The Effects of Embodiment and Social Eye-Gaze in Conversational Agents

Dimosthenis Kontogiorgos, Gabriel Skantze, Andre Pereira, and Joakim Gustafson
(diko@kth.se, gabriel@speech.kth.se, atap@kth.se, jocke@speech.kth.se)
KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

The adoption of conversational agents is growing at a rapid pace. Agents however, are not optimised to simulate key social aspects of situated human conversational environments. Humans are intellectually biased towards social activity when facing more anthropomorphic agents or when presented with subtle social cues. In this work, we explore the effects of simulating anthropomorphism and social eye-gaze in three conversational agents. We tested whether subjects' visual attention would be similar to agents in different forms of embodiment and social eye-gaze. In a within-subject situated interaction study (N=30), we asked subjects to engage in task-oriented dialogue with a smart speaker and two variations of a social robot. We observed shifting of interactive behaviour by human users, as shown in differences in behavioural and objective measures. With a trade-off in task performance, social facilitation is higher with more anthropomorphic social agents when performing the same task.

Keywords: human-computer interaction, social agents, conversational artificial intelligence, smart speakers, social robots

Introduction

With conversational AI and domestic technology on the rise, several questions remain open on how humans engage with interactive agents when in different forms of embodiment and social behaviour. A wide range of interaction modalities has been researched for agents, that come in various forms such as smart speakers (Alam, Reaz, & Ali, 2012), and social robots (Breazeal, Dautenhahn, & Kanda, 2016). 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 (Breazeal & Fitzpatrick, 2000; Mizoguchi, Sato, Takagi, Nakao, & Hatamura, 1997), social robots enable different manifestations of interaction similar to how humans interact with each other (Shibata, Tashima, & Tanie, 1999).

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 (Marakas, Johnson, & Palmer, 2000; Moon & Nass, 1996). While interactions between humans include many subtle social cues that we take for granted, 'face-to-face' interactions are still considered to be the gold standard of communication when interacting with either humans or conversational agents (Adalgeirsson & Breazeal, 2010). Therefore, agents need to employ anthropomorphic designs and a rich set of social behaviours to be considered as socially intelligent partners in interactions.

Figure 1: Situated interaction with a human-like social robot.

Many social robots do employ these elements, especially the ones with a human-like design, and provide the possibility of generating non-verbal social behaviours in their interactions with humans (Fong, Nourbakhsh, & Dautenhahn, 2003). Many of these behavioural elements are subtle social cues (e.g. gaze shifts and facial expressions), 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. However, generating and interpreting these cues, induces higher levels of cognitive load (Torta, Oberzaucher, Werner, Cuijpers, & Juola, 2013) and may therefore increase interaction time. This suggests that human-like conversational agents that can express patterned non-verbal behaviours can cause social facilitation in users, but may be less efficient in task performance.

In this paper we contribute to this emerging field with a two-fold empirical evaluation of the elements of: 1) *anthropomorphic design* and 2) *non-verbal social behaviour* in conversational agents. We study whether a human-like face (i.e. a social robot), capable of displaying non-verbal cues, shifts interactive behaviour in comparison to a voice-only conversational agent (i.e. a smart speaker), that does not employ these multimodal features. Our contribution consists of a user study that was conducted with participants interacting with a smart speaker and a social robot collaborating in dialogue. To comprehend the effects of the comparison further, we test whether it is the anthropomorphic face or the social eye-gaze 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 was therefore to investigate the following research question:

- What are the effects in human behaviour when simulating anthropomorphism and social eye-gaze in conversational agents?

Related work

Conversational agents have become ubiquitous, and they are embedded in various forms and embodiments, from smart 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 the perception of social presence and facilitation in agents. Studies have compared agents in digital screens to social robots (Torta et al., 2013; Kidd & Breazeal, 2008) and have shown that anthropomorphic agents that are physically co-located are generally preferred and perceived to be more socially present than their virtually embodied versions (Kennedy, Baxter, & Belpaeme, 2015; Lee, Jung, Kim, & Kim, 2006; Kidd & Breazeal, 2004; Bainbridge, Hart, Kim, & Scassellati, 2008; Koda & Ishioh, 2018; Jung & Lee, 2004; Thellman, Silvervarg, Gulz, & Ziemke, 2016), or remote video representations of the same agents (Powers, Kiesler, Fussell, Fussell, & Torrey, 2007; Wainer, Feil-Seifer, Shell, & Mataric, 2006). Other studies have shown that social robots' perceived situation awareness is higher (Luria, Hoffman, & Zuckerman, 2017), and by adding non-verbal cues, the same agent is perceived more socially present (Pereira, Prada, & Paiva, 2014; Goble & Edwards, 2018).

