Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service

EFFIE LAI-CHONG LAW, Department of Computer Science, Durham University, UK
ASBJØRN FØLSTAD, SINTEF, Norway
NENA VAN AS, boost.ai, Norway

Trust in chatbots can be shaped by various factors such as humanlikeness in terms of visual appearance and conversational content, and conversational performance in terms of the chatbot's ability to avoid conversational breakdown. The literature is inconclusive concerning the effect of humanlikeness and conversational performance on trust, especially their interaction effect. To examine the relations among these variables, we conducted a 2x3 (humanlikeness x conversational performance) factorial experiment with 251 participants, who were asked to perform three tasks with a chatbot for an online bank under one of the six conditions. Participants completed a questionnaire measuring trust and commented on trust 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 repairs 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 tasks but non-significant inter-task correlations on trust for the two groups. 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. Thematic analysis showed that inter-group contrasts could be found for the theme 'underlying functionality' and 'affective responses.' Implications for research, practice and future work were drawn.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: chatbot, trust, human-likeness, conversational performance, breakdown, repair

ACM Reference Format:
Effie Lai-Chong Law, Asbjørn Følstad, and Nena van As. 2022. Effects of Humanlikeness and Conversational Breakdown on Trust in Chatbots for Customer Service. In *Nordic Human-Computer Interaction Conference (NordiCHI '22), October 8–12, 2022, Aarhus, Denmark*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3546155.3546665

1 INTRODUCTION

Chatbots are increasingly offered as a user interface to customer service. Chatbots are software agents that interact with users through natural language [16] and may, hence, be seen as a supplement to chat-based service offered by human support personnel. In a Gartner industry report [35], one third of surveyed companies either had chatbots implemented for customer service already or were planning for this within a year. Furthermore, in early adopting sectors such as financial services, nearly half of the leading companies have been found to take up chatbots for customer service [53]. Industry reports from the US market suggest that more than one third of customers have engaged with chatbots for customer service within the retail domain and one fifth within utilities [52], and the chatbot market is predicted to continue to grow exponentially [43].

This is the accepted version (postprint) of a paper presented at NordiCHI '22. The version of record can be found at https://doi.org/10.1145/3546155.3546665.

In spite of the recent uptake and growth predictions, chatbots are still a relatively novel technology in the context of customer service. While a substantial proportion of potential chatbot users have already had first-hand experience with chatbots for customer support purposes, just as many may still have their first chatbot interaction ahead of them. For the expected future growth of chatbots to be realised, it is important that users in need of customer service can regard chatbots as a trusted means for providing information and support required [40]. Hence, for chatbot developers and providers, it is critical to understand the factors that determine user trust.

Trust in information technology has emerged as an area of much research interest [37]. Trust in chatbots entails particularly interesting research challenges as such trust may be impacted by not only system-oriented aspects like functionality and reliability, but also the humanlike character of the chatbot [41]. Humanlikeness has been found to be conducive to trust [45]. In one study, variation in chatbot features to enhance perceptions of anthropomorphism has been found to be as important to trust as chatbot performance [55]. Furthermore, studies of social robots and cognitive agents suggest that humanlikeness may be conducive to trust resilience, that is, the ability to maintain user trust in the face of performance issues [23].

At the same time, studies also suggest that though humanlikeness indeed may have some impact on user experience, its relevance for trust in current chatbots for customer service may be of far less relevance than, for example, the chatbot's expertise and responsiveness [40]. Such suggestions about a lack of actual impact of humanlikeness on trust in current chatbots for customer service are substantiated by the relative prevalence of conversational breakdowns in such chatbots [1]. Industry reports have found chatbots' inconsistency in interpreting and responding appropriately to user requests to be a main detractor for future use [27].

In this context, practitioners and researchers alike need insight into the relative importance of factors impacting users' trust in chatbots. In particular, there is a lack of understanding about the impact of chatbot appearance, in terms of humanlikeness, relative to that of conversational performance, in terms of the ability to avoid or mitigate conversational breakdowns on trust. Several open research questions in the literature along this line are:

- Are chatbots' humanlike features as important for trust as their ability to reliably provide support? Or is trust dominated by the chatbot's conversational performance rather than their humanlikeness?
- Does chatbot humanlikeness strengthen trust resilience in the face of performance issues?

