Effects of Prior Experience, Gender, and Age on Trust in a Banking Chatbot with(out) Breakdown and Repair

Effie Lai-Chong Law[1], Nena van As[2] and Asbjørn Følstad[3]

[1] Durham University, Department of Computer Science, DH1 3LE Durham, UK
lai-chong.law@durham.ac.uk
[2] Boost.ai, Grenseveien 21, 4313 Sandnes, Norway
nena.van.as@boost.ai
[3] SINTEF, Forskningsveien 1, 0373 Oslo, Norway
Asbjorn.Folstad@sintef.no

Abstract. Trust is an attitudinal construct that can be sensitive to prior experience, gender, and age. In our study, we explored how trust in a banking chatbot might be shaped by these user characteristics. Statistical analysis of 251 participants, who interacted with one of six chatbots defined by humanlikeness (high/low) and conversational performance (no breakdown, breakdown with repaired, breakdown without repair), showed that the user characteristics of gender and age did not significantly impact trust, but prior experience did. Trust resilience was found across the gender and age groups. The effect of users' prior experience on their trust in a chatbot which they have never used holds implications for research and practice. Future studies on the effect of cultural context, longer interaction episodes, and more diverse application contexts on trust in chatbots are recommended.

Keywords: Chatbot, Artificial Intelligence (AI), Trust, Age, Gender, Prior experience, Breakdown, Repair

1 Introduction

Chatbots, text-based conversational agents powered by artificial intelligence (AI), are gaining inroads in an ever-expanding scope of sectors. People from all walks of life with different demographic backgrounds interact with chatbots, albeit to different extents, for banking, shopping, healthcare consultancy, and other online services [12]. Despite the increasing sophistication of the technologies underpinning the design and development of chatbots, including natural language processing, machine learning algorithms, human-robot interaction, and speech emotion recognition [30], communication breakdowns with chatbots still happen frequently [3]. Some attempts to repair breakdowns succeed, for instance, by asking users to rephrase requests so that their intents can better be identified, but some fail. Such failure to repair leads to frustration and confusion in users, whose trust in the chatbot of interest can be so severely undermined that they reject the chatbot altogether.

Two significant factors influencing trust in chatbots are *humanlikeness* and *conversational performance*. While several empirical studies have recently been conducted to investigate how trust could vary with these two factors (Section 2.2 and Section 2.3), other non-technological factors have captured less research effort. Impressions formed in previous interactions with fellow humans, products, and services, be they technology-based or not, can shape people's attitudes and behaviours in subsequent encounters with entities having some similar traits. This phenomenon, from the psychological research perspective, is generally referred to as *cognitive bias* [13]. Specifically, positive and negative transfer of opinions and perceptions built upon experiences in previous events to a current one can be known as *halo effect* and *horn effect*, respectively [34, 39]. In the field of Human-Computer Interaction (HCI), the halo effect of beauty to usability in different products was systematically studied and confirmed (e.g. [20]). However, to the best of our knowledge, little research on the halo (or horn) effect of trust across computing products/services has been conducted.

Apart from prior interaction experience, demographic variables, especially *gender* and *age*, can play a significant role in influencing the level of trust in people as well as technology. For instance, based on some neuropsychological and behavioural data, it was found that male trusted interaction objects (human or nonhuman entity) more than female who were more risk aversive, as observed in the context of trust-sensitive games (e.g., [6, 45]). This corroborates the arguments pertinent to gender difference in predisposition to trust [47]. Specifically, based on their analysis of the neuroimaging data on eleven heterosexual dyads playing a multi-round binary trust game, Wu and colleagues [45] found that men trusted their partner more than women, that the payoff level moderated the effect of gender on trust, and that women were more sensitive to social risk while trusting. Furthermore, in understanding the motivation underlying behaviours in an investment game exhibited by the two genders, Buchan and colleagues [6] found that men trusted interacting entities more than women; men than women emphasized more the relationship between expected return and trusting behaviour; women felt more obligated both to trust and reciprocate.

In addition, Haselhuhn and colleagues [19] had intriguing findings on gender difference in *trust dynamics*. The authors reported that following a trust violation, women were both less likely to lose trust and more likely to restore trust in a transgressor than men. Toader and colleagues [44] examined the impact of chatbot error on trust with gender as a moderating factor, which was manipulated in terms of avatar's gender but not user's gender. They found that the chatbot with a female avatar was much more forgiven when committing errors compared to one with a male avatar. In contrast, two other studies did *not* find any gender differences in trust in functional chatbots, one for online shopping [23] and the other for student support [33].

The effect of users' age on their attitudes towards chatbots has been examined in a small number of studies. Terblanche and Kidd [43], based on the adapted Technology Acceptance Model (TAM) questionnaire, found that age did *not* play a significant role in determining the level of perceived risk for deploying non-directive reflective

coaching chatbot. They further reported that older adults' intention to use the chatbot was influenced by the effort expected to invest in using it whereas younger adults valued more the usefulness and level of enjoyment of the chatbot. Goot and Pilgrim [15], based on the intriguing socioemotional selectivity theory, conducted interviews with older adults and younger ones on attitudes towards customer service chatbots. They found that the motivation for the chatbot use was contrasting. While older adults would appreciate chatbots with "human touch", their younger counterparts intended to use chatbots that enable them to avoid human contact.

Based on the literature reviewed, we were motivated to explore the following research question as part of a larger empirical study investigating the issue of trust in customer service chatbots [26]:

> *What is the respective effect of (a) prior experience, (b) gender, (c) age on the perceived trust and interaction qualities of the chatbots characterised by humanlikeness and conversational performance?*

2 Related Work and Hypotheses

In this section, we first present an overview on the work related to the design and development of chatbots, especially on the two attributes – *Humanlikeness* and *Conversational Performance*. Note that the effects of these two attributes on the fluctuation of trust levels are published in a conference paper [26]. Nonetheless, it is necessary to present the relevant descriptions in this paper to contextualise the analyses to be reported subsequently. It is also important to point out that the data and results included here are *not* covered in [26] where the analysis results on demographic variables are *not* reported to keep it more focused. Towards the end of this section, we delineate the three main hypotheses of our study.

