invokes subjects a greater feeling of sociability, assuming they attribute that agent the role of a more socially present partner in conversation, which was also confirmed on the self-reported social presence dimensions.

5.1 Sociability and task performance

Subjects spent less time in the task with SS and AMR than AMSR. However, when comparing message understanding measures within conditions, no statistical difference was found. Intuitively, all agents were understood the same, regardless of differences in turns and clarification questions.

Voice-only controlled interfaces are ubiquitous and intuitive, and provide a modality ideal for accessing devices while being engaged in another task [10]. However, the sense of control is lower when using voice controlled devices, which may be explained by their limited situation awareness [34].

Task performance is nevertheless dependent on the nature of the task; in more task-oriented domains, such as emergency management, interactions may be efficiency-prone. A user, may want to get the task done as quickly as possible, and get frustrated when having to speak longer than necessary. In these cases, following established social norms, and providing robots with improved mechanisms to do so, may make users distracted from the task at hand and more focused on the social interaction [20, 24]. However, other tasks such as in the home-care domain are very dependent on social cues and interaction value. In these environments, face-to-face collaboration might be favourable due to the usage of more natural and familiar channels of communication.

We were, able to verify hypothesis **[H1]** that AMSR displays higher levels of sociability, as shown in the gaze, co-presence and attentional allocation dimensions, however with the cost of task performance, if one defines such as completing the task in the shortest amount of time possible. In this study, increased time in dialogue length had a positive effect to the robot's perceived sociability, in contrast to [33]. We therefore reject hypothesis **[H3]**, that there will not be differences in task performance. While very few errors in the task were done with the agents, subjects spent more time and were slower to act with AMSR.

5.2 Anthropomorphism

The results support **[H4]** reflecting that the anthropomorphic social robot (AMSR) would be preferred for the task. We saw a wide difference between AMSR and SS, however AMR also rated higher than SS, which aligns with the fact that there is a relation to anthropomorphic agents with familiarity, in terms of natural means of communication.

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 in the agent preference. Social robots are at time of writing still emerging platforms and not as common and commercially available, as smart speakers are. However, 2/3 of the participants were not able to identify the difference between the two anthropomorphic agents, while they still preferred AMSR for the task. This indicates that the non-verbal cues were subtle and asserted familiarity with the device, but did not happened with AMR which was the same device.

5.3 Non-verbal social behaviour

We reject hypothesis **[H2]** reflecting that no differences in sociability would be found between SS and AMR. Our assumption was that anthropomorphic facial 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 social interaction with users. While this is partially supported, we saw statistical differences in social presence dimensions between SS and AMR, indicating that an anthropomorphic agent is perceived to be more co-present in the room than a non-anthropomorphic one, even without a set of non-verbal behaviours.

One could expect that AMR would display higher levels of sociability than SS, as it raises expectations with its anthropomorphic design, however we did not observe any statistical differences in eye gaze, conversational turns and interaction time from SS. A human-like design is not enough to establish rapport with humans; human-like behaviour may be expected as well, when anthropomorphic designs are manifested. Intuitively, a turn-taking gaze mechanism invokes subjects a greater feeling of social presence, assuming they attribute that agent the role of a socially intelligent partner in conversation.

Our assumption is that AMSR has *joint attention* afforded as an embodied phenomenon in its actions, giving the impression it is more aware on the situatedness of the task. 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.

The results also suggest that smart speakers, while embodied, do not facilitate the same turn-taking mechanisms as social robots with non-verbal behaviour, likely due to the lack of eye-gaze and other non-verbal behaviours. Finally, while literature has shown that people are perceived as being more intelligent and more friendly, when they engage in direct eye contact [4, 14], we did not find any statistical differences in the agents' perceived intelligence to support **[H5]** in these findings.

6 CONCLUSION

In this paper, we discussed the trade-off between sociability and task performance when empirically controlling for anthropomorphism and social behaviour parameters on guided tasks with virtual assistants. 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, may need more or less sociability versus the value of task performance. Not every agent needs to be anthropomorphised or to communicate with nonverbal behaviour; teasing out these variables and how they affect performance and sociability is the focus of this study.

Tying these findings to the agents' differences in embodiment is of course one possible interpretation. To understand which of the independent variables contributed to the general preference of the robot, we concluded that while an anthropomorphic physical embodiment increases the perceived social presence, a set of non-verbal behaviours also increase the conversational turns and attentional allocation to the agent. We also saw that the users' 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.

Further research should be conducted in different HRI scenarios, to investigate variability in the nature of the task and its relation to social engagement between humans and agents. In sum, despite potential task performance drawbacks, situation aware social robots hold a good interaction paradigm for enabling improved social interactions with users.

ACKNOWLEDGMENTS

We would like to acknowledge the support from the Swedish Foundation for Strategic Research project FACT (GMT14-0082). We would also like to thank the anonymous reviewers for their valuable comments in earlier versions of this paper.

REFERENCES

[1] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*.

[2] Muhammad Raisul Alam, Mamun Bin Ibne Reaz, and Mohd Alauddin Mohd Ali. 2012. A review of smart homes - Past, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics* (2012).

[3] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *ACM/IEEE international conference on Human-robot interaction*.

[4] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[5] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN*.

[6] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If You Ask Nicely: A Digital Assistant Rebuking Impolite Voice Commands. In *International Conference on Multimodal Interaction*.

[7] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. In *Springer handbook of robotics*.

