The Effect of Audiences on the User Experience with Conversational Interfaces in Physical Spaces

Heloisa Candello
IBM Research
hcandello@br.ibm.com

Claudio Pinhanez
IBM Research
csantosp@br.ibm.com

Mauro Pichiliani
IBM Research
mpichi@br.ibm.com

Paulo Cavalin
IBM Research
pcavalin@br.ibm.com

Flavio Figueiredo
Fed. Univ. of Minas Gerais
flaviovdf@dcc.ufmg.br

Marisa Vasconcelos
IBM Research
marisaav@br.ibm.com

Haylla Do Carmo
IBM Research
hayllat@br.ibm.com

ABSTRACT

How does the presence of an audience influence the social interaction with a conversational system in a physical space? To answer this question, we analyzed data from an art exhibit where visitors interacted in natural language with three chatbots representing characters from a book. We performed two studies to explore the influence of audiences. In Study 1, we did fieldwork cross-analyzing the reported perception of the social interaction, the audience conditions (visitor is alone, visitor is observed by acquaintances and/or strangers), and control variables such as the visitor's familiarity with the book and gender. In Study 2, we analyzed over 5,000 conversation logs and video recordings, identifying dialogue patterns and how they correlated with the audience conditions. Some significant effects were found, suggesting that conversational systems in physical spaces should be designed based on whether other people observe the user or not.

CCS CONCEPTS

• **Human-centered computing** → Empirical studies in HCI; Empirical studies in interaction design.

KEYWORDS

Conversational interfaces, audience effects, chatbot design.

ACM Reference Format:
Heloisa Candello, Claudio Pinhanez, Mauro Pichiliani, Paulo Cavalin, Flavio Figueiredo, Marisa Vasconcelos, and Haylla Do Carmo. 2019. The Effect of Audiences on the User Experience with Conversational Interfaces in Physical Spaces. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300320

CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300320

1 INTRODUCTION

Conversational machines are being increasingly employed in physical spaces for both private and public usage. Examples include hotel lobbies and store showrooms [15], car dashboards [24], and home devices [40]. With such machines, or chatbots, human interactions may happen in the presence of an audience, be it friends, family (e.g., while on a road trip), or simply strangers and bystanders (e.g., in a hotel lobby).

Previous studies have found that humans tend to change their normal conversation behavior when in front of others. In such contexts, people sometimes resort to using long and complicated words, uttering jokes, quoting from obscure authors, and, in general, pretending to be smarter, wittier, or funnier than in private conversations [4, 35]. Also, when in the presence of others, some people may enhance the emission of dominant responses [8], according to the status of those in the audience [8, 10]. However, some people react in the opposite way, becoming more shy than normal, failing to complete sentences, getting nervous, or even stuttering.

Understanding such changes in behavior are important because it may be necessary to design the machine conversation systems to handle those situations where people change their usual behavior to accommodate the presence of an audience. We were motivated to study this kind of behavior change by some initial observations we made of visitors experiencing an art exhibit where they interacted with a group of chatbots either alone or in front of acquaintances and/or strangers. For example, we observed some people trying to amuse their friends by trying to "break" the machine with impossible questions; asking questions related to local politics and sports to provoke the other visitors; and uttering deep and complicated questions to show off to others their knowledge about the subject of the artwork.

Figure 1: The physical setup of *Coffee with the Santiagos*.

In any of those cases, we found that the art exhibit could have been designed to better handle the presence of an audience. For example, the system could have a higher threshold of guessing the right answer to complex questions when an audience is present. It could, for instance, assume that a visitor asking a complex question in front of an audience is less a situation where she is looking for knowledge and more like the system is being made fun of. While in the former case the appropriate response could be trying to find a good answer as hard as it could, whereas in the group situation it could be simply deflecting the question.

Beyond art exhibits, as conversational systems become more ubiquitous, similar situations will be common in more down-to-earth scenarios. For instance, a conversation speaker (like Amazon's *Echo* or Google's *Home*) could benefit to adjust its behavior to handle audience effects. It could be less prune to making jokes to avoid making anyone in the audience uncomfortable, or, even worse, feeling ridiculed in front of acquaintances. In other words, by recognizing the audience context, the conversational system may be designed to answer in a more appropriate form for a situation of group social interaction, adapting to and enhancing the overall experience of users and their audiences.

However, such considerations and strategies only make sense if we understand whether and how users change their behavior when conversing with machines in front of other people. Do they feel more embarrassed, powerful, or wittier by an audience when dialogging with a machine instead of a person? Are the changes different if the audience is comprised of acquaintances or strangers? To shed a light on such questions, we went further and performed two studies on the art exhibit and its visitors. This was a setting where single or multiple visitors freely conversed in a physical space with three text-based chatbots representing characters from a well-known 19th century book in Brazil. No control on how visitors interacted with the space or the chatbots was in place, with the exception that they had to do it through a single tablet. Images from the exhibit are depicted in Figure 1.

To explore changes in conversational behavior of people due to the presence of an audience, we investigated the visitor perceptions of the three agents' social skills and the user's engagement with agents with the artwork. In the majority of the situations the interaction happened in front of other visitors, some of them known to the users, but also often in front of strangers. In our first study, we conducted 92 semi-structured interviews with visitors, after observing their behavior at the exhibit. Analyzing this data, we were able to determine that, in some specific situations, it was very likely that the audience presence was affecting the user experience of the visitors. In a second study, we analyzed the conversation logs of more than 5,000 sessions. Coupled with a silent video of the audience interaction, which we used to manually determine the occurrence and type of audience, we were able to explore changes in conversation patterns which could be related to the presence of other people around the visitor. The two studies seem to provide evidence of audience effects, and that designers should be taken into account audience effects in conversational systems in physical spaces. Moreover, our findings seem to indicate that those effects are modulated by many factors, including gender, knowledge about the content of the exhibit, and whether there were strangers in the audience.

The next sections describe in detail the related work, the experimental setup, the two studies, their findings, and our

main conclusions. Finally, we discuss some design implications, indicating how our findings may guide the design of conversational systems in physical spaces.

2 RELATED WORK

In this session, we describe the previous work as a background for our study, both in the scope of social interaction with chatbots and in the context of audience effects.

Social Interaction with Chatbots: With the recent advances in conversational and natural language technologies, interest has increased on how humans interact with conversational systems, here referred generically as chatbots, and on how social presence and context may play a key role in understanding the dynamics of the interaction [29, 37].

Social presence is described as the social connection and involvement between two or more people in an interaction often developing and maintaining some sort of personal relationship [41]. The perception of social presence is sometimes connected to the anthropomorphism of physical robots, chatbots, and avatars. In particular, anthropomorphism is a prevailing topic of Embodied Conversational Agents (ECAs), a special case of embodied agents in which the agents provide human-like capabilities of face-to-face dialogue.