2.1 Trust in Chatbots

Trust is typically understood as the willingness of a trustor to "accept vulnerability based on positive expectations of the intentions or behaviour of the other" [37]. Several models of trust in technology exist (e.g. [8, 18, 25, 28, 29]). They typically consider trust as determined by a set of underlying factors representing beliefs about the trustworthiness of the trustee. In a review of trust-building factors in embodied conversational agents, [36] identified social intelligence, communication style, performance and humanlikeness as among the factors impacting agent trustworthiness. [22] also found that chatbot humanlikeness leads to increased trust and adoption, contributing to customer loyalty. Research on cognitive agents and social robots has studied how humanlikeness may lead to 'trust resilience', that is, upkeep of trust in spite of undesirable system outcomes [9]. Similarly, [17] found that humanlike design cues conveyed by social robots can strengthen user trust and positively impact user preference regardless of operation failure. Nonetheless, findings on the relative effects of humanlikeness and conversational performance on trust in chatbots remain inconsistent (e.g., [31, 46]).

2.2 Humanlikeness of Chatbots

Many AI-powered systems are designed to mimic human behaviour, verbal as well as non-verbal. The extent to which a chatbot is perceived to be humanlike shapes the user experience [21], intention to use [42], and goal attainment that the chatbot is aimed to enable [4]. The phenomena of the Turing test [27] and uncanny valley effect [7] are associated with the humanlikeness of such AI-based conversational interactions. In fact, Rapp and colleagues, in their review of chatbot research [35], found that more than 25% of the studies addressed the topic of humanness. Furthermore, several design features of chatbots have been found influencing the perceived level of humanlikeness (i.e., anthropomorphism), including conversational style [21], visual representation and initial self-presentation [2, 14], informal language [2], and features hinting at chatbot intelligence such as backchanneling [14] and conversational relevance [40].

2.3 Conversational Performance of Chatbots

In the context of customer service, we define '*conversational performance*' as the chatbot's ability to provide relevant and helpful responses to users' requests. This interpretation is supported by certain industry reports. Accordingly, efficient and effective access to help can motivate users to engage with chatbots [10] whereas getting stuck in a conversation without progress or receiving irrelevant responses can undermine the chatbot use [12]. Despite the advances of machine learning methods, especially large language models deployed in GPT-3 and BERT, human-chatbot interactions involve breakdowns [12], which often occur even in human-human interactions [38]. Conversational breakdown in chatbots may happen when the chatbot fails to predict any user intent for the user request. It typically triggers a fallback response as a common attempt to conversational repair where the chatbot states that it has not understood and asks the user to rephrase [11, 16, 31].

2.4 Chatbots for Customer Service

One of the rapidly growing application areas of chatbots is customer service [35]. The banking chatbot we created for our empirical study (Section 3) is a typical example. Basically, customer service chatbots are deployed to respond to frequently asked questions (FAQs) posed by customers [41] and integrated into customer websites as alternative text-based information source [1]. User interactions with customer service chatbots are generally short. Technically, chatbots can be rule-based or AI-based. The former relies on pre-defined decision trees whereas the latter utilises statistical data-driven methods to infer user intents based on prediction models. Specifically, a user enters a request in a chatbot in free text from which an intent is predicted [21]. The chatbot responds according to the intent inferred by conveying to the user one or more messages that may meet the request. The user may refine the chatbot's response through selecting one of answer options, presented as buttons or menu items. The

content and prediction models of customer service chatbots can be complex, especially when the scope of user intents is diverse [48].

2.5 Research Hypotheses

In this subsection, first we reiterate the three key insights discussed in Introduction, corresponding to the three parts of the main research question:

(a) The halo and horn effects have not been applied to analyse the phenomenon of people's trust in chatbots.
(b) Results of some studies of trust in interpersonal relationships and technologies, including chatbots, suggest that some gender-specific patterns could be observed. In general, male tend to trust interacting objects, be they animate or inanimate, more than their female.
(c) There seem age-dependent factors influencing people's trust in chatbots with older adults relying more on the perceived humanness of chatbots.

However, as the number of the related studies is limited, the observed patterns and factors remain inconclusive and more empirical research is required.

We integrate the insights to formulate the following *null* hypotheses (H), indicating the non-conclusive directions as derived from our analysis of the related work.

Hypothesis 1: There are no significant differences between users with different prior chatbot experience in their overall trust in the banking chatbots characterised by specific levels of humanlikeness and conversational performance.

Hypothesis 2: There are no significant differences between male and female in their overall trust in the banking chatbots characterised by specific levels of humanlikeness and conversational performance.

Hypothesis 3: There are no significant differences between younger and older users in their overall trust in the banking chatbots characterised by specific levels of humanlikeness and conversational performance.

3 Methods

Our empirical study employed a 2x3 factorial design with Humanlikeness (yes / no) and Conversational Performance (no breakdown, breakdown with repair, breakdown without repair) as IVs. This resulted in six groups of participants of which two did not experience breakdown (Table 1). We go into details on the operationalization of each IV level below.

Table 1: Six variants of chatbots tested with six groups of participants

	No Breakdown	Breakdown with Repair	Breakdown without Repair
Humanlikeness: Yes	Group 1	Group 2	Group 3
Humanlikeness: No	Group 4	Group 5	Group 6

3.1 Instrument – Customer Service Chatbot Variants

For our study, we developed a customer service chatbot representing a fictitious bank called "Boost Bank", using a dedicated platform for virtual agents [26] where user messages are processed by an AI-powered intent prediction model. The chatbot was modified into six variants characterised by the combination of two attributes. Each version of the chatbot deployed an equal number of open-ended as well as button-based answer options for the participants.

Conversational performance was operationalized in terms of the presence (or absence) of *breakdown* and *repair* for one of the three tasks as shown in Fig.1. Breakdown and repair followed the 'repeat' pattern of [3] where breakdown involved the chatbot failing to understand the user request and asking the user to reformulate, and repair involved the chatbot understanding the users' reformulated request and providing a relevant response. Each version was evaluated by different groups of participants.