Anthropomorphic agents take advantage of design elements afforded in their shape and movements (Gomez, Szapiro, Galindo, & Nakamura, 2018). Social robots in particular, raise expectations on how sophisticated they are in their actions and how socially intelligent they are perceived. A very human-like agent will make humans expect a higher degree of interaction and *social facilitation*, which is essential when designing the physical appearance of a social 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 (Straub, 2016).

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 (Dourish, 2004). While conversational interfaces manifest intent recognition using language and dialogue, social robots as embodied interfaces, communicate intentions with the use of multimodal cues, and additionally encourage users to anticipate joint actions and shared intent in the same physical space (Luria et al., 2017).

Non-verbal behaviour is used for communication and social coordination. The more human-like the agents' responses, the more they are attributed as social actors (Nass & Steuer, 1993). Social eye-gaze in particular, refers to the communicative cues of eye contact between humans and is classified to 4 main archetypes (Admoni & Scassellati, 2017): 1) *Mutual gaze* where interlocutors attention is directed at each other, 2) *Joint attention* where interlocutor's focus their attention on the same object, 3) *Referential gaze* which is directed to an object and often comes with referring language, and 4) *Gaze aversions* that typically avert from the main direction of gaze -i.e. the interlocutors face.

(a) Smart Speaker (b) Social Robot

Figure 2: The conversational agents used in the study.

The current work differs and in part extends the discussed studies. First, we simulate both anthropomorphism and non-verbal behaviour in the same study, and second we apply the comparison in only physically present voice-based agents, where we discuss the implications of *social eye-gaze* against *task performance*. Is a human-like face sufficient to cause social facilitation or is non-verbal social behaviour also needed when interacting with conversational agents?

Method

In order to investigate the impact of anthropomorphism and social eye-gaze in this study, we chose three conversational agents and a human trial. All agents engaged in human-agent interaction using the same dialogue policy and simulated situation awareness of human actions (Figure 1).

Experimental conditions

1. The *Human Agent (H)*. 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. That way, subjects got familiar with the task and we were able to reduce the learning curve.

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

3. The *AnthropoMorphic Robot (AMR)* is an embodied conversational agent (Figure 2b) in the form of a robotic head with a human-like face, that as the SS uses only speech to interact and no other modalities. We used a back-projected human-like robotic head with three degrees of freedom called

Furhat. The robot was stationary and did not use any head movements, but statically looked at the subject. The robot had a TTS of equivalent quality to Echo, speaking the same pre-scripted utterances. The reason for choosing a robotic head instead of a full-body embodied robot is that it limits the modalities of communication, making it easier to control for comparison to a smart speaker.

4. 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 the ingredient during a referring expression and a turn-taking gaze mechanism.

Hypotheses

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

- ☐ **H1.** We expected that a robot with non-verbal social behaviour will be perceived to be more socially present (Pereira et al., 2014). *The AMSR will cause more social facilitation with human users than the SS and the AMR.*
- ☐ **H2.** While non-vebal behaviour should cause more social facilitation, a human-like design without non-verbal cues should not induce the same differences. *Differences in social facilitation will not apply between the SS and AMR.*
- ☐ **H3.** *There will not be any difference in task performance across the agents.*
- ☐ **H4.** As a conversational partner, *the AMSR will generally be preferred for the task.*

Experimental design

A within-subject design was used in a study (Kontogiorgos, Pereira, Andersson, et al., 2019) where participants interacted with all four agents. To test our hypotheses, we manipulated two independent variables [*embodiment* and *social eye-gaze*], in three conditions [*Smart Speaker (SS)*, *AntropoMorphic Robot (AMR)*, *AntropoMorphic Social Robot (AMSR)*], presented in different orders to participants using a Latin Square, and a *human trial* that was always first.

Task

We asked subjects to cook 4 variations of fresh spring rolls without providing the recipes; they had to find out the recipes while interacting with the agents. Different varieties of ingredients and amounts were used. The 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 (varied across conditions).