Studies with ECAs have provided evidence that they can induce social-emotional effects comparable to those in human-to-human interactions [38, 43]. Previous work found that people conversing with ECAs or interacting with robots show social reactions such as social facilitation or inhibition [3, 38, 50], a tendency to socially desirable behavior [20, 39, 43], and increased cooperation [32]. For example, analyses of users' utterances while interacting with a museum agent [19, 20] showed semblance with human-to-human communication, with similarities in the amount of greetings and farewells, common phrases (such as "How are you?"), and human-likeness questions (e.g., "Do you have a girlfriend?").

In general, system which exhibit human-like traits tend to improve the quality of the user experience with them., Cafaro et al. [6] found that smile, gaze and proxemics are important for conversational museum guide agents, implying that those agent influenced user's interpretation of agent's extraversion and affiliation and impacted on the user's decisions about further encounters.

Although the degree of veracity in the dialogue often improves the quality of the interaction, it might have the opposite effect: the uncanny valley effect [30, 44] where people are averse to a high degree of human similarity has also been observed. Experiments, such as [18, 32], have validated this hypothesis by observing the user's emotion engagement strategies towards agents of varying human likeness.

In this study, we contextualize our study object, the art exhibit, as containing three embodied chatbots. Even though the chatbots did not have a physical body they have a clear physical presence provide by scenographic elements (see Figure 1): female and male hats hanging above chairs around a table unmistakably embodied the chatbots.

Audience Effects:. Seminal work on drive-producing effects of the presence of an audience [8] uncovered specific group interaction behaviors, which led to theories and design frameworks for spectatorship (e.g. [4, 35]). Among the implications and findings of audience effects are the impact of behavior and views of bystanders on the response to an interaction, which has been known to influence engagement, either being related to attention, interest, or affective feelings [4, 8, 35].

One of the early studies of audience effects concluded that proximity and presence of audience enhance the emission of dominant responses [8], i.e. responses governed by strong verbal habits at the expense of responses governed by weaker ones. Active audiences who looked and interacted with the subjects directly affected individual performance measured by the average number of responses in a word recognition task. In 1982, Michaels et al. performed a classical study on social facilitation showing that the performance of good pool players improved 14% in front an audience while bad pool players had a dramatic decrease of 30% [28].

Love and Perry [23] studied the behavior and views of bystanders in response to a proximal mobile telephone conversation held by a third party. In their experiments, subjects demonstrated noticeable changes in body posture when viewing and listening to a confederate attending a call. The influence of audience has been also studied in video gaming, where researches explored audience aspects including age [49], size and distance of the interactor [18], typologies of spectatorship [27], player performance and perceived game difficulty [49], co-located/remote and virtual/real audience [11], cheering [16], supportiveness [5], activeness [21], and social aspects [9]. Overall, the findings report that different characteristics and behaviors of audience have positive and negative impacts, sometimes affecting the entire gameplay experience.

Spectator experience design has been proposed by Reeves et al. [36], which produced a taxonomy that uncovers design strategies based on interface manipulations and their resulting outcomes. Audience participation in public spaces has also been studied from the point of view of interaction and engagement in many domains, such as education [47], sports [7], and arts [21]. One common observed practice which directly affects the experience is the honey-pot effect, where interaction with a screen in public can drive social clustering and further engagement [4]. Furthermore, Group interaction helps to explain how users understand and react to displays in public settings.

Although previous works explored audience and spectatorship effects in games, sports, arts and other domains, to

the best of our knowledge no research efforts have been made to study the experience of audience effects in scenarios where the main interaction is conversing with chatbots in a physical space. Finally, given that our setting is an art exhibit, we use the terms *visitor* and *user* interchangeably. In our study, a person is both a visitor of the exhibit as well as the user of the physical chatbot architecture described next.

3 THE EXPERIMENTAL SETUP

In this section, we begin by describing the physical artwork which was part of a large art exhibition of a Brazilian arts center. Next, we discuss our experimental setup and the research questions tackled using the data from the exhibit.

The Artwork: Coffee with the Santiagos

The study reported in this paper was done in the context of an artwork called *Coffee with the Santiagos* by three Brazilian artists. It recreated a dining room of the 19th century populated with physical representations of characters from one of the most well known and acclaimed novels in Brazilian literature: *"Dom Casmurro"* by Machado de Assis (the novel was originally published on 1899).

The choice of the book was made to facilitate the visitors experience by engaging them with familiar material and characters. Also, the novel is known for not being clear about some events of the story, specially whether the main character, *Bentinho*, was betrayed by his wife *Capitu* with his best friend *Escobar*. *Bentinho* is tormented with the resemblance of his son with *Escobar*, but the actual betrayal is never described in the book or admitted by *Capitu*. Therefore, we expected that visitors could interact within the context of the book from the start, and often they did so by asking the *Capitu* chatbot whether she had betrayed or not her husband (which she vehemently denied).

Figure 1 shows images from the exhibit. In the center of the space there was a large table with cups, saucers, and a teapot arranged as if the characters were having coffee. Around the table there were four chairs, three of which were"occupied" by the main characters of the book, represented by floating hats. Attached to the fourth chair there was a tablet and a headphone which allowed visitors to interact with the bots. Sounds of barking dogs, horse carriages, and singing birds were played as environmental sound.

The utterances from the characters and the visitor were seen as animated text projections on the table, as if departing from the cup in front of each character or the visitor (see right image of Figure 1), in an effect similar to White and Small's *Poetic Garden* [48]. The path of the text on the table allowed it to be read independently of the position of the visitor around the table. The projector also changed the color of the cup associated with the current speaker to help visitors understand from which character the utterance was coming

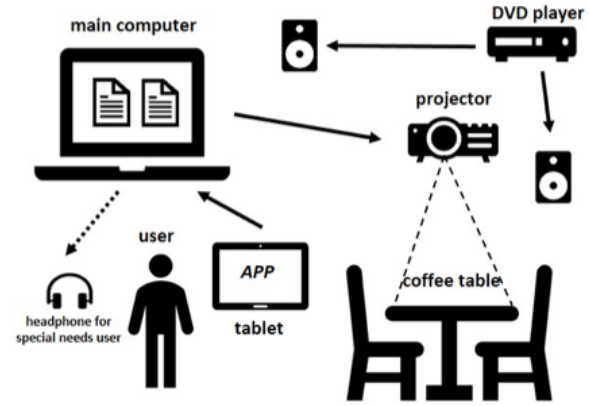

Figure 2: Physical architecture of *Coffee with the Santiagos*.

and varied the decorative motif of the saucers by projecting images of 19th century watercolors on them.