These questions motivated us to conduct a 2x3 factorial experiment to investigate the effects of humanlikeness and conversational performance on trust in chatbots for customer service. Altogether 251 participants were recruited for an online study where they were asked to interact with one of six chatbot conditions before reporting on their experience in a questionnaire. The chatbot conditions varied on humanlikeness (yes/no) and conversational performance (no breakdown / breakdown with repair / breakdown without repair). The dependent variables, gauged by the questionnaire, included measures of trust and anthropomorphism. Furthermore, participants were asked to give free-text comments on factors influencing their trust in chatbots for customer service.

The study makes an important contribution to the body of knowledge on factors impacting trust in chatbots for customer service. Specifically, our study provides insights into the effects of humanlikeness and conversational performance as well as the interaction between these two factors, complementing and challenging existing studies on trust resilience. Furthermore, we developed a comprehensive coding scheme to categorize the participants' qualitative feedback on their chatbot interaction experiences. The scheme can serve as a useful tool for future chatbot research.

The remainder of the paper is structured as follows: First, we present relevant background concerning chatbots for customer service, humanlikeness, conversational performance and trust. We then detail our research hypotheses before presenting our findings and discussing them. Finally, we discuss implications for theory and practice and point out relevant future research.

2 BACKGROUND

2.1 Chatbots for Customer Service

Customer service is an important application domain for chatbots [35], In fact, it is a domain that has seen sustained growth since the onset of the most recent wave of interest in chatbot research [11, 53]. Chatbots for customer service are typically set up to answer frequently asked questions by customers [51] and may therefore serve as a first line of support for companies [21]. In a recent literature review on text-based chatbots, [42] found customer service to be the most frequently represented application domain in chatbot research, followed by health, e-commerce, and education.

Chatbots for customer service typically resemble text-based chat applications and are made available to users as part of customer websites with social media or messaging platforms [1]. A range of platforms is available for chatbots for customer service [44]. The chatbots are typically set up utilising what [38] refers to as statistical data-driven approaches where user intents are inferred on the basis of underlying AI-based prediction models. Users are typically encouraged to enter their requests in free text from which the specific user intent is predicted [25]. Based on the predicted user intent, the chatbot provides the corresponding action, typically sending the user one or more messages conveying relevant content to their request. Users may then refine the chatbot responses through selecting between predefined answer alternatives, presented as buttons or quick replies .

The content and prediction models of chatbots for customer service can be relatively complex. For example, [28] reported on a chatbot in telecom domain including more than 2700 user intents. A public sector chatbot included about 6000 user intents [17]. Correctly predicting user intents among this diversity is challenging and requires continuous and resource demanding maintenance of the underlying prediction model [56]. At the same time, user interactions with chatbots for customer service chatbots may be relatively brief.

2.2 Chatbot Humanlikeness

Interactions with chatbots closely resemble interactions with human beings through chat applications. In this paper, we use the term 'humanlikeness' to refer to characteristics in a chatbot and its interactions that resemble those of a human conversation partner. Furthermore, we refer to users' perceptions of such humanlikeness as 'anthropomorphism'. That is, humanlikeness may be objectively characterised whereas anthropomorphism is subjectively perceived in the eye of the beholder. In the literature at large, the terms humanlikeness and anthropomorphism are at times used interchangeably, but we prefer to make an explicit distinction between them.

The humanlikeness of conversational interaction with computers has been found alluring to users, theorists, and practitioners since the early days of computing [54]. The fascination for chatbot humanlikeness is illustrated in the aim for chatbots to pass the Turing test [34], as well as the concern for chatbot humanlikeness to trigger what in social robotics is referred to as the uncanny valley effect [8]. In their review of chatbot research, [42] found that more than a quarter of the studies had the theme of humanness.