To ensure participants would engage with the agents more, they were told that if they followed the recipe with the correct ingredients and amounts, they would take the food with them at the end of the experiment. Counting the time participants took to cook the recipe served as a measure of the time they spent engaging with each agent. 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 (typically 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.

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 (Kontogiorgos, Pereira, & Gustafson, 2019). Given a human action or utterance, an appropriate response was selected from the dialogue policy. Driven by the possible set of actions, agent utterances are aggregated to higher level *dialogue acts* that describe the current state of the conversation. The dialogue acts model user actions, user utterances and any changes in the environment. An example dialogue:

USER : [FINISHED ACTION] What's next?
AGENT : [INSTRUCTION] Next, take three pieces of lettuce and put it in the spring roll.
USER : [CLARIFICATION-Q] Where is the lettuce?
AGENT : [CLARIFICATION-A] The lettuce is the green thing in the middle.
USER : [STARTED ACTION] Uh, yes!

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 decision making. The social behaviours were designed to maintain a socially contingent interaction with the subjects, and in order to keep the dialogues between the subjects and the agents consistent for comparison across conditions. The human wizard had to select the appropriate agent response, triggered only by the state of the environment and user actions. The wizard 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 generate, given the current dialogue act. The wizard therefore indicated only the current dialogue act in conversation.

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 (Kendon, 1967). Employing such a behaviour in agents, leads to human-like conversational turn-taking where each participant waits for the speaker's utterance before taking an action (Skantze, Hjalmarsson, & Oertel, 2014). 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.

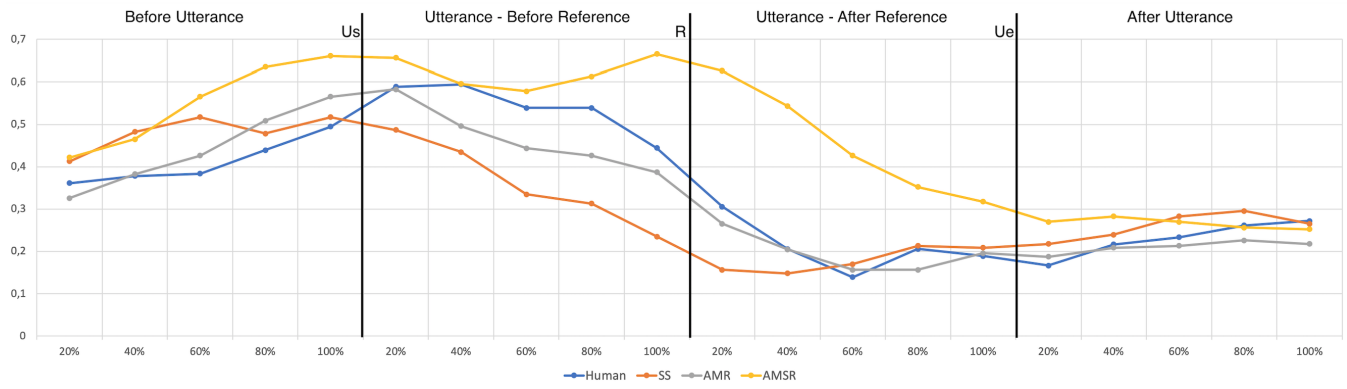

Figure 3: Gaze proportion to the agent during an agent instruction. Each phase of the instruction is normalised in time: Before the utterance - During the Utterance and before the reference - During the utterance and after the reference - After the utterance. The x axis shows the relative time of the instruction and the y axis the eye-gaze proportion across all participants per condition.

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 subject to establish the end of the turn and pass the floor to the subject. The agent gazed at referent objects right before they were mentioned (500ms before).

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. In the second phase, they cooked a recipe with the help of an agent. They repeated that phase 3 times with a new agent every time (counter-balanced). In the third phase, participants filled an exit questionnaire. During the agent trials, participants were alone in the room, and a human wizard 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.

The human instructor throughout the trial 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. The agents were situated on the sides of the table, with only the agent relevant to the task visible.

Participants

Participants 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 (stdev=5). The experiment was in English, and all participants were fluent (mean 5.8, stdev=0.7). 17 had interacted with a robot before and 20 had interacted with smart speaker. 13 had interacted with both a smart speaker and a robot before, while 6 with none of the two. Overall, their experience with digital technology was 4.8 (stdev=1.6) and their cooking skills were 5.0 (stdev=1.2). 24 had never cooked spring rolls recipes before. All scales above are 1-7.