The physical architecture of the *Coffee with the Santiagos* is shown on Figure 2. The main interaction with the system was performed with a custom tablet application which captures the visitor's input (name, gender, and utterance/question) and sends it to a main computer. The IBM *Watson Assistant* API service was employed to obtain the character response for the visitor's utterance from a set of pre-defined answers captured from dialogues of the original book. Other web API services were also called to generate the narrated audio of the questions (for visitors with special needs). Both characters' and visitor texts were displayed on the table by a projector and the generated audios were played on the headphone. The ambient soundtrack was reproduced on loop by a DVD player connected to sound speakers.

When visitors arrived at the exhibit they could see the characters conversing with each other and an inviting message on the tablet. If they decided to interact, they had to enter their names and gender, select the character to which they wanted to send a message, and then type the message (with the aid of auto-correction and completion). The visitor's message was projected on the table and was followed by a reply from the selected character. Replies from characters were mostly based on actual dialogue sentences from the book. If the utterance of the visitor was deemed to be beyond the scope of the book or not recognized by the chatbot, the character tried to divert the question by asking "More coffee?" to the visitor.

For some randomly selected visitors, the replies would also contained a direct address to the visitor in the form of a vocative, such as "*No such a thing, dear Maria.*" The use of direct address intended to enhance the impression that the character was talking back to the visitor and was part of another study being conducted at the art exhibit.

Research Questions

We carried out an exploratory study in the wild during the last three weeks of the artwork in August of 2017. In this period, 5,100 people interacted with the exhibit. Those participants were videotaped (without sound) and close to 10,000 questions were logged. Out of these visitors, 92 were observed while interacting with the exhibit and asked to participate in a semi-structured interview (described later).

We performed two studies to understand the effect of audiences on the visitor's experience and in particular on how it affects the visitor's utterances with the chatbots; and the visitors' perceptions of the social interaction with them. As the interaction happens in a physical space, the users and the audience are always co-located. Formally, we pose our two research questions as:

RQ1: What are the effects of audiences on the users' perceptions of social interaction with chatbots?

RQ2: Does the presence of audiences influence the type and content of user's questions directed to chatbots?

We conducted the observations and semi-structured interviews to answer RQ1 in Study 1, and we analyzed the collected conversation-logs and video recordings to answer RQ2 in Study 2. In both studies we classified the interaction sessions into four non-mutually exclusive audience conditions based on the videotape and notes from the observation studies. The four audience conditions are:

(A) No audience: the visitor was either alone or no one was observing her/his experience in the artwork.

(B) Observed by acquaintances: the visitor was accompanied by friends or acquaintances who sometimes also shared the tablet.

(C) Observed by strangers in the queue: the visitor, *with acquaintances or not*, had strangers observing her experience from the artwork queue.

(D) Observed by strangers standing around the table: the visitor *with acquaintances or not*, had strangers observing her experience from around the table.

The (C) and (D) conditions were determined by considering the proximity of the audience to the users of the tablet. Condition (C) was annotated by researchers (authors) that were co-located with the exhibit while observing visitors. Condition (D) was analyzed from the video recordings as it was done in [23]. Unfortunately condition (C) was not determinable from the video since the queue was out of the field of view of the camera, and therefore only the first study examines this condition.

Table 1: Self-metric answers of participants in Study 1.

	Dis.	Neu.	Agr.
Q5: I felt part of the conversation	18	9	65
Q6: The characters only talked to each other	64	5	23
Q7: The characters talked to me	16	5	71
Q8: The characters answered my questions	30	3	59
Q9: The characters answer about any subject	36	23	33
Q10: I asked everything I wanted	18	4	70

4 STUDY 1: AUDIENCE EFFECTS ON SOCIAL INTERACTION

The objective of the first study (addressing RQ1) is to understand the visitors' experience and examine whether audience affects or not the perception of their social interaction with the art exhibit chatbots. For this study, we focus on the 92 users which had their interaction observed and completed semi-structured interviews.

Procedure

During the last three weeks of the exhibition period, some of the authors of the paper randomly selected visitors from the art exhibit to observe. As soon as such visitors had finished their interaction with the artwork, they were invited to participate in an interview. In total, 92 participants agreed to participate in the study and signed a consent form before starting the interview. The semi-structured interview was designed to be short and took between 8 and 10 minutes to be completed. The questions in the interview were:

Q1. Are you familiar with the story of the book?
Q2. Please tell us how you would describe your experience with this exhibit for a friend who will not be able to visit it.
Q3. How do you think the exhibit works?
Q4. Did you hear about this exhibit before visiting it today?

Participants also answered self-metric statements (Q5 to 10 on Table 1) by choosing one option among five: *totally disagree, partially disagree, neutral, partially agree,* and *totally agree.* The perception of social interaction was measured using those self-reported metric questions which explore how participants had a sense of belonging to the talk vs. a sense of being an outsider from the talk (Q5 to Q8); the sense of perception on the chatbots' scope of knowledge (Q9); and the sense of satisfaction (Q10).

At the end of the interview, participants had the chance to share any other thoughts they wished. Additionally, participants answered demographic questions. Data collection was stopped when findings were becoming repetitive, achieving data saturation [45]. Interviews were audio-recorded, transcribed, and analyzed using a mix-methods software package. The qualitative data gathered was coded in categories. The

Table 2: Demographics of the participants in Study 1.

Participants		92
Age group	16-26	46
	27-37	19
	38-48	16
	49-59	7
	60-72	4
Gender	Female	50
	Male	42
Familiarity with book plot	Familiar	56
	More or Less	20
	Not Familiar	16

Table 3: Number and percentage of participants according to the audience condition in Study 1. Notice that a user observed by acquaintances may also be observed by strangers standing in queue (B ∩ C) or around the table (B ∩ D).

condition	A	B	C	D	B ∩ C	B ∩ D	all
# of users	4	65	51	73	34	55	92
% of users	4%	71%	55%	79%	37%	60%	100%

coding scheme emerged from the data by applying an inspired grounded-theory approach [45]. A thematic network was applied as an analytical tool to understand better the themes emerged from the conversation logs [1]. Additionally, we also collected the conversation logs of the 92 participants from the system log and integrated this data to this analysis, so we could understand better their experience with the chatbots.

Demographics of Participants

Our participants' demographics are shown on Table 2. 65 of our participants were aged between 16 to 35 years old (71%), with gender being roughly balanced. 76 of the participants had familiarity with the story (83%). The duration of the participants' interaction experience with the artwork was between 5 and 8 minutes. In Table 3 we present the audience types. Notice that 65 participants (71% of the total) were observed by acquaintances (B). Participants who were observed by strangers were in one or two conditions, either they were being watched from a queue of people waiting to use the tablet (C) or they were watched by other visitors from around the table (D). For the purpose of this analysis, we preferred to leave those conditions separated due to the different levels of audience effects that may arise in each condition [23].

In this sample only four participants were interacting alone and did not have any audience (A). For this reason, we focused the analysis of this study more on the (B), (C) and (D) conditions, and in particular on understanding whether being with acquaintances or not makes a difference on the social interaction with the chatbots. We also consider gender, age, and familiarity with the plot as independent variables.