[8] Cynthia Breazeal and Paul Fitzpatrick. 2000. That certain look: Social amplification of animate vision. In *AAAI*.

[9] Herbert H Clark. 2005. Coordinating with each other in a material world. *Discourse studies* (2005).

[10] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design*.

[11] Paul Dourish. 2004. *Where the action is: the foundations of embodied interaction*.

[12] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. Hey Google is it OK if I eat you?: Initial Explorations in Child-Agent Interaction. In *Conference on Interaction Design and Children*.

[13] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* (2003).

[14] Chris Fullwood and Gwyneth Doherty-Sneddon. 2006. Effect of gazing at the camera during a video link on recall. *Applied Ergonomics* (2006).

[15] Henry Goble and Chad Edwards. 2018. A Robot That Communicates With Vocal Fillers Has... Uhhh... Greater Social Presence. *Communication Research Reports* (2018).

[16] Randy Gomez, Deborah Szapiro, Kerl Galindo, and Keisuke Nakamura. 2018. Haru: Hardware Design of an Experimental Tabletop Robot Assistant. In *International Conference on Human-Robot Interaction*.

[17] Zenzi M Griffin. 2001. Gaze durations during speech reflect word selection and phonological encoding. *Cognition* (2001).

[18] Edward Twitchell Hall. 1966. *The hidden dimension*. Vol. 609.

[19] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

[20] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Computer supported cooperative work*.

[21] Younbo Jung and Kwan Min Lee. 2004. Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE* (2004).

[22] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the Design Space of Rendered Robot Faces. In *International Conference on Human-Robot Interaction*.

[23] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* (1967).

[24] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics* (2015).

[25] Cory D Kidd and Cynthia Breazeal. 2004. Effect of a robot on user perceptions. In *IROS*.

[26] Cory D Kidd and Cynthia Breazeal. 2008. Robots at home: Understanding long-term human-robot interaction. In *IROS*.

[27] Scott R Klemmer, Björn Hartmann, and Leila Takayama. 2006. How bodies matter: five themes for interaction design. In *Designing Interactive systems*.

[28] Tomoko Koda and Takuto Ishioh. 2018. Analysis of the Effect of Agent's Embodiment and Gaze Amount on Personality Perception. In *4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*.

[29] Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexandersson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze, and Joakim Gustafsson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *LREC*.

[30] Dimosthenis Kontogiorgos, Andre Pereira, and Joakim Gustafson. 2019. The trade-off between interaction time and social facilitation with collaborative social robots. In *The Challenges of Working on Social Robots that Collaborate with People, CHI 2019*.

[31] Dimosthenis Kontogiorgos, Elena Sibirtseva, Andre Pereira, Gabriel Skantze, and Joakim Gustafson. 2018. Multimodal Reference Resolution In Collaborative Assembly Tasks. In *International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*.

[32] Kwan Min Lee, Younbo Jung, Jaywoo Kim, and Sang Ryong Kim. 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. *International Journal of Human-Computer Studies* (2006).

[33] Diane J Litman and Shimei Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction* (2002).

[34] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *CHI Conference on Human Factors in Computing Systems*.

[35] George M Marakas, Richard D Johnson, and Jonathan W Palmer. 2000. A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies* (2000).

[36] Rachel Metz. 2017. Growing Up with Alexa. https://www.technologyreview.com/s/608430/growing-up-with-alexa/

[37] Antje S Meyer, Astrid M Sleiderink, and Willem JM Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition* (1998).

[38] Hiroshi Mizoguchi, Tomomasa Sato, Katsuyuki Takagi, Masayuki Nakao, and Yotaro Hatamura. 1997. Realization of expressive mobile robot. In *Robotics and Automation*.

[39] Youngme Moon and Clifford Nass. 1996. How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication research* (1996).

[40] Clifford Nass and Jonathan Steuer. 1993. Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research* (1993).

[41] David G Novick, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *ICSLP 96*.

[42] Andre Pereira, Rui Prada, and Ana Paiva. 2014. Improving social presence in human-agent interaction. In *SIGCHI Conference on Human Factors in Computing Systems*.

[43] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. 2018. What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *International Conference on Human-Robot Interaction*.

[44] Aaron Powers, Sara Kiesler, Susan Fussell, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *International conference on Human-robot interaction*.

[45] Daniel C Richardson, Rick Dale, and Natasha Z Kirkham. 2007. The art of conversation is coordination. *Psychological science* (2007).

[46] Michael S. Rosenwald. 2017. How millions of kids are being shaped by know-it-all voice assistants. https://www.washingtonpost.com/local/how-millions-of-kids-are-being-shaped-by-know-it-all-voice-assistants/

[47] Takanori Shibata, Toshihiro Tashima, and Kazuo Tanie. 1999. Emergence of emotional behavior through physical interaction between human and robot. In *Robotics and Automation*.

[48] John Short, Ederyn Williams, and Bruce Christie. 1976. The social psychology of telecommunications. (1976).

[49] Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication* (2014).

[50] Ilona Straub. 2016. 'It looks like a human!' The interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & society* (2016).

[51] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*.

[52] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. virtual agent embodiment and effects on social interaction. In *International Conference on Intelligent Virtual Agents*.

[53] Elena Torta, Johannes Oberzaucher, Franz Werner, Raymond H Cuijpers, and James F Juola. 2013. Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. *Journal of Human-Robot Interaction* (2013).

[54] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *CHI Conference on Human Factors in Computing Systems*.

[55] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006*.