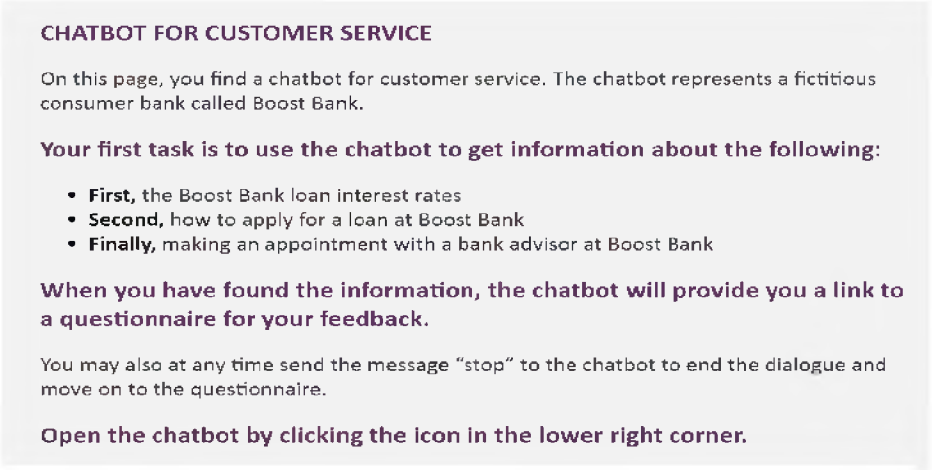

Fig 1. The instruction page of the chatbot

Humanlikeness was operationalized in terms of cues in chatbot appearance and conversational style (Fig. 2). Specifically, the humanlike chatbot, in contrast to the non-humanlike chatbot, had a humanlike avatar image [14], presented itself with a human name [2, 14], and an informal conversational style [2], including greetings and pleasantries. as well as first and second person pronouns.

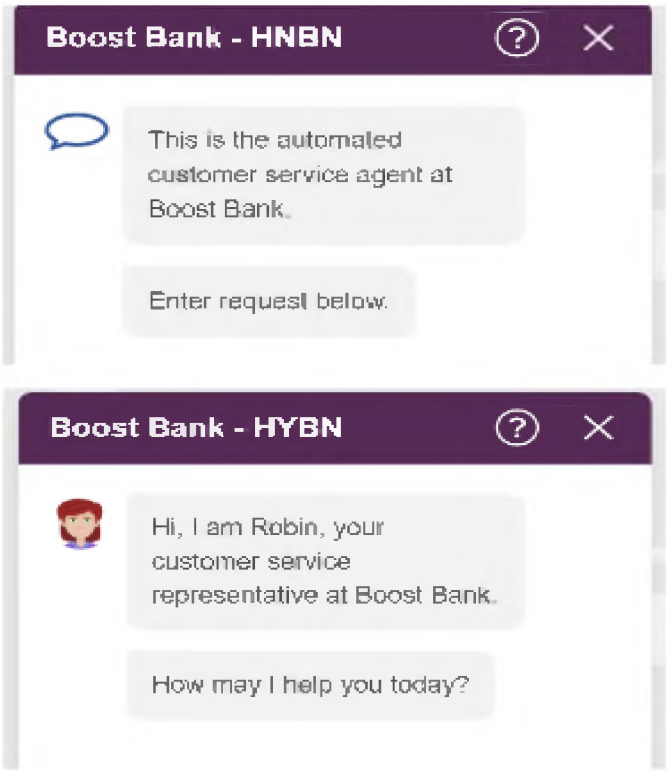

Fig. 2. Chatbot humanlikeness implementation with different greeting styles.
Upper (abstract icon, impersonal style; Humanlikeness - No);
Lower (avatar with name, personal conversational style; Humanlikeness - Yes).
HYBN = Humanlikeness No Breakdown No; HYBN = Humanlikeness Yes Breakdown No

3.2 Measurement: Post-intervention Questionnaire

After completing the tasks with the chatbot, participants were asked to complete a questionnaire with four parts: Part 1 on measuring trust in the chatbots, overall as well as task-specific; Part 2 on qualitative feedback on trust in the chatbots; Part 3 on reliability, anthropomorphism, and social presence; Part 4 on demographics and prior chatbot experience. Each item, where applicable, is measured with a 7-point Likert scale with 1 (Strongly disagree) and 7 (Strongly agreed). Table 2 shows the items that are relevant to this paper.

Table 2: Post-intervention questionnaire items

Variable	Items	Source
Overall Trust (OT)	OT1: When in need of customer service, I feel I can depend on the chatbot OT2: I can always rely on the chatbot to provide good customer service OT3: I feel I can count on the chatbot for my customer service needs	[25]
Task-specific Trust TT1, TT2, TT3	Considering the chatbot's answer on [Task 1/2/3], I feel I can depend on it. I can rely on the support provided by the chatbot on [Task 1/2/3]. I feel I can count on the chatbot for questions on [Task 1/2/3]	Home-grown
Prior Chatbot Experience	***Prior Use Preference (PF)*** PF1: I frequently use chatbots for customer service PF2: I use chatbots for customer service when this is provided as a service alternative PF3: I have used chatbots for customer service for a long time ***Prior Chatbot Satisfaction* (SAT)** SAT1: Chatbots for customer service typically provide good help SAT2: In general, chatbots for customer service are an efficient way to get support SAT3: I usually find chatbots for customer service pleasant to use. ***Prior Use Frequency (FQ)*** Five options: • More than 10 times • 5-10 times • 3-4 times • 1-2 times • Never	Home-grown
Demographic	Gender (female, male, prefer not to say) Age (free text) Country of residence (free text) Education (three options)	Home-grown

3.3 Participants

Altogether 251 participants were recruited via the crowdsourcing platform Prolific. Among them, 178 were female, 69 male and 4 preferred not to say. For country of residence, the distribution was: 128 UK, 106 US, 5 Canada, 5 Ireland, 4 South Africa, and 1 from Australia, Hungary, and Mexico each. Most of the participants (n=226) had higher education level and the rest had high school level. The average age was 35.7 years old (SD=12.1, range: 18-68). The majority (n=112) of participants were under 30 years old (Table 3).

Table 3: Distribution of participant ages

Age Range	18-20	21-30	31-40	41-50	51-60	61-68
Frequency	17	95	72	34	23	10

Each participant was randomly assigned to one of the six groups and given a unique code to log into the website where they carried out the tasks with the chatbot (Fig. 1). On the cover page, participants were informed about the study's tasks, that data collection was fully anonymous, that data would be used for research purposes, and that they would agree to participate and enter the study by clicking the 'next' button. On average, they spent 5.8 mins (SD =4.0, range: 2.8-23.9) in completing the three tasks.

4 Results

In this section, we present our empirical findings in the order of the three hypotheses (Section 2.5), which correspond to the three parts of the main research question: effects of *Prior Experience* first, then those of *Gender*, and end with effects of *Age*.