In a review of research on social robots and chatbots, [4] found that design features impact perceptions of anthropomorphism, which in turn impact intention to use. Furthermore, they identified that the impact of anthropomorphism depends on robot-related mediators (e.g. intelligence), functional mediators (e.g. usefulness), and relational mediators (e.g. trust). Chatbot design features found to impact humanlikeness and perceptions of anthropomorphism include conversational style [25], visual representation and initial self-presentation [2, 20], informal language [2], and features hinting at chatbot intelligence such as backchanneling [20] and conversational relevance [50]. In turn, chatbot human-likeness influences user perception and behaviour on a range of dimensions, including hedonic user experience [25], brand perception [2], user sentiment [10], user compliance [1], transaction conversion [48], and intention to use [33].

2.3 Chatbot Conversational Performance

The conversational performance of chatbots for customer service is highly important to chatbot success. By 'conversational performance' we mean the chatbot's ability to provide relevant and helpful responses to users' requests. In a questionnaire survey study with chatbot users, productivity was found to the most frequently reported motivation for chatbot use [5]. Furthermore, industry reports have noted the importance of efficient and effective access to help for users' motivation to engage with chatbots [14], and experiences of being stuck in a conversation without progress or receiving irrelevant responses have been reported as key factors discouraging for chatbot use [27].

While most conversations with well-crafted chatbots for customer service are likely to entail relevant and helpful chatbots responses, such conversations may also involve breakdown. Across a small series of case reports, [19] found chatbots for customer service to provide false positive responses to 14-28 % of user messages and to provide no help in 16-25% of chatbot conversations, depending on available features for breakdown mitigation. Due to the complexity in language interaction, troubles in communication and interpretation are prevalent – even among human conversationalists [49]. Conversational breakdown in chatbots may happen when the chatbot fails to predict any user intent for the user request or when the chatbot erroneously predicts a user intent not reflecting the actual content of the user request. The former breakdown typically triggers a fallback response where the chatbot states that it has not understood and asks the user to rephrase. The latter is considered a false positive response and, while noticeable by the user, is not discoverable as such by the system unless the user reacts negatively [18].

As conversation is prone to breakdown, conversational repair is a critical capability of any conversationalist – human or machine [22].[39] described conversational repair as a key conversational UX (user experience) pattern, and different approaches to conversational repair have been suggested [3]. Arguably, the most prevalent approach to conversational repair still is the fallback intent where a chatbot expresses failure to understand and invites the users to ask again. In this study, we address conversational performance as a chatbot's ability to avoid conversational breakdown and, if breakdown occurs, to provide efficient mitigating action.

2.4 Trust in Computers and Chatbots

Given the importance of trust for the acceptance and uptake of new technology, trust in computers has evolved as an area of substantial interest [37]. 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" [47]. Several models of trust in technology exist (e.g. [9, 30, 37]). Grounded in a broadly acknowledged model of trust in organisations [36], these models typically consider trust as determined by a set of underlying factors representing beliefs about the trustworthiness of the trustee. For example, [37] consider trust to depend on user perceptions of system functionality, helpfulness, and reliability. That is, trust in technology is considered grounded, at least partly, in a cognitive assessment of the technology and its applicability for the task at hand.

Due to the relative novelty of chatbots for many user groups, and the need to understand factors impacting user uptake of this technology, trust in chatbots has emerged as an area of substantial research interest [31]. Consequently, the body of knowledge in this area is growing rapidly and recently several literature reviews have addressed trust in chatbots, conversational agents, and/or social robots as one of their review topics [4, 24, 26, 45]. In their review of trust-building factors in conversational agents, [45] identified social intelligence, communication style, performance and humanlikeness as among the factors impacting agent trustworthiness. For the chatbot context, the study by [45] is however limited as most of the reviewed studies concern interaction with embodied agents. In their review of research on chatbots and customer loyalty [26] found that chatbot humanlikeness leads to increased trust and adoption, which contributes to customer loyalty.

As the saying goes, trust is hard earned and easily lost. Resonating this, researchers have investigated factors which might sustain user trust in the face of unwanted system outcomes. In particular, research on cognitive agents and social robots has studied how humanlikeness may lead to so-called 'trust resilience', that is, upkeep of trust in spite of undesirable system outcomes [12]. In a series of experiments comparing user trust in traditional computer systems, humanlike cognitive agents, and human agents, [12] found evidence for such trust resilience in that humanlikeness was found to provide a buffer against reduced trust due to automation failure. Similarly, in a study of social robots, [23] found that humanlike design cues can strengthen user trust and positively impact user preference in spite of operation failure. While the current literature suggests trust in chatbots to be impacted by both conversational performance and humanlikeness, the relative impact of the two, and possible interaction effects, warrant further research, especially when the findings of the related studies diverge with regard to the relative effects that these chatbot characteristics may have (e.g., [40, 55]).