Results

We present the main findings along two main themes: a) *visual attention*, and b) *interaction time*. Repeated measures analyses of variance (ANOVA) and post-hoc tests with Bonferroni corrections were carried out to test statistical differences across conditions. We report the behavioural and objective measures on visual attention during the agent utterances, interaction times (Table 2), and finally, notable insights from qualitative data.

Visual attention

Using motion capture, we detected subjects' head pose over time and measuring their visual angle (Kontogiorgos et al., 2018), we extracted proportional eye-gaze to the agent and the task table during different phases of the robot's utterances: a) before the robot speaks an utterance, b) during the utterance right before a reference to an object is uttered, c) during an utterance right after the reference has occured, and d) after the utterance. The four phases of proportional eye-gaze to the agent are presented in figure 3. Before and after the utterance phases are in 2 second intervals.

Each phase is first normalised per subject to reduce subject variability and then, each phase interval mean is used for comparison (Table 1). It is important to note that while agent conditions were counter-balanced in order, the human trial was always first to get familiar with the task.

Eye-gaze to the agent before the utterance. A repeated measures ANOVA to test the effect of gaze before the robot instruction showed a significant main effect, Wilks' Lambda = .674, $F(3,27) = 4.35$, $p = .013$). Post-hoc tests with a Bonferroni correction, and p value adjusted for multiple comparisons, revealed that gaze towards AMSR is statistically different than gaze to the AMR condition (p=.022) and to the Human trial (p=.029). No other statistical differences were found in pairwise comparisons.

	Human Trial		SS		AMR		AMSR	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Before Utterance	.4008	.1822	.4740	.2083	.4476	.2083	**.5472**	.2028
During Utterance (Before Reference)	.5402	.2131	**.3609**	.2321	.4659	.2057	**.6239**	.2114
During Utterance (After Reference)	.2119	.1749	.1836	.1605	.2006	.1320	**.4514**	.1851
After Utterance	.2273	.1570	.2524	.1372	.2072	.1301	.2599	.1633

Table 1: Mean eye-gaze to agent in different phases of the agent utterances. Each phase is normalised per subject and each phase interval mean is used for comparison.

	SS	AMR	AMSR
Task time (sec)	212.6 ± 7.93	217.2 ± 7.75	232.9 ± 8.52

Table 2: Interaction time in seconds: Each cell shows mean and standard error of the mean.

Eye-gaze to the agent during utterance (before the reference). A repeated measures ANOVA on the gaze before the reference showed a significant main effect, Wilks' Lambda = .483, F(3,27) = 9.65, $p < .001$). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons revealed that gaze towards AMSR is statistically different than gaze to SS ($p<.001$) and AMR (p=.001). SS was also different than the Human trial (p=.022). No other statistical differences were found in pairwise comparisons between the rest of the conditions.

Eye-gaze to the agent during utterance (after the reference). A repeated measures ANOVA on the gaze after the reference revealed a significant main effect, Wilks' Lambda = .316, F(3,27) = 19.49, $p < .001$). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons showed that gaze towards AMSR is statistically different than gaze to all other conditions ($p<.001$) and to the Human trial ($p<.001$). Here as well, no other statistical differences were found in pairwise comparisons between the rest of the conditions.

Eye-gaze to the agent after the utterance. A repeated measures ANOVA on the gaze after the robot instruction showed no statistically significant difference across conditions, Wilks' Lambda = .859, F(3,27) = 1.47, p = .244).

Interaction time

As indicators to task performance we measured *task time* (time from first to last agent action). We tested for comparison in time within the sequence of the conditions, and no statistical difference was found, meaning the condition sequence did not affect task performance (subjects were not significantly faster in the last trial). When compared across conditions however, a repeated measures ANOVA revealed a significant main effect, Wilks' Lambda = .739, F(2,28) = 4.94, p = .014). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons showed a significant effect between AMSR to SS (p=.041) and AMR (p=.023) but there was no evidence of a difference between SS and AMR (Table 2).