4.1 Effects of Humanlikeness and Conversational Performance on Trust: A Synopsis

As mentioned earlier, results on the effects of the two factors – humanlikeness and conversational performance – on trust are published elsewhere [26]. Nonetheless, when presenting and discussing the effects of prior experience, gender, and age on trust, it is relevant to contextualise them with reference to these factors.

Results of between-group analysis showed that for the task with seeded breakdowns there were significant differences in trust across the six groups with the lowest ratings for the two groups experiencing breakdowns without repair, and that humanlikeness did not impact the extent to which the trust level changed. Results of within-group analysis showed significant differences in trust across the three tasks (Fig. 1). These observations challenge the effect of humanlikeness on trust while supporting the notion of trust resilience as the participants did not spill the impaired trust over the subsequent task (for details see [26]).

4.2 Effect of Prior Experience on Trust (Hypothesis 1)

The participants' prior experience with chatbots was measured through three variables: *Prior Use Preference*, *Prior Use Frequency*, and *Prior Chatbot Satisfaction* (Table 2). The variables were measured with 7-point Likert scales. To investigate the effect of these variables on the participants' trust, it was beneficial to conceptualise these as different grouping variables rather than scales. We applied the same analysis approach for the effect of *Gender* and *Age* (Section 4.3 and Section 4.4). To this end, we regrouped participants into three ranges for each of these variables: Low, Middle, and High. The ranges were based on the 33rd and 66th percentiles of the ratings (see Tables 4-6 for details).

Specifically, 3*2*3 ANOVAs (*[Prior variables]*Humanlikeness*Conversational Performance*) were performed, where *[Prior variables]* include *Prior Use Preference*, *Prior Use Frequency* or *Prior Chatbot Satisfaction* (Table 2). The DV was Overall Trust.

Results showed that *Prior Use Preference* significantly impacted the participants' Overall Trust ($F_{(2,233)} = 21.920$, $p < .001$, $\eta^2 = .158$). However, no significant interaction effects were observed. Means and standard deviations for Overall Trust across the three ranges of participants' ratings for *Prior Use Preference* are shown in Table 4.

Table 4: Mean (SD) of Overall Trust across the three rating ranges of *Prior Use Preference*

Group	Range	n	Overall Trust
Low	1.00-3.67	91	3.71 (.16)
Middle	3.68-5.33	74	4.08 (.17)
High	5.34-7.00	86	5.13 (.14)

Furthermore, results showed that *Prior Use Frequency* did not have any significant impact on the participants' Overall Trust ($F_{(2,233)} = 1.917$, $p = .149$, $\eta^2 = .016$). No significant interaction effects were observed here either. Means and standard deviations for the Overall Trust across the three ranges of participant's ratings for *Prior Use Frequency* are presented in Table 5.

Table 5: Mean (SD) of Overall Trust across the three rating ranges of *Prior Use Frequency*

Group	Range	n	Overall Trust
Low	<5 times	64	3.95 (0.19)
Middle	5-10 times	90	4.44 (0.16)
High	>10 times	97	4.44 (0.16)

Finally, results showed that *Prior Chatbot Satisfaction* significantly impacted the participants' levels of Overall Trust ($F_{(2,233)} = 65.456$, $p < .001$, $\eta^2 = .360$). No significant interaction effects were observed. Means and standard deviations for Overall Trust across the three ranges of participants' ratings for *Prior Chatbot Satisfaction* are presented in Table 6.

Table 6: Mean (SD) of Overall Trust across the three rating ranges of *Prior Chatbot Satisfaction*

Group	Range	n	Overall Trust
Low	1.00 -3.67	91	3.28 (0.14)
Middle	3.68 -5.33	74	4.32 (0.13)
High	5.34 -7,00	86	5.41 (0.14)

To further investigate why out of the three measures on prior experience only *Prior Use Frequency* did not have a significant impact on Overall Trust, bivariate Spearman correlations among the three components, factored by gender and age, were computed. Some intriguing findings were obtained.

Significant *positive* correlations between *Prior Use Preference* and *Prior Chatbot Satisfaction* were found, irrespective of gender or age groups (Table 7). In other words, the results suggested that participants who tended to choose to use customer service chatbots when available, were satisfied with the experience. At the same time, a significant *negative* correlation was found between *Prior Chatbot Satisfaction* and *Prior Use Frequency*, i.e., the more participants used such chatbots the less satisfied they became. Interestingly, for male participants this correlation was not significant; nor was it significant for the younger or middle age group (Table 7).

Table 7: Bivariate correlations among the three components of prior experience: *Prior Chatbot Satisfaction* (Satisfaction), *Prior Use Preference* (Preference) and *Prior Use Frequency* (Frequency) by *Gender* and *Age* groups.

	All	**Female**	**Male**	**Younger**	**Middle**	**Older**
	N = 251	N=178	N=69	N=112	N=72	N=67
Satisfaction vs. Preference	0.64 $p<.001$	0.640 $p<.001$	0.624 $p<.001$	0.677 $p<.001$	0.607 $p<.001$	0.611 $p<.001$
Satisfaction vs. Frequency	-0.195 $p=.002$	-0.230 $p=.002$	-0.083 $p=.496$	-0.168 $p=.076$	-0.210 $p=.076$	-0.251 $p=.041$

4.3 Effect of Gender on Perceived Trust (Hypothesis 2)

To analyse the main effect of *Gender* (female, male - participants who reported "prefer not to say" were excluded for this analysis) and its interaction effects with *Humanlikeness* (no, yes) and *Conversational Performance* (no breakdown, breakdown with repair, breakdown without repair) of the chatbots, a 2*2*3 ANOVA was performed with Overall Trust as DV.

Results showed that the main effects of the three IVs were not significant for Overall Trust. The interaction effects were also non-significant.

Concerning the notion of gender-related "trust dynamics" or "trust resilience" (Section 1 and 2), we examined how the level of trust varied with the tasks and gender. When breakdown occurred, the impact on the task-specific trust (i.e., TT2; Trust in Task 2) was obvious (Table 8). Interestingly, while there were obvious drops in TT2, the level of trust bounced back for TT3 for both genders, albeit to a slightly larger extent for male. We performed 2*2*3 ANOVAs on TT1-TT2 (i.e., trust difference between Task 1 and Task 2) and TT2-TT3 (i.e. trust difference between Task 2 and Task 3). The main effect of *Conversational Performance* was significant, but non-significant for *Humanlikeness* or *Gender*. None of the interaction effects were significant. This suggested both female and male demonstrated trust resilience.