Thematic Network Analysis

Initially, we describe here the trends and themes that emerged from the user experience of being in the context of the artwork. The thematic network approach was applied to organize the themes and understand the overall experience. The source of this analysis were the open-ended questions (Q1 to Q4) from the 92 semi-structured interviews, notes from observation studies and video recordings, and the rationality reported by the participants when they answered the self-reported questions (Q5 to 10).

The semi-structured interviews and observation studies unveiled a set of emerging themes which helped to create a picture of the visitors' experience with the artwork. In total 125 basic codes emerged from the analysis, which were grouped and organized into four main organizing themes, identified as important factors affecting the visitors' social interaction with chatbots. The organizing themes are: curiosity and novelty; interest on the plot; expected chatbots answers; and audience effects.

Curiosity and novelty: It seems that the scenographic elements created an atmosphere which attracted participants to interact (as in [31]). The decorative hats representing each chatbot as seating at the table, the 19th century dining room complete with wall-paper, tapestries, and surrounding sound created an atmosphere which attracted interaction. When reporting their experience to the researcher (Q2) it was evident that several scenographic elements provoked their curiosity. (P27) "*The soundtrack was flawless. [...] And I also noticed the animation of the words on the table and the lights when someone responds, and it lights up and highlights colors inside the cups, I found this to be very cool!*". Participants also pointed to the utterances projected on the table and tried to figure out what the best question to ask was when they are accompanied by acquaintances. Others reported that the scenography and the empty space in the chairs reminded them of ghosts similar to a *table-turning séance*. (P14) "*I thought the visual was very interesting, it called my attention. You see those hats, it seems like you're talking to the beyond. It was intriguing... very curious.*"

Expected chatbot answers: Most of the users seem to have a mental model of how the chatbots were supposed to answer them. Some of the participants wanted only to test the technology behind the exhibit and were not focused on the content of the plot. (P4) "*For me, I was more curious to see what happened then ... I did not develop several questions, I was not too involved with the intention of having a conversation.*". P(40) expressed his intention when he asked the chatbots a topic which he presumed the chatbots would not be able to answer him: "*I asked when we will have peace in the world. I guess I was*

not too fair on him, I was very picky. I confess I did a very tricky question. I wanted to challenge the software.". However, for others the technology was invisible. (P42) "*I enjoyed! [laughs], I didn't try to imagine how it was working.*" In those cases, participants reported as if they were immersed in a parasocial interaction [14], as highlighted by the language they used to describe their experience and the questions they asked. They also applied human social interaction rules [25, 34], using personal pronouns and asking questions related to human nature of the chatbots. As it was evident in the words of (P1): "*Her (Capitu) response was genuine. The answer for the other two (Bentinho and Escobar), were neutral. As hers was more elaborated I felt that she had thought to answer that. It was not just a game*". Participants also consulted the chatbots as oracles. (P35) "*Do you believe in free love?*" and reported their emotional states to the bot. (P23) "it I'm scared. Are you someone who scares people?"

Interest on the plot and characters: Most of the participants seem to have only the betrayal question in mind and reported they were satisfied with the number of questions they asked. Participants who knew the story usually interacted the same amount of time as the ones that did not know the plot, usually asking two to three questions. As mentioned before, some chatbot utterances projected on the table contained the name of the participants (ex. "it Dear Paulo"). In those situations, participants were engaged with the dialogue such as (P43): "*I wrote my name and the exhibit interacts with me, so to the point that as if I was there, because she (Capitu) speaks my name, she calls me lady and such. And it was very interesting because you can really ask any questions, I thought you had pre-selected questions. Do you know what I mean?*". Participants also tested the chatbots to validate their answers when they knew the plot and had several opinions about which character should be blamed for the betrayal. Others only have one question they wanted to know, whether the betrayal had happened or not[1].

Audience effects: We identified that people waiting in the queue influenced the participants' interaction. Participants reported they would have asked more questions, if there was not a queue waiting for the interaction. As described by (P23): "*[. . .] If I had known that I was not disturbing I would have asked more questions, I had a lot of those thoughts: oh, my God there's someone else waiting to see. I already had two interactions, so it was already ... it was already good*". When visitors were observed by acquaintances (B), it tended to enhance certain behaviors of appreciation and visitors appropriated the exhibit to communicate their feelings. (P29) wrote to his sweetheart: "*I love you, Mrs. Tatiana*". Finally, we also saw participants asking questions not related to the plot when in the presence of strangers (C and D) such as (P53): "*Why do you think people are so cold?*"; and (P40): "*When are we going to have peace in the world?*".

[1]This "fundamental" question is never answered in the book *Dom Casmurro*.

Statistical Analysis of the Self-Reported Questions

We performed a statistical analysis based on the self-reported questions (Q5 to 10). The answers to the self-reported questions (Q5 to Q10) were joined from the 5 categories into 3 to better perform a statistical analysis to complement the qualitative findings: *D= totally disagree, partially disagree; N= Neutral;* and *A= agree, totally agree*. Results are shown in Table 1. Those questions served as response variables, capturing the users' perception of social interaction with the chatbots. They were cross-sectioned with relation to age, gender, familiarity with the plot, and the different audience conditions (A to D) using contingency tables. To understand the table, each column represents answers from Q5-10 (disagree, neutral and agree), or responses. The rows capture different participant conditions (e.g., gender or familiarity). The cells thus counts the intersection between rows and columns (e.g., females who answered agree on Q6).

Our statistical analysis is based on the *Fisher exact tests*. The null hypothesis of the test states that responses are uniformly distributed on the contingency table. Rejecting this hypothesis serves as evidence that the different conditions or demographic variables led participants to indicate that the social interaction between bots in each of the four conditions (A, B, C, D) affected their experience.

From our 92 participants, only 4 were in condition (A). Because of this, our study mostly focuses on the other conditions (B to D). Similarly, when isolating some of the variables (e.g. females), the number of participants naturally reduced. Nevertheless, for all of the cases where we found some statistical significance ($p < 0.05$), we had at least $n = 32$ participants. Before continuing, we point out that we did not find any statistical significance in our sample when analyzing the effects of audience considering age and also on (Q10). Most participants (77%) reported they asked all the questions they wanted, therefore they were satisfied.

Observed by acquaintances (B): In this condition participants were observed by acquaintances, sometimes sharing the interaction experience with them ($n = 65$). Here, 59% (45 out of 65, $p < 0.05$) of participants felt that the characters answered their questions more often (Q8). In the other conditions (C and D), that is, who experience the exhibit alone or only observed by strangers, this percentage was of higher (78% out of $n = 27$). This may indicate that participants in condition (B) were less able to pay attention to the chatbot answers or that they were less focused on the conversation with them, perhaps distracted by the acquaintances. However, the same participants in condition (B) who had previous knowledge of the plot (51 out of 65) significantly disagreed (78.4%, $p < 0.05$) that the characters only talked to each other

(Q6), showing it might be a higher degree of engagement with the artwork when sharing the experience with acquaintances for users who understand better the context. Gender did not have any effect in this category.