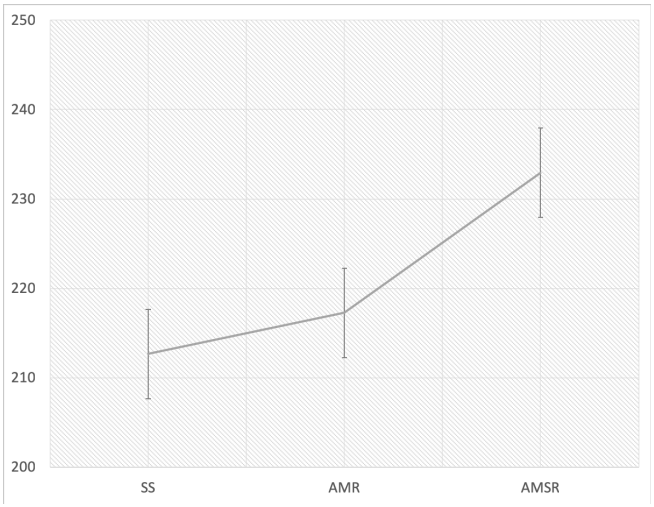

Figure 4: Mean *interaction time* per condition. Error bars indicate standard error of the mean (n = 30).

Qualitative data

The post-experimental questionnaire included asking participants to choose their preferred agent for the task and questions to elaborate on the preference. Participants were also asked to identify the differences of the three agents to understand if they are aware of what is tested in the experiment.

Perceived differences between agents. Out of the 30 participants, 18 replied this question. While the differences in the agent embodiment were obvious between [SS] and [AMR/AMSR], 66% of the participants did not notice a difference between [AMR] and [AMSR]. Asking participants further, 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. 69% of the participants preferred AMSR, while 24% preferred AMR and 7% preferred SS (χ^2 = 17.862, $p < .001$). Looking further at the participants 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.

Discussion

In an experiment with human subjects, we found a lack of positive effects in task performance on interactions with anthropomorphic social agents. Nonetheless, our findings show a higher degree of social facilitation in conversation with AMSR, as determined by subjects' visual attention and agent preference. Our strongest finding therefore is a trade-off between interaction time and social facilitation.

Anthropomorphism

The agents we compared represent different levels of embodiment in conversational agents. The most preferred agent for

the task had an anthropomorphic embodiment and a set of social eye-gaze behaviours. While AMSR was preferred, task time was increased by 10% with this agent in comparison to the less anthropomorphic in physical embodiment SS. We saw that participants looked at AMSR longer after the referent word was uttered 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 social facilitation, assuming they attribute that agent the role of a more socially present partner in conversation.

Non-verbal social behaviour

AMSR has joint attention afforded as an embodied phenomenon in its actions. Eye-gaze here is attributed as a social function where it regulates turn-taking, closer to how humans do when they interact with each other. AMSR therefore gave the impression that it is aware of the situatedness of the task.

In cases, it is possible the user may be distracted from the task through agent social behaviour because more attention is required to the agent's behaviour. While face-to-face collaboration is favourable due to its natural mediated channels of communication, interpreting social cues and maintaining attention is a timely and cognitively demanding process.

Social behaviour and task performance

Social behaviour is timely and counter-intuitive to task performance with more attentive agents. Nevertheless, task performance is certainly dependent on the nature of the task; in more task-oriented domains, such as emergency management, interactions may be more efficiency-prone. A human user may want to get the task done as quickly as possible, and get frustrated when having to interact longer than necessary. However, other tasks such as in the home-care domain are very dependent on social cues and interaction value.

As mentioned, referring expressions to objects did not contain any ambiguities in language (i.e. "this one here"). Therefore gaze from AMSR did not add value to task success but was attributed to a social function, as humans typically gaze at objects before mentioning them in language. Our purpose in the gaze condition (AMSR) was therefore not to increase task performance but to observe the social functions of gaze behaviour across agents.

We were able to verify hypothesis **[H1]**, that AMSR will cause social facilitation, as shown in the visual attention and preference dimensions, however with the cost of task performance. We did verify **[H2]** in the assumption that SS and AMR will not be different in social facilitation. In fact, a human-like design is not enough to establish rapport with human users; human-like behaviour may be expected too, when more anthropomorphic designs are manifested. The results also suggest that smart speakers, while embodied, do not facilitate the same turn-taking mechanisms as social robots do, likely due to the lack of non-verbal behaviours.

The results support **[H4]** reflecting that AMSR would be preferred for the task. We saw a wide difference between AMSR and SS, however AMR was also rated higher than SS, which may align with the fact that there is a relation to anthropomorphic agents with human-like designs, in terms of natural means of communication.