Table 8; Mean Task-specific Trust (TT) per task for two genders under different conditions

	Female					
	Humanlike			**Non-humanlike**		
Conversational Performance	**TT1**	**TT2**	**TT3**	**TT1**	**TT2**	**TT3**
No Breakdown	6.01	***5.77***	5.74	5.15	***4.91***	5.35
Breakdown with Repair	5.27	***4.76***	5.31	5.37	***4.66***	5.32
Breakdown without Repair	6.08	***1.45***	5.22	5.06	***1.1***	4.94
	Male					
	Humanlike			**Non-humanlike**		
Conversational Performance	**TT1**	**TT2**	**TT3**	**TT1**	**TT2**	**TT3**
No Breakdown	5.86	***5.05***	5	5.4	***4.53***	5.2
Breakdown with Repair	5.69	***5.21***	5.59	5.21	***4.31***	0.31
Breakdown without Repair	5.5	***1.38***	5.07	5.52	***2.09***	5.45

4.4 Effect of Age on Perceived Trust (Hypothesis 3)

As indicated in Table 3, the distribution of ages was skewed towards the younger ones. To address this issue, we regrouped participants into three age brackets: Younger (18-30 years old, n = 112), Middle (31-40 years old, n =72), Older (41-68 years old, n = 67). Similar to the analysis on the effect of *Gender* (Section 4.3), a 3*2*3 ANOVA (*Age*Humanlikeness *Conversational Performance*) was performed with Overall Trust as a DV.

Results showed that *Age* did not play any significant role in influencing Overall Trust ($F_{(2,231)}$ =.759, p =.469). None of the interaction effects among the three IVs were significant. Regardless of age brackets, participants had lowest trust when they experienced breakdowns in both human-like and non-humanlike conditions.

Table 9 illustrates the observation that the three age groups gave similar ratings for Overall Trust with the means leaning towards neutrality (i.e., 4 out of 7). We also applied the same analysis of trust dynamics to the three age groups (cf. Section 4.3 for *Gender*). Table 10 displays the descriptive statistics. Results of 3*3*2 ANOVAs showed that the only significant main effect was *Conversational Performance*.

Table 9: Mean (SD) of the four variables across three age groups

Group	Range (years)	N	Overall Trust
Younger	18 - 30	112	4.21 (1.61)
Middle	31 - 40	72	4.39 (1.50)
Older	41 - 68	67	4.41 (1.55)

Table 10: Mean Task-specific Trust (TT) per task for three age groups under different conditions of the chatbot *Conversational Performance*

Younger						
	Humanlike			**Non-humanlike**		
Conversational Performance	**TT1**	**TT2**	**TT3**	**TT1**	**TT2**	**TT3**
No Breakdown	5.87	***5.56***	5.69	4.88	***4.59***	4.90
Breakdown with Repair	5.15	***4.75***	5.32	5.20	***3.8***	4.98
Breakdown without Repair	5.96	***1.23***	5.32	4.85	***1.24***	4.85
Middle						
	Humanlike			**Non-humanlike**		
Conversational Performance	**TT1**	**TT2**	**TT3**	**TT1**	**TT2**	**TT3**
No Breakdown	6.03	***5.93***	5.73	5.34	***4.79***	5.49
Breakdown with Repair	5.89	***4.83***	5.39	5.54	***5.28***	5.72
Breakdown without Repair	6.00	***1.17***	5.64	5.25	***1.52***	5.00
Older						
	Humanlike			**Non-humanlike**		
Conversational Performance	**TT1**	**TT2**	**TT3**	**TT1**	**TT2**	**TT3**
No Breakdown	6.13	***5.47***	5.33	5.83	***5.13***	5.5
Breakdown with Repair	5.48	***5.10***	5.48	5.24	***4.97***	5.42
Breakdown without Repair	5.42	***1.97***	4.57	5.49	***1.23***	5.44

However, the three-way interaction effects (*Age*Conversational Performance* Humanlikeness*) for both TT1-TT2 ($F_{(4,233)} = 3.57$, $p = 0.008$) and TT2-TT3 ($F_{(4,233)} = 2.49$, $p = .044$) trust differences were significant (Fig. 3a; Fig 3b). These suggested that the three age groups changed the level of trust from task to task significantly under different chatbot conditions. For instance, for the Middle age group, the TT2-TT3 value of 4.47 for the group 'breakdown without repair and humanlike' was higher than the corresponding values of 4.09 and 2.6 for the Younger and Older age groups.