Observed by strangers in the queue (C): In this condition participants were observed by strangers in the exhibit space usually next to or behind them in a queue ($n = 51$), with or without acquaintances. This group of users mostly indicated (45 out of 51, 88.2%, $p < 0.01$) that the characters talked directly to them (Q7).This is an interesting effect of participants who, maybe inconvenienced by people waiting in the queue, may have tried to focus more in the interaction with the chatbots. Participants in condition (C) who were **also** familiar with the plot ($n = 44$) indicated that the chatbots talked to them (Q7), (41 out of 44, 93.2%, $p < 0.05$). These same participants agreed that they felt part of the conversation (Q5), (34 out of 44, 77.3%, $p < 0.05$), and that the characters answered their questions (Q8), (35 out of 44, 79.5%, $p < 0.05$). When cross-sectioning with gender variables, female participants ($n = 32$) felt more often the characters talked to them (Q7) (25 out of 32, 87.5%, $p < 0.05$), though we did not find any significant result for males.

Observed by strangers standing around the table (D): In this condition participants were observed by strangers while typing their questions on the tablet ($n = 73$), some also observed by acquaintances. No statistical significance was found within the distributions of the self reported answers of (D), or in comparison with participants not in this condition. However, we found significant findings when we considered some of the independent variables (gender and knowledge of the plot). Of the users in this condition who **also** had knowledge of the plot ($n = 60$), 40% (24) of them disagreed that the characters answered about any subject (Q9) ($p < 0.05$). We looked in the conversation logs and found that most of the questions asked here were out of scope questions, but featuring a curiosity on the chatbots' opinion about the story. For instance, (P46) "*Capitu, what do you think of Bentinho's brain sickness?*", poses a question for which the book does not have an answer. (P19) goes deeper into the plot "*Who is the father of your son?*". (P24) humanizes the chatbots "*What is your favorite color?*".

Similar results were observed when we examined (Q9) considering gender. Of the females in this condition of having an audience of strangers around the table ($n = 41$), 43.9% (16) disagreed that the characters answered about any subject ($p < 0.05$). Many of the female questions relied on curiosity about the relationship among characters. In the same question (Q9), only 21.9% (7 out of $n = 32$) of the male participants felt the same as females ($p < 0.05$). From the conversation logs, we found that often male questions were in the first person, giving opinions, showing off, such as doing love declarations or trying to be humorous: (P11) "*Am I handsome?*"; (P41) "*Yes, I think.*"; or (P59) "*I only have coffee with cigarettes, do you?*" Therefore, we found evidence that male participants disagreed less than females that the chatbots answered about any subject, but it seems that the type of out-of-scope questions made by each gender was not the same. This suggests that the social interaction may have been different comparing male and female participants in the condition of strangers standing around the table.

5 STUDY 2: AUDIENCE EFFECTS ON USER INTERACTION

In this second study, we tackle our second research question *RQ2: Does the presence of audiences influence the type and content of user's questions directed to chatbots?* by exploring a larger dataset consisting approximately 5,000 utterances logged during three weeks at the exhibition, combined with information extracted from the silent video recordings of the interactions. In this analysis, we employed both manual coding and machine learning tools to analyze the dataset using a *semi-supervised* approach.

Coding the Audience Conditions

Given that our ultimate goal is to understand audience effects, we filtered our initial dataset to consider only the user interactions where we could be sure to determine the audience conditions (A, B, and D) from the video. Unfortunately, we could not determine condition (C), when users were observed by strangers in a queue, because the queue was not visible in the video recordings. In the previous study, this was possible because researchers were observing the participants in the field.

To code the different audience conditions (A, B, and D), independent human coders analyzed the video recordings from the exhibit. A sample of 54 hours of video (four weekdays and two weekends) was analyzed. This sample was chosen since it captured different types of public (e.g., weekday visitors and weekend visitors). Based on those recordings, sessions were determined to be in one or more of the three settings A, B, and D. Then three different human coders manually observed and annotated features of the movement across the room, communication cues, proximity, presence of a companion, and closeness among the visitors while a participant interacted with the tablet. In the end, we considered 633 sessions which were subdivided into 1,542 same-user interactions (each session sometimes included several consecutive users). Out of those user interactions, 240 were coded as (A), no audience; 1200 were coded as (B), observed by acquaintances; and, 102 were coded as (D), observed by strangers standing around the table. Considering only those interactions, our resulting dataset contained approximately 5,000

user utterances. Since the majority of utterances were questions to the chatbots, we use from now on the latter term instead of the former. Each question was then coded in a *semi-supervised* manner as discussed next.

Clustering User Questions into Topics

After coding the audience condition from the videos, we proceeded to cluster the topics of the questions from the users. Given the rich and distinct way users asked questions, instead of manually coding each sentence individually, we initially used a semi-supervised methodology to cluster the full set of 5,000 user utterances into 32 clusters of semantically-similar sentences. In our approach, we first employed a clustering algorithm which found an initial set of meaningful topics. Next, a manual, open-coding of the clusters was performed by two coders to validate the automatic clustering. We compared different methods such as *K-Means* [22], *Spectral Clustering* [26], and *Hierarchical Dirichlet Process* [46]. We proceed by considering only the approach which showed the best results according to the two coders who inspected the results, which was *Spectral Clustering*.

Clustering Methodology: Given the set of user questions, our clustering methodology is based on converting this set to an affinity space S. That is, each utterance q_j, where $1 \leq j \leq N$, is represented as a N-dimensional vector, where position i represents the similarity of question q_j to question q_i. This process resulted in a $N \times N$ square matrix, denoted S, where all the values in the diagonal are equal to 1. For computing the similarity, we employed *Spacy* [42], a popular natural language processing tool which determines the similarity between two sentences as the Euclidean distance between the average *word vectors* [33] of each sentence. Word vectors refers here to the known technique of finding semantic similarity across words by embedding each word into a low K-dimensional space (usually $K = 100$). The average word vector of a sentence can thus be used as a proxy for the semantics of that sentence. This technique is known to have limitations for large texts. However, the sentences employed by users to interact with chatbots had on average 4 words (with a standard deviation of 2 words). For such short texts, average word vectors are known to perform well in machine learning tasks, and particularly in clustering [17]. In our analysis, we employed *Glove Portuguese* vectors [33].