3 METHODOLOGY

3.1 Instruments

3.1.1 Chatbot design. A bespoke chatbot for customer service representing a fictitious customer bank called Boost Bank was developed for our empirical studies. The chatbot was developed using a dedicated platform for virtual agents (https://www.boost.ai/product/platform) where user messages are processed by an AI-based intent prediction model.

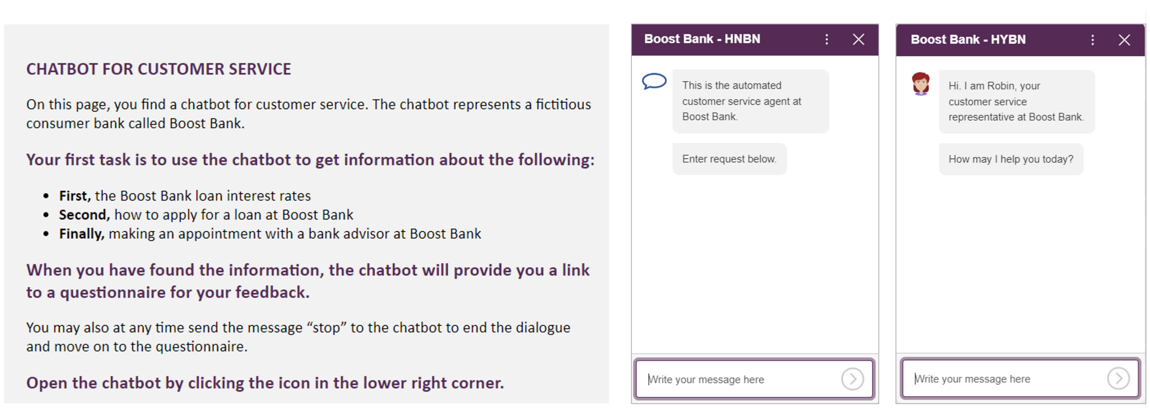


Fig. 1. The instruction page of the chatbot (left). The introduction with a non-humanlike icon (middle) and a humanlike avatar (right).

The chatbot was morphed into six variants differentiated by two factors: humanlikeness and conversational performance. Humanlikeness was operationalized in terms of icon cues in chatbot appearance and conversational style (Section 2.2). Specifically, the humanlike chatbot, in contrast to the non-humanlike chatbot, had a humanlike avatar image [20], presented itself with a human name [2, 20], and an informal conversational style [2] including greetings and pleasantries. as well as first and second person pronouns. Conversational performance was operationalized in terms of the presence (or absence) of breakdown and repair (Section 2.3) for one of three tasks. 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 entailed the chatbot's understanding of the users' reformulated request to provide a relevant response. Each variant was evaluated by different groups of participants (Section 3.3).

Figure 1 displays the general instruction page and the introductory lines that users see after they click a non-humanlike icon or a humanlike avatar. Table 1 shows the dialogues under the different conditions. Specifically, the last three rows show how the chatbot responds to a communication breakdown where it does not understand the user's request with repair or without repair.

The chatbot was designed to handle three tasks in a fixed sequence: 'Loan interest rates'; 'How to apply for a loan'; 'Bank advisor appointment'. Each task corresponded to a specific intent in the chatbot intent prediction model and associated with a set of actions in the chatbot. The specific intents and actions were based on a generic intent prediction model for customer service in consumer banking but with content and training data adapted for the purposes of this study. In designing the experiment, we injected the breakdown in the second task. Under the 'with repair' condition, after the user makes the second attempt the conversation continues onto the second task with a relevant response. Under the 'no repair' condition, the second task is left incomplete irrespective of the number of attempts. After that, the third task is launched when the user enters the related request.