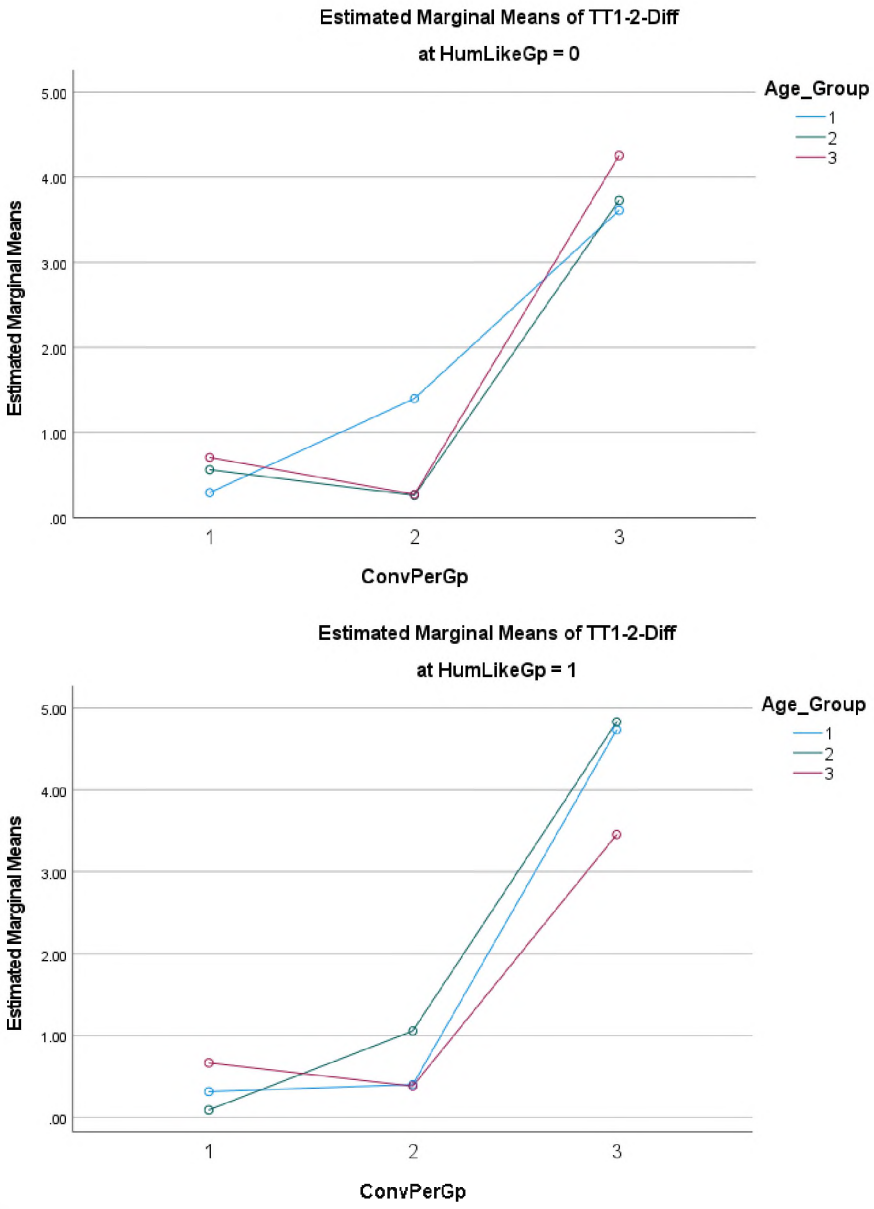

Fig. 3 (a): Significant three-way interaction effects. The upper graph shows the trust difference between Task 1 and Task 2 (TT1-2-Diff) under the condition of *Humanlikeness* = No. The lower graph shows the trust difference between Task 1 and Task 2 (TT1-2 Diff) under the condition of *Humanlikeness* = Yes. *ConvPerGp* = Conversational Performance (1 = No breakdown; 2 = Breakdown without repair; 3 = Breakdown with repair)

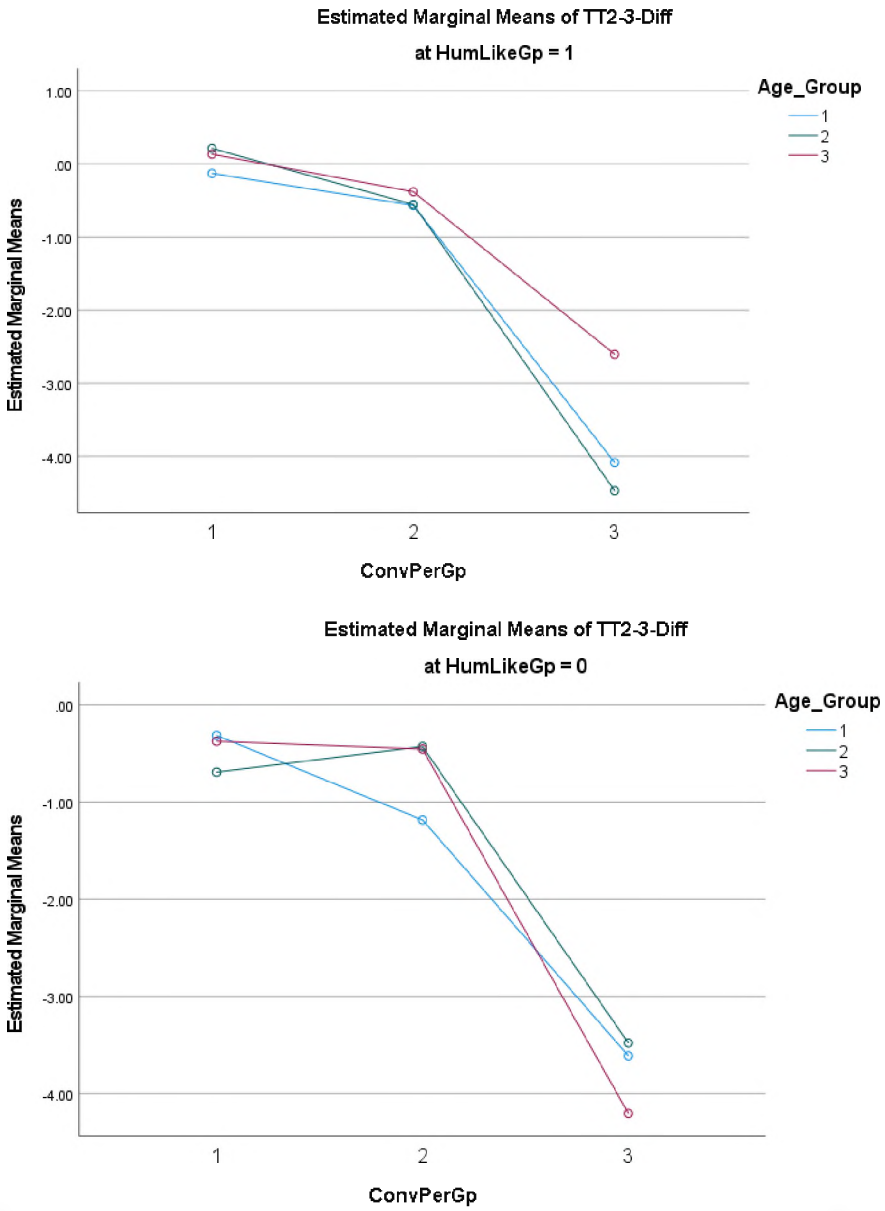

Fig. 3 (b): Significant three-way interaction effects. The upper graph shows the trust difference between Task 2 and Task 3 (TT2-3-Diff) under the condition of *Humanlikeness* = No. The lower graph shows the trust difference between Task 2 and Task 3 (TT2-3-Diff) under the condition of *Humanlikeness* = Yes. *ConvPerGp* = Conversational Performance (1 = No breakdown; 2 = Breakdown without repair; 3 = Breakdown with repair)

5 Discussion

Based on the analysis results, we can address the main research question of this work (Section 1). Of the three components of prior chatbot experience, both *Prior Use Preference* and *Prior Chatbot Satisfaction* had a significant effect on Overall Trust whereas *Prior Use Frequency* had no significant effect. Hence, we can only partially accept the null Hypothesis 1. *Gender* and *Age* did not have any significant effect on Overall Trust. Hence, we accept the two null Hypothesis 2 and Hypothesis 3.