After computing the affinity space S, clustering was conducted based on two steps. First, we filtered out rows in S where all values were equal to zero, except the value 1 in the diagonal. That is, all the questions which had null similarity to the other questions were separated and put into a special, single cluster denoted "Anything Else". Then, the remaining rows in S were clustered with the *Spectral Clustering* algorithm [26]. This process resulted in 32 clusters, which were then manually corrected by two human coders.

From Clusters of Questions to Topic Clusters: To associate each cluster with a meaningful topic, the same two human coders named each of the 32 clusters separately. Then, the coders met and discussed what they had found as meaningful identifications for the clusters and did some validation steps (e.g., manually merging or splitting clusters which presented similar or different content, respectively).

After agreeing on the clustering, the coders worked on the filtered dataset of 1,542 interactions together to clusterize further. In the end, four main topic clusters were identified: (S1) questions *out of scope* of the original book; (S2) questions *about characters* of the book; (S3) *greetings* such as *Hello* and *Goodbye*; and, (S4) *reaction to failure*, which corresponds to the user reaction when the chatbots deflected questions they did know how to answer with *More Coffee?* (for example, a user replied that indeed she wanted some coffee). After the coding was performed, we measured the *Cohen Kappa* agreement [12] between the coders. Overall we found a reasonable score of 0.78 (strong agreement), with $p < 0.001$.

The number of questions per topic cluster was: (S1) 271 out of scope questions; (S2) 978 questions about characters of the book; (S3) 101 greetings; and (S4) 165 reactions to the dialogue failure (that is, to the *More coffee?* utterance). The coders also found 27 utterances where the text was gibberish or isolated numbers and discarded them. We also note that 896 of such questions were from females while 646 were from males. Unlike the field study, we did not gather an age variable from the tablet. Understanding age effects on is thus left as future work.

Statistical Analysis

After the clustering into the four main topic clusters was performed, we then proceeded to determine whether there were audience effects on the user interactions. To do so, we employed *random effects logistic models* [13] where the response variable was the type of topic cluster (out of scope, about characters, greetings, or reactions to failure). The explanatory variables were: direct address, gender, and audience condition (A, B, and D). Recall that a direct address is defined as a message from the chatbot to the user in the form of a vocative, containing the user's name, which was used in some utterances from the chatbots. We employed logistic regression models via a *Bayesian MCMC sampler* using *Bambi* [2]. Every variable was coded as a categorical. Given the imbalance in topic clusters S1-S4 for the interactions, on every model we added an intercept to capture the settings where the effect is simply due to the number of interactions in the cluster. Notice that by predicting the topic cluster of

Table 4: Average value (μ) and the lower (↓95%) and upper HPD (↑95%) values for significant explanatory variables in each topic cluster.

	μ	↓95%	↑95%
(S1) Out of Scope			
Intercept	-2.11	-3.04	-1.30
Male Gender & Aud. D	-2.06	-3.98	-0.03
(S2) About Characters			
Intercept	1.66	0.80	2.46
(S3) Greetings			
Intercept	-3.32	-4.56	-2.01
(S4) Reaction to Failure			
Intercept	-4.48	-6.26	-2.78
Audience B	1.74	0.18	3.60
Direct Address	1.88	0.38	3.61
Direct Address & Aud. B	-1.74	-3.35	-0.16

the utterance, we unveil the factors that may have led users to choose a topic of interaction.

In Table 4 we show the results of the model for each topic cluster. The table displays only the statistically significant explanatory variables. To determine those variables, we looked into the highest posterior density (HPD) with a significance level of 95%[2]. Explanatory variables whose HPD contained the value of 0 were deemed as insignificant. That is, their effect cannot be determined to the either positive or negative. The table shows the average value for each variable (μ) as well as the lower (↓95%) and upper HPD (↑95%). One way to read the values from the table is to consider that the effect varies from the lower to the upper HPD, with μ being the expected value. Positive values indicate that the predictor tends to increase the change of the category, negative ones show the opposite effect.

From Table 4, we can initially see that S2 (about characters) and S3 (greetings) have the intercept as the only statistically significant explanatory variable. This indicates no variable, including audience conditions, can predict the topic cluster of those user utterances. We would expect a lower (for greetings) or higher (for about characters) number of interactions in each cluster due to the imbalances in interactions.

However, in the S1 topic cluster, corresponding to out of scope questions, we can see that male users when observed by strangers around the table, condition (D), decreased the number of out of scope questions. Since the effect is negative, a decrease of out of scope questions is expected. This maybe be related to the dominance effect [8].

Next, for S4, reaction to failure, we can initially see that most sentences are not of this topic (negative intercept), which is expected, due to the imbalances. However, we here saw clearly an audience effect on the user interaction, since there is significant increase in reaction to failures when the user is being observed by acquaintances (B). This perhaps can be explained by the honey-pot effect [4]. Also, we can observe that direct address increases the chance of a reaction to failure. This suggests that chatbot designers may choose to add a direct address in situations where the chatbot does not know the answer, as an attempt to further engage the user.

However, and more interesting, the effect changes direction when both conditions are true, that is, when users are directly addressed by failing chatbots in the presence of acquaintances they tend to react less to failure. We hypothesize that a direct address to a single audience member in such situations may constrain the shared experience. However, since there were a limited number of questions in this specific scenario, we believe those are interesting issues to be explored by future work.

6 DISCUSSION

In both Study 1 and Study 2, we found evidence of audience effects on how people interact with chatbots in public spaces. Regarding both of our research questions *RQ1: What are the effects of co-located audiences on the users' perceptions of social interaction with chatbots? RQ2: Does the presence of audiences influence the type and content of user's questions directed to chatbots?*

In Study 1, we saw that different audience conditions produced some significant effects on the users' self-reported user experience as measured by our questionnaire. When users were observed by acquaintances (B), users felt less that the characters answered their questions (Q8) than users in the other conditions, sometimes engaging in direct conversation with the people they knew through the artwork. Conversely, users were observed by strangers waiting in the queue (C) reported more often that the characters were talking to them (Q7) than users in the other conditions. A possible explanation is that those users may have tried to focus more on the conversation with the chatbots because they felt pressured by people waiting behind them.

We also saw in Study 1 that users who had previous knowledge about the story depicted in the artwork (as measured by Q1), seem to be more strongly affected by audience conditions. Such users, when observed by acquaintances (B), significantly disagreed that characters talked among themselves (Q6), showing an increased sense of belonging in the talk. Similarly, users familiar with the plot when observed by strangers in the queue (C) were significantly more likely to perceive the chatbots talking to them (Q7), to feel part of the conversation (Q5), and to believe that the characters answered their questions (Q8); and they tended to disagree that the characters were talking to each other (Q6). When users

[2]The HPD is the Bayesian analogous of the Confidence Interval.

know about the context and were observed by strangers standing around the table (D), there was a significant increase in the perception that the characters could talk about any subject (Q9).