3.1.2 Questionnaires. A questionnaire with items adapted from the related work on measuring user trust of different aspects and another one with homegrown items for demographic data were employed for this study. Each item, where applicable, is measured with a 7-point Likert scale with 1 (Strongly disagree) and 7 (Strongly agreed).

Trust General (TG). : It consists of three items adapted from [30]: (TG1) When in need of customer service, I feel I can depend on the chatbot; (TG2) I can always rely on the chatbot to provide good customer service; (TG3) I feel I can count on the chatbot for my customer service needs.

Task-specific Trust (TT): It consists of three items adapted from [30]: (TTi1) Considering the support I received from the chatbot on [task], I feel I can depend on it; (TTi2) I can rely on the support provided by the chatbot on [task]; (TTi3) I feel I can count on the chatbot for my questions on [task], where the [task] is substituted by the task-specific terms and i = 1, 2 or 3. These close-ended items are followed by an open-ended free-text one with the instruction "In the chatbot you just used, what is most important for you to trust its information and advice? (Please, answer in your own words, at least 15 words)"

Trusting Belief Reliability (TBR): It consists of three items: (TBR1) The chatbot is a very reliable solution; (TBR2) The chatbot will not fail its users; (TBR3) The chatbot is extremely dependable. They are adapted from [30] who identify five other constructs of trusting beliefs. But they are not included here to optimise the length of the questionnaire, mitigating the risk of fatigue effect. Reliability is deemed most relevant for our study.

Perceived Anthropomorphism (PA): Three pairs of contrasting adjectives (adapted from [2]) are presented with a 7-point semantic differential scale: Machine-like vs. Humanlike; Unnatural vs. Natural; Artificial vs. Lifelike.

Table 1. Dialogues of the chatbot under different conditions

Humanlike	Non-humanlike
Chatbot: Hi. I am Robin, your customer service representative at Boost Bank. How may I help you today? **User**: loan interest rate **Chatbot**: I am happy to help you with loan interest rates. At Boost Bank we are proud/happy to offer competitive interest rates. Which type of loan are you interested in? • Home • Vehicle • Flexible **User**: [Clicks button] **Chatbot**: Thank you. Let me give you a brief overview of our Home / Vehicle / Flexible loan interest rates: • If the loan is for 70% of the total value or less, you get our best interest rate of 3% • If the loan is for more than 70% of the total value, interest rates are slightly higher at 4%	**Chatbot**: This is the automated customer service agent at Boost Bank. Enter your request below. **User**: Loan interest rate **Chatbot**: Boost Bank offers competitive loan interest rates. For information, select one of the following. • Home • Vehicle • Flexible **User**: [Clicks button] **Chatbot**: Home / Vehicle / Flexible loan interest rates are as follows: • Loans less than or equal to 70% of total value: interest rate 3% • Loans greater than 70% of the total value: interest rate 4%
User: How do I apply for a loan? **Chatbot**: Here at Boost Bank, we make loan applications fast and easy. In the near future, you will even be able to apply for loans right here in the chat. For now, you can send in your loan application by getting in touch with one of our friendly bank advisors.	**User**: How do I apply for a loan? **Chatbot**: Loan application is fast and easy. In the near future this can be done in the chat. Currently, loan applications are to be submitted via a bank advisor.
User: Contact bank advisor **Chatbot**: I would love to help you get in touch with one of our advisors. For which area do you need advice? • Banking • Loans • Insurance **User**: [Clicks button] **Chatbot**: Thank you for making the selection. I have now registered your request for an appointment with one of our advisors. Have a great meeting!	**User**: Contact bank advisor **Chatbot**: For appointment with bank advisor, select the relevant area of advice • Banking • Loans • Insurance **User**: [Clicks button] **Chatbot**: The request for an appointment with a bank advisor is now registered.
If chatbot does not understand **Chatbot**: I am sorry that I was not able to understand your question. You may try to ask again using slightly different words. Please keep questions brief and to the point.	**If chatbot does not understand** **Chatbot**: Request not identified. Consider entering the request in other words. Requests should be brief and concise.
If chatbot understands the second time (with repair) the regular conversation on "Apply for a loan"	**If chatbot understands the second time (with repair)** the regular conversation on "Apply for a loan"
If chatbot does not understand – second time (No repair) **Chatbot**: I am sorry, but it seems that I am not able to help you with this question. Is there anything else I can help you with?	**If chatbot does not understand – second time (No repair)** **Chatbot**: Unable to respond to request. A new request may be entered.