5.1 Hypothesis 1: Prior experience

Concerning the effect of prior experience on trust in chatbots for customer service, we found that both *Prior Use Preference* and *Prior Chatbot Satisfaction* significantly impacted trust. There are some interesting things to note about these variables, however. The measures were taken after the participants had interacted with the chatbots of this study, in a post-intervention questionnaire. Note that it was a procedural arrangement rather than any intentional experimental manipulation. Nevertheless, it was plausible that the positive or negative user experience the participants had in this study influenced their recall of satisfaction with some other chatbots - mixing prior experiment experience with the actual experiment experience.

Putting this in perspective, we can say that for chatbot users a halo or horn effect on trust (of positive / negative transfer) (Section 1) from previous chatbot experiences is detectable. This finding is of high interest to research as well as service providers, as it suggests the importance of being aware of the experience a participant or user brings with them into a research setting - or a real-world usage situation. If users with more positive prior experiences with chatbots are prone to have higher levels of trust in a chatbot with which they have not previously interacted, it is important to researchers to check for this user characteristic. Likewise, for service providers, it will be important to understand the experience users bring with them into their chatbot interactions: it allows explaining why some chatbots face a more sceptical use base than others and may help designing prompts or marketing strategies to address these earlier experiences.

The implication is to identify means to assess prior experience from chatbot users and how to use such data once gathered. A design suggestion is to present a welcome message of the chatbot, asking a simple question: "Is this your first time talking with a chatbot?" which, depending on the response, can lead to a different sort of conversation, where there would be a set of predefined options to allow users to express their like or dislike in chatbots in general and train AI to recognise that. That way the bot could again tailor its responses to the user.

5.2 Hypothesis 2: Gender

Results of our study confirmed the findings of existing studies, albeit small in number, that no significant gender difference in trust in chatbots (e.g. [23, 44]) could be found. Nevertheless, more empirical research needs to be conducted to substantiate

this observation. On the contrary, our findings could not confirm the previous work on gender difference in trust in the context of human-human interaction such as that female are more trust resilient than male (e.g. [6, 45, 47]; Section 1). This observation may challenge the prevailing assumption that models on human trust in AI-powered systems can (or even should) be grounded in their counterparts on interpersonal trust (e.g. [18]). But anthropomorphising AI systems like chatbots does not necessarily imply that users interact in the same manner as they typically do with fellow humans. One implication is that we should adopt inductive approaches, integrating as well as extrapolating what is empirically observed to inform the development of an alternative model of trust in human-AI interaction.

5.3 Hypothesis 3: Age

Concerning the effect of *Age* on the level of trust in chatbots, our findings confirmed the work of [43] that age did not play a significant role in the form of any main effects. Nevertheless, the coaching chatbot examined in [43] did not impart any knowledge to users but rather gave them space to reflect through conversational stimulation whereas the chatbot used in this study was directive by conveying specific information requested to users. Clearly, more research on different types of chatbots is needed, especially given the observed interaction effects due to age differences. Furthermore, the analyses of [18, 43] on the different motivations underpinning younger and older adults for their acceptance and intention to use chatbots are intriguing. While older adults may appreciate more emotional than practical value from chatbot interactions, which may be appreciated more by their young counterparts [18], it is critical that trustworthy AI-powered chatbots can convey a strong sense of fairness, respect, and transparency to users, irrespective of their ages or gender.

5.4 Limitations

There are some limitations of our work. Trust is a culture-sensitive construct. People's propensity to (dis)trust objects, human or non-human objects, can be shaped by the sociocultural environment where they grow up. While potentially interesting to explore the effect of culture or social environment on trust, our data collection did not consider including this as demographic variables, given the concern that it would be difficult to get a balanced distribution of relevant subgroups with the sample size of 200-300. This can be addressed in our future work. Another limitation is that the three tasks could be completed in a relatively short period of time, which is rather common in chatbot interaction (e.g., a comparable duration in [42]). Nonetheless, the effect of gender, age and prior use experience on trust might be more detectable with longer interaction episodes. In the same vein, the application context and associated tasks, which are online banking services, can have a strong impact on trust. In our future work, we aim to explore other contexts such as healthcare and education.

6 Conclusion

AI-powered systems like chatbots are increasingly prevalent in many sectors. It is critical to ensure the trustworthiness of the systems, which should be developed with effective algorithms and human-centred design approaches. Prospective users of a trustworthy system can only benefit if they accept and adopt the system. Hence, it is deemed important to examine systematically factors influencing trust in AI-powered systems. Trust is an attitudinal construct that can be sensitive to demographic variables such as gender, age, and prior experience interacting with similar entities, human as well as non-human. Based on the results of our empirical study on the effect of different characteristics of customer service chatbots, these demographic variables did not play any significant role in influencing trust in the chatbots. This observation lent further evidence to the conclusion of some existing work while defying the others. Overall, the landscape of trust in AI is evolving as well as diversifying. To state the obvious: more research needs to be conducted to gain insights into the design of AI-powered systems that improve the quality of huma lives in a fair and safe manner.

References

[1] Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-based Chatbots in Customer Service and their Effects on User Compliance. Electronic Markets 31, 2, 427–445.

[2] Theo Araujo. 2018. Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in Human Behavior, 85, 183–189.

[3] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.

[4] Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. 2021. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. Journal of the Academy of Marketing Science 49, 4, 632–658.

[5] Matthew Brzowski and Dan Nathan-Roberts. 2019. Trust measurement in human–automation interaction: A systematic review. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 63. SAGE: Los Angeles, CA, 1595–1599.

[6] Nancy R. Buchan, Rachel TA Croson, and Sara Solnick, 2008. Trust and gender: An examination of behavior and beliefs in the Investment Game. Journal of Economic Behavior & Organization, 68, no. 3-4, 466-476.

[7] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. In the shades of the uncanny valley: An experimental study of human– chatbot interaction. Future Generation Computer Systems, 92, 539–548.

[8] Cynthia L Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. International Journal Of Human-Computer Studies 58, 6, 737–758.

[9] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22, 3, 331.

[10] Drift. 2018. The 2018 State of Chatbots Report. Technical Report. https://www.drift.com/blog/chatbots-report/

[11] Asbjørn Følstad and Cameron Taylor. 2019. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In International Workshop on Chatbot Research and Design. Springer, 201–214.

[12] Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. Quality and User Experience, 6, 1, 1–17.