In some way, previous content knowledge seems to boost the audience effects on the perception of social interactions, both when a familiar audience surrounded users or observed by strangers. We also saw that the audience effects on such users were more pronounced when considered in conjunction with gender and the kind of questions males and females posted to validate bots scope of knowledge.

In Study 1 we also saw evidence that audience effects are different according to the gender of the user, particularly considering the presence of strangers. Female users observed by strangers in the queue (C) felt more that the characters talked to them (Q7), while male users did not report that. When observed by strangers around the table (D), female users perceived that the chatbots did not talk about other subjects (Q9) to a greater extent and significantly more than male users. Our pieces of evidence seem to point towards that the male gender will 'behave better' when acquaintances did not observe them. In contrast, the qualitative analysis of Study 1 shows that when males ask out of scope questions, their sentences intensify the behavior of showing off with more frequency. Similar results exist in the literature [8].

Study 2 uncovered other types of audience effects, complementary to the ones detected in Study 1. In particular, user reactions to chatbot failures (the topic cluster *Reaction to failure*) were more common when users were observed by acquaintances (B). We also saw those male participants had a higher tendency to ask sentences in the scope when they were in an audience of strangers around the table (D).

Study 2 also showed enhanced audience effects when there was a direct address in the chatbot response. In general, we found that direct address increased the likelihood of the user reacting to a chatbot failure. However, a shared experience with acquaintances combined with direct address tended to reduce the reaction to failure.

In summary, the two studies seem to have gathered enough evidence to support positive answers to both our research questions *RQ1* and *RQ2*, and point towards the need of considering audience effects when designing conversational experiences in physical spaces. As shown in the findings, those audience effects can be modulated by many factors, including the audience being composed of acquaintances or strangers; the existence of a waiting queue; the knowledge of the context of the interaction; gender of the users; and use of direct address.

7 RECOMMENDATIONS TO DESIGNERS

In this paper, we discussed the effects of co-located audiences on chatbot interactions. The findings, discussed in the previous section, suggest several design recommendations:

DR1: Designers should consider the user's previous knowledge of content as it tends to affect the social interaction with machines, in particular when users have audiences.

DR2: Designers should consider that the presence of strangers in a queue waiting to interact with a physical conversational system, may affect how users will experience the system.

DR3: Designers should consider gender effects when crafting public interactions with conversational systems, including how to handle answers to out of scope questions.

DR4: Designers should consider tailoring and using *direct address* in some cases of chatbot utterances according to the presence of an audience. Respecting the user expectation of being considered to the bot. In general, chatbots should use the direct address, such as vocatives or pronouns, to acknowledge either all the participants in the audience or should not use them.

8 FUTURE WORK

The results and findings of our studies not only answered some basic questions about audience effects in conversational systems but also uncovered new and exciting issues. In particular, we believe it is necessary to investigate some of the gender effects which arose in both of our studies, and also whether other social demographical factors impact public human experiences with chatbots. Also, investigating how our findings transfer to other, non-physical settings such as online chatbots seems to be a promising line of future work; and to more task-oriented contexts such as in retail or hospitality.

Moreover, researchers might also want to use and improve our mixed-methods approach to understand the implications of social interaction with chatbots and to analyze the conversation spectrum. In Study 1, the use of Thematic networks helped unveil the main themes emerged from the user experience and in conjunction with statistical analysis identify the visitors' perception of social interaction with chatbots. In Study 2, the clustering approach with human coders added value and precision to classify the user question topics and the statistical log-analysis to determine the audience effects on visitors' interactions. We hope other designers and researchers also find our methodological approach useful to apply in similar projects.

9 LIMITATIONS OF THE WORK

We performed this study in Brazil, and thus our findings may not transfer to other cultures. However, it is essential to

take into account that this initial study looked into variables which exist in any culture such as gender and social issues in shared experiences. Another limitation is that our study was conducted in an art exhibition context, and therefore the findings may be different in task-oriented shared experience scenarios such as hospitality and retail.

ACKNOWLEDGEMENTS

We thank the participants from our fieldwork study for their support in providing the information required to perform our study. We also thank the anonymous reviewers and David R. Millen for the insightful comments and reviews that led to this paper. Flavio Figueiredo is sponsored by personal grants from Brazil's National Council for Scientific and Technological Development (CNPq).

REFERENCES

[1] Jennifer Attride-Stirling. 2001. Thematic networks: an analytic tool for qualitative research. *Qualitative Research* 1, 3 (2001), 385–405. https://doi.org/10.1177/146879410100100307 arXiv:http://qrj.sagepub.com/content/1/3/385.full.pdf+html

[2] Bambi 2019. BAyesian Model-Building Interface (BAMBI) in Python. Retrieved Jan 04, 2019 from https://github.com/bambinos/bambi

[3] Jim Blascovich. 2002. The Social Life of Avatars. Springer-Verlag, Berlin, Heidelberg, Chapter Social Influence Within Immersive Virtual Environments, 127–145. http://dl.acm.org/citation.cfm?id=505799.505807

[4] Harry Brignull and Yvonne Rogers. 2003. Enticing People to Interact with Large Public Displays in Public Spaces. In *INTERACT*.

[5] Jennifer L. Butler and Roy F. Baumeister. 1998. The trouble with friendly faces: skilled performance with a supportive audience. *Journal of personality and social psychology* 75 5 (1998), 1213–30.

[6] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First Impressions in Human–Agent Virtual Encounters. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 24 (Aug. 2016), 40 pages. https://doi.org/10.1145/2940325

[7] Alan D. Chatham and Florian Mueller. 2013. Adding an interactive display to a public basketball hoop can motivate players and foster community. In *UbiComp*.

[8] N. B. Cottrell, D. L. Wack, G. J. Sekerak, and R. H Rittle. 1968. Social facilitation of dominant responses by the presence of an audience and the mere presence of others. *Journal of Personality and Socia Psychology* 9, 3 (1968), 245–250.