Most participants were more familiar with smart speakers than with social robots, which may indicate a novelty effect in the agent preference. Social robots are at time of writing emerging platforms and not as common and commercially available as smart speakers. However, we found that 2/3 of the participants were not able to identify the difference between AMR and AMSR, while they still preferred AMSR for the task. This indicates that the non-verbal behaviours used were subtle and asserted familiarity with the device.

Finally, we reject hypothesis **[H3]** reflecting that no differences would be found in task performance. Our assumption is that anthropomorphic facial features, without non-verbal behaviours is not enough to create more socially contingent interactions than SS: it is a combination of the two features that creates social facilitation to users.

Conclusion

In this paper, we presented a trade-off between task performance and social behaviour with conversational agents. Our contribution lies on an empirical evaluation of the anthropomorphic and non-verbal behaviour parameters of agents in task-oriented dialogues. This is particularly important to applications in which agents engage in a variety of tasks, and depending on the nature of the task, may need more or less social facilitation 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 interaction time and social behaviour is the focus of this paper. To fully address the aspect of a potential novelty effect, longitudinal studies need to be designed where users' experiences are tested in long-term interactions with social robots and smart speakers. Potentially, increased familiarity with AMSR could decrease gaze time to levels similar to a more familiar social agent (i.e. another human).

To understand which of the independent variables contributed to the general preference of the robot, we concluded that while an anthropomorphic physical embodiment affects social behaviour, a set of non-verbal behaviours also increase the interaction time with the agent. Further research should be conducted in a variety of HRI scenarios, to investigate variability in the nature of the task and its relation to social facilitation between human users and agents. In sum, despite the task performance shortcomings of social and situation aware robots, they do hold a good interaction paradigm for enabling social facilitation with users.

Acknowledgements

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 earlier versions of this paper.

References

Adalgeirsson, S. O., & Breazeal, C. (2010). Mebot: a robotic platform for socially embodied presence. In *International conference on human-robot interaction.*

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*.

Alam, M. R., Reaz, M. B. I., & Ali, M. A. M. (2012). A review of smart homes - past, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics*.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *Ro-man.*

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. In *Springer handbook of robotics.*

Breazeal, C., & Fitzpatrick, P. (2000). That certain look: Social amplification of animate vision. In *Aaai.*

Dourish, P. (2004). *Where the action is: the foundations of embodied interaction*.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*.

Goble, H., & Edwards, C. (2018). A robot that communicates with vocal fillers has... uhhh... greater social presence. *Communication Research Reports*.

Gomez, R., Szapiro, D., Galindo, K., & Nakamura, K. (2018). Haru: Hardware design of an experimental tabletop robot assistant. In *International conference on human-robot interaction.*

Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE*.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*.

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. In *Iros.*

Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *Iros.*

Koda, T., & Ishioh, T. (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.*

Kontogiorgos, D., Avramova, V., Alexandersson, S., Jonell, P., Oertel, C., Beskow, J., ... Gustafsson, J. (2018). A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Lrec.*

Kontogiorgos, D., Pereira, A., Andersson, O., Koivisto, M., Gonzalez Rabal, E., Vartiainen, V., & Gustafson, J. (2019). The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *International conference on intelligent virtual agents.*

Kontogiorgos, D., Pereira, A., & Gustafson, J. (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.*

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (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*.

Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing social robot, screen and voice interfaces for smart-home control. In *Chi conference on human factors in computing systems.*

Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies*.

Mizoguchi, H., Sato, T., Takagi, K., Nakao, M., & Hatamura, Y. (1997). Realization of expressive mobile robot. In *Robotics and automation.*

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? psychological responses to personality types in human-computer interaction. *Communication research.*

Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*.

Pereira, A., Prada, R., & Paiva, A. (2014). Improving social presence in human-agent interaction. In *Sigchi conference on human factors in computing systems.*

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.*

Shibata, T., Tashima, T., & Tanie, K. (1999). Emergence of emotional behavior through physical interaction between human and robot. In *Robotics and automation.*

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*.

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

Thellman, S., Silvervarg, A., Gulz, A., & Ziemke, T. (2016). Physical vs. virtual agent embodiment and effects on social interaction. In *International conference on intelligent virtual agents.*

Torta, E., Oberzaucher, J., Werner, F., Cuijpers, R. H., & Juola, J. F. (2013). Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. *Journal of Human-Robot Interaction*.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2006). The role of physical embodiment in human-robot interaction. In *Roman 2006.*