Social Presence: It comprises four items adapted from [29]: (SP1) I felt like I was engaged in an active dialogue with the chatbot; (SP2) My interaction with the chatbot felt like a back-and forth conversation; (SP3) I felt as if the chatbot and I were involved in a mutual task; (SP4) The chatbot was efficient in responding to my activities.

Previous chatbot usage: All these items are homegrown. Previous experience with chatbots for customer service: (EXP1) I frequently use chatbots for customer service; (EXP2) I use chatbots for customer service when this is provided as a service alternative; (EXP3) I have used chatbots for customer service for a long time. General satisfaction with such chatbots: (GSAT1) Chatbots for customer service typically provide good help; (GSAT2) In general, chatbots for customer service are an efficient way to get support; (GSAT3) I usually find chatbots for customer service pleasant to use.

Demographics: Age (free text); Gender (three options); Country of residence (free text); Education (three options)

In summary, the total number of close-ended items is 32, and three open-ended (free text) questions.

3.2 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. Majority (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). 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 (Figure 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 minutes (SD= 4.0, range= 2,8-23,9) in completing the three tasks.

3.3 Research Hypothesis

The study employed a 2x3 factorial experimental design with two independent variables (IVs). IV1: Humanlikeness (two levels) and IV2: Conversational Performance (three levels) (Table 2).

Dependent variables (DVs) are self-reported measures with different instruments (Section 3.1.2). Ten null hypotheses are formulated following the conventional statistical approach. The main effect of each IV and their interaction effect (between-subject) on the trust ratings are examined in H1-H5 and on the non-trust qualities in H8-H9. H6 examines (within-subject) whether the participants in each Group rate the trust differently across the three tasks. H7 examines how the trust ratings are related. H10 examines how the past chatbot experiences account for the variance of trust in the chatbot of this study.

Table 2. Six variations of the dedicated chatbot for customer service

IV1 / IV2	No breakdown (n=77)	Breakdown with Repair (n=85)	Breakdown without Repair (n=89)
Humanlike (n=121)	Group 1 (n=39)	Group 2 (n=40)	Group 3 (n=42)
Non-humanlike (n=130)	Group 4 (n=38)	Group 5 (n=45)	Group 6 (n=47)

H1 There are no significant differences in general trust (**TG**) in the chatbot across the six conditions.
H2 There are no significant differences in the rating of trust in the chatbot on the first task of loan interest (**TT1**) across the six conditions.
H3 There are no significant differences in the rating of trust in the chatbot on the second task of applying for a loan (**TT2**) across the six conditions.
H4 There are no significant differences in the rating of trust in the chatbot on the third task of booking an advisor appointment (**TT3**) across the six conditions.
H5 There are no significant differences in the rating of trusting belief reliability (**TBR**) in the chatbot across the six chatbot conditions.
H6 There are no significant differences in trust in the chatbot across TT1, TT2 and TT3 under each of the six conditions.
H7 There are no significant correlations among the five measures of trust - TG, TT1, TT2, TT3 and TBR – for the whole sample and per group.
H8 There are no significant differences in the rating of perceived anthropomorphism (**PA**) of the chatbot across the six conditions.
H9 There are no significant differences in the rating of social presence (**SP**) of the chatbot across the six conditions.
H10 There are no significant effects of the past use experience (**EXP**) and past general satisfaction (**GSAT**) with chatbots for customer service on the general trust (**TG**) in the chatbot for this study.

4 RESULTS

In this section, first we report on the findings based on quantitative data, which are analysed with statistical methods to verify the ten hypotheses. Then we present the findings based on qualitative data, which are analysed with thematic analysis, leading to a coding scheme with nine themes/subthemes.

4.1 Quantitative Data Analysis

The instruments employed (Section 3.1.2) consist of three to four items, which are assumed to measure the same latent variable. To verify the assumptions, we performed the reliability tests for nine variables (Table 3). Results show that all Cronbach’s alphas are 0.9 or above (except EXP), and the inter-item correlations were generally high. Overall, it was justifiable to use the average scores for verifying the hypotheses of the study.