[9] Travis Cox, Marcus Carter, and Eduardo Velloso. 2016. Public DisPLAY: Social Games on Interactive Public Screens. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction (OzCHI '16)*. ACM, New York, NY, USA, 371–380. https://doi.org/10.1145/3010915.3010917

[10] Naomi Ellemers, Cathy Dyck, Steve Hinkle, and Annelieke Jacobs. 2000. Intergroup Differentiation in Social Context: Identity Needs versus Audience Constraints. *Social Psychology Quarterly* 63 (04 2000), 60–74. https://doi.org/10.2307/2695881

[11] Katharina Emmerich and Maic Masuch. 2018. Watch Me Play: Does Social Facilitation Apply to Digital Games?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 100, 12 pages. https://doi.org/10.1145/3173574.3173674

[12] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical methods for rates and proportions; 3rd ed.* Wiley, Hoboken, NJ. https://cds.cern.ch/record/1254063

[13] Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1 ed.). Cambridge University Press. http://www.amazon.com/Analysis-Regression-Multilevel-Hierarchical-Models/dp/052168689X/ref=sr_1_1?s=books&ie=UTF8&qid=1313405184&sr=1-1

[14] David C. Giles. 2002. Parasocial Interaction: A Review of the Literature and a Model for Future Research. *Media Psychology* 4, 3 (aug 2002), 279–305. https://doi.org/10.1207/s1532785xmep0403_04

[15] Shang Guo, Jonathan Lenchner, Jonathan Connell, Mishal Dholakia, and Hidemasa Muta. 2017. Conversational Bootstrapping and Other Tricks of a Concierge Robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. ACM, New York, NY, USA, 73–81. https://doi.org/10.1145/2909824.3020232

[16] Dennis L. Kappen, Pejman Mirza-Babaei, Jens Johannsmeier, Daniel Buckstein, James Robb, and Lennart E. Nacke. 2014. Engaged by Boos and Cheers: The Effect of Co-located Game Audiences on Social Player Experience. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play (CHI PLAY '14)*. ACM, New York, NY, USA, 151–160. https://doi.org/10.1145/2658537.2658687

[17] Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1411–1420. https://doi.org/10.1145/2806416.2806475

[18] E. S. Knowles. 1983. Social physics and the effects of others: Tests of the effects of audience size and distance on social judgments and behavior. *J Pers Soc Psychol* 45, 6 (1983), 1263–1279.

[19] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 329–343.

[20] Nicole Krämer, Gary Bente, and Jens Piesk. 2003. The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behaviour in a TV/VCR application1. *IMC Workshop 2003, Assistance, Mobility, Applications* (01 2003).

[21] Celine Latulipe, Erin A. Carroll, and Danielle Lottridge. 2011. Love, Hate, Arousal and Engagement: Exploring Audience Responses to Performing Arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1845–1854. https://doi.org/10.1145/1978942.1979210

[22] S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (March 1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[23] Steve Love and Mark Perry. 2004. Dealing with Mobile Conversations in Public Places: Some Implications for the Design of Socially Intrusive Technologies. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. ACM, New York, NY, USA, 1195–1198. https://doi.org/10.1145/985921.986022

[24] Josh Lowensohn. 2015. Elon Musk: cars you can drive will eventually be outlawed. https://www.theverge.com/transportation/2015/3/17/8232187/elon-musk-human-drivers-are-dangerous

[25] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[26] Ulrike Luxburg. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing* 17, 4 (Dec. 2007), 395–416. https://doi.org/10.1007/

s11222-007-9033-z

[27] Bernhard Maurer, Ilhan Aslan, Martin Wuchse, Katja Neureiter, and Manfred Tscheligi. 2015. Gaze-Based Onlooker Integration: Exploring the In-Between of Active Player and Passive Spectator in Co-Located Gaming. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '15)*. ACM, New York, NY, USA, 163–173. https://doi.org/10.1145/2793107.2793126

[28] J.W. Michaels, J.M. Blommel, R.M. Brocato, R.A. Linkous, and J.S Rowe. 1982. Social facilitation and inhibition in a natural setting. *Replications in social psychology* 2 (1982), 21–24.

[29] Robert J. Moore, Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski. 2017. Conversational UX Design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 492–497. https://doi.org/10.1145/3027063.3027077

[30] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robot. Automat. Mag.* 19 (2012), 98–100.

[31] Heather L. O'Brien and Elaine G. Toms. 2008. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *J. Am. Soc. Inf. Sci. Technol.* 59, 6 (April 2008), 938–955. https://doi.org/10.1002/asi.v59:6

[32] S. Parise, S. Kiesler, L. Sproull, and K. Waters. 1999. Cooperating with life-like interface agents. *Computers in Human Behavior* 15, 2 (1999), 123–142.

[33] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

[34] Martin Porcheron, Joel E. Fischer, Moira McGregor, Barry Brown, Ewa Luger, Heloisa Candello, and Kenton O'Hara. 2017. Talking with Conversational Agents in Collaborative Action. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, New York, NY, USA, 431–436. https://doi.org/10.1145/3022198.3022666

[35] Stuart Reeves. 2011. *Designing Interfaces in Public Settings: Understanding the Role of the Spectator in Human-Computer Interaction* (1st ed.). Springer Publishing Company, Incorporated.

[36] Stuart Reeves, Steve Benford, Claire O'Malley, and Mike Fraser. 2005. Designing the spectator experience. In *CHI*.

[37] Stuart Reeves, Martin Porcheron, Joel E. Fischer, Heloisa Candello, Donald McMillan, Moira McGregor, Robert J. Moore, Rein Sikveland, Alex S. Taylor, Julia Velkovska, and Moustafa Zouinar. 2018. Voice-based Conversational UX Studies and Design. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article W38, 8 pages. https://doi.org/10.1145/3170427.3170619

[38] Raoul Rickenberg and Byron Reeves. 2000. The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/332040.332406

[39] Paul W. Schermerhorn, Matthias Scheutz, and Charles R. Crowell. 2008. Robot social presence and gender: Do females view robots differently than males? *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2008), 263–270.

[40] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[41] John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. John Wiley and Sons Ltd.

[42] spaCy 2019. spaCy. Retrieved Jan 04, 2019 from http://spacy.io

[43] Lee Sproull, Mani Subramani, Sara Kiesler, Janet H. Walker, and Keith Waters. 1996. When the Interface is a Face. *Hum.-Comput. Interact.* 11, 2 (June 1996), 97–124. https://doi.org/10.1207/s15327051hci1102_1

[44] Megan Strait, Lara Vujovic, Victoria Floerke, Matthias Scheutz, and Heather L. Urry. 2015. Too Much Humanness for Human-Robot Interaction: Exposure to Highly Humanlike Robots Elicits Aversive Responding in Observers. In *CHI*.

[45] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. In *Handbook of qualitative research*. 273–285.

[46] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *NIPS*.

[47] Jane Webster and Hayes Ho. 1997. Audience Engagement in Multimedia Presentations. *SIGMIS Database* 28, 2 (April 1997), 63–77. https://doi.org/10.1145/264701.264706

[48] Tom White and David Small. 1998. An Interactive Poetic Garden. In *CHI 98 Conference Summary on Human Factors in Computing Systems (CHI '98)*. ACM, New York, NY, USA, 335–336. https://doi.org/10.1145/286498.286804

[49] Laura K. Wolf, Narges Bazargani, Emma J. Kilford, Iroise Dumontheil, and S J Blakemore. 2015. The audience effect in adolescence depends on who's looking over your shoulder. In *Journal of adolescence*.

[50] Sarah Woods, Kerstin Dautenhahn, and Christina Kaouri. 2005. Is someone watching me? - consideration of social facilitation effects in human-robot interaction experiments. In *CIRA*. IEEE, 53–60.