[13] Joseph P. Forgas, J. and Laham, S. M. 2017. Halo effects. In R. F. Pohl (Ed.), Cognitive illusions: Intriguing phenomena in thinking, judgment and memory (pp. 276–290). Routledge/Taylor & Francis Group.

[14] Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Computers in Human Behavior, 97, 304–316.

[15] Margot J Goot and Tyler Pilgrim, 2019. Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. In International Workshop on Chatbot Research and Design, pp. 173-186. Springer, Cham.

[16] Erika Hall. 2018. Conversational design. A Book Apart New York.

[17] Adriana Hamacher, Nadia Bianchi-Berthouze, Anthony G Pipe, and Kerstin Eder. 2016. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In Proceedings of 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). 493–500.

[18] Peter A Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2021. Evolving trust in robots: specification through sequential and comparative meta-analyses. Human factors 63, 7, 1196–1229.

[19] Michael P. Haselhuhn, Jessica A. Kennedy, Laura J. Kray, Alex B. Van Zant, and Maurice E. Schweitzer. 2015. Gender differences in trust dynamics: Women trust more than men following a trust violation. Journal of Experimental Social Psychology, 56, 104-109.

[20] Marc Hassenzahl, 2004. The interplay of beauty, goodness, and usability in interactive products. Human–Computer Interaction, 19, no. 4, 319-349.

[21] Isabel Kathleen Fornell Haugeland, Asbjørn Følstad, Cameron Taylor, and Cato Alexander Bjørkli. 2022. Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. International Journal of Human-Computer Studies, 161, 102788.

[22] Liss Jenneboer, Carolina Herrando, and Efthymios Constantinides. 2022. The Impact of Chatbots on Customer Loyalty: A Systematic Literature Review. Journal of theoretical and applied electronic commerce research 17, 1, 212–229.

[23] Dharun Lingam Kasilingam, 2020. Understanding the attitude and intention to use smartphone chatbots for shopping. Technology in Society, 62, 101280.

[24] Guy Laban and Theo Araujo. 2019. Working together with conversational agents: the relationship of perceived cooperation with service performance evaluations. In International Workshop on Chatbot Research and Design. Springer, 215–228.

[25] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. Journal of the Association for Information Systems 16, 10, 1.

[26] Effie L-C Law, Asbjørn Følstad, and Nena van As. 2022. Effects of humanlikeness and conversational breakdown on trust in chatbots for customer service. In Proceedings of Nordic Human-Computer Interaction Conference (NordiCHI '22). Aarhus, Denmark, ACM.

[27] Catherine L Lortie and Matthieu J Guitton. 2011. Judgment of the humanness of an interlocutor is in the eye of the beholder. PLoS One 6, 9, e25085.
[28] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. Academy of management review 20, 3, 709–734.
[29] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. ACM Transactions on management information systems (TMIS) 2, 2, 1–25.
[30] Michael McTear. 2020. Conversational ai: Dialogue systems, conversational agents, and chatbots. Synthesis Lectures on Human Language Technologies 13, 3, 1–251.
[31] Chelsea M. Myers, Luis Fernando Laris Pardo, Ana Acosta-Ruiz, Alessandro Canossa, and Jichen Zhu. 2021. "Try, Try, Try Again:" Sequence Analysis of User Interaction Data with a Voice User Interface. In Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21). ACM, New York, NY, USA, Article 18, 1–8.
[32] Cecilie Bertinussen Nordheim, Asbjørn Følstad, and Cato Alexander Bjørkli. 2019. An initial model of trust in chatbots for customer service—findings from a questionnaire study. Interacting with Computers 31, 3, 317–335.
[33] Joonas A Pesonen, 2021. "'Are You OK?' Students' Trust in a Chatbot Providing Support Opportunities. In International Conference on Human-Computer Interaction, pp. 199-215. Springer, Cham.
[34] Mary Katherine Radeke and Anthony John Stahelski, 2020. Altering age and gender stereotypes by creating the Halo and Horns Effects with facial expressions. Humanities and Social Sciences Communications 7, no. 1, 1-11.
[35] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human- chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. International Journal of Human-Computer Studies, 151, 102630.
[36] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic re- view: Trust-building factors and implications for conversational agent design. International Journal of Human–Computer Interaction 37, 1, 81–96.
[37] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. Academy of management review 23, 3, 393–404.
[38] Emanuel A Schegloff. 1991. Conversation analysis and socially shared cognition. In Socially Shared Cognition, Lauren B Resnick, John M Levine, and Stephanie D Teasley (Eds.). American Psychological Association, Washington, DC, US, 150– 171.
[39] Marie-Sophie Schönitz. 2019. The horn effect in relationship marketing: a systematic literature review. In Proceedings of the 48th European Marketing Academy, 8378
[40] Ryan M Schuetzler, Justin Scott Giboney, G Mark Grimes, and Jay F Nunamaker Jr. 2018. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. Decision Support Systems, 114, 94–102.
[41] Amir Shevat. 2017. Designing bots: Creating conversational experiences. O'Reilly Media, Inc., Boston, US.
[42] Mark P Taylor, Kees Jacobs, KVJ Subrahmanyam, et al. 2019. Smart talk: How organizations and consumers are embracing voice and chat assistants. Technical Report. Capgemini SE.
[43] Nicky Terblanche and Martin Kidd, 2022. Adoption Factors and Moderating Effects of Age and Gender That Influence the Intention to Use a Non-Directive Reflective Coaching Chatbot. SAGE Open, 12, no. 2, 21582440221096136.
[44] Diana-Cezara Toader, Grațiela Boca, Rita Toader, Mara Măcelaru, Cezar Toader, Diana Ighian, and Adrian T. Rădulescu, 2019. The effect of social presence and chatbot errors on trust. Sustainability, 12, no. 1, 256.

[45] Yan Wu, Alisha SM Hall, Sebastian Siehl, Jordan Grafman, and Frank Krueger, 2020. Neural signatures of gender differences in interpersonal trust. Frontiers in Human Neuroscience, 14, 225.

[46] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. ACM Transactions on Internet Technology (TOIT) 17, 1, 1–20.

[47] Rachid. Zeffane, 2020. Gender, individualism–collectivism and individuals' propensity to trust: A comparative exploratory study. Journal of Management & Organization 26, no. 4, 445-459.

[48] Juliana JY Zhang, Asbjørn Følstad, and Cato A Bjørkli. 2021. Organizational factors affecting successful implementation of chatbots for customer service. Journal of Internet Commerce, 1–35.