Table 3. Results of the reliability tests for nine self-reported variables

Variable	TG	TT1	TT2	TT3	TBR	PA	SP	EXP	GSAT
Nr. items	3	3	3	3	3	3	4	3	3
Cronbach’s α	0.96	0.95	0.98	0.97	0.94	0.93	0.90	0.89	0.94
Inter-item correlation	0.87-0.92	0.87-0.89	0.93-0.97	0.88-0.93	0.81-0.87	0.78-0.83	0.57-0.79	0.65-0.8	0.8-0.9

In addition, normality tests were applied to the data broken down by the six groups. Results of Shapiro-Wilk tests show mixed observations. For six variables, at least three out of the six groups were normally distributed, e.g., for SP all except Group 1 were normally distributed. In contrast, none of the six groups for TT1 and TT3 were normally distributed whereas two groups for TT2 were. Given these inconsistent patterns, we performed both parametric and non-parametric tests to ensure consistent conclusions would be drawn. In the case of convergent results, we would report the statistical results of the parametric tests only; otherwise, we would report both.

Descriptive statistics of the nine variables broken down by groups are shown in Table 4, which are also displayed graphically to illustrate the changes (Figure 2).

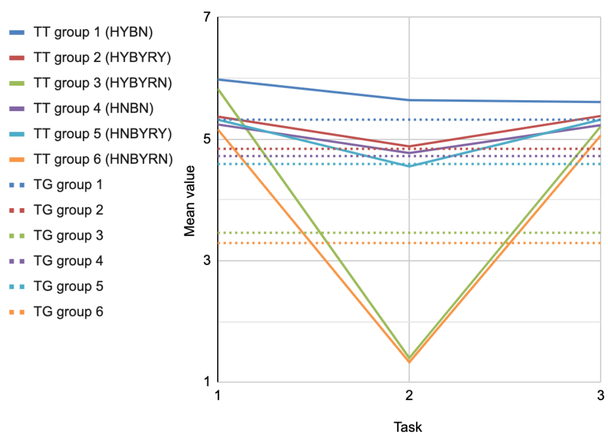
The six mnemonic group labels denote the combinations of the levels of the two IVs: Group 1, **HYBN** (**H**umanlikeness **Y**es, **B**reakdown **N**o); Group 2, **HYBYRY** (**H**umanlikeness **Y**es, **B**reakdown **Y**es, **R**epair **Y**es) and so on. The overall trust levels (N=251) as measured by TG (mean = 4.31, SD =1.56) and TBR (mean = 4.1, SD = 1.55) were modest and could even be considered as neutral.

Table 4. Means and SD (in brackets) of the nine variables for the six experimental groups

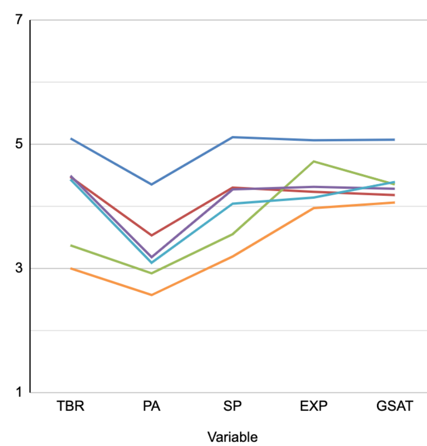
Group	TG	TT1	TT2	TT3	TBR	PA	SP	EXP	GSAT
1: HYBN	5.32 (0.23)	5.98 (0.87)	5.64 (1.30)	5.61 (1.37)	5.09 (1.41)	4.35 (1.54)	5.11 (1.74)	5.06 (1.56)	5.07 (1.49)
2: HYBYRY	4.84 (1.57)	5.37 (1.20)	4.88 (1.22)	5.38 (1.25)	4.47 (1.13)	3.53 (1.29)	4.30 (1.48)	4.23 (1.56)	4.18 (1.56)
3: HYBYRN	3.46 (0.25)	5.83 (1.29)	1.40 (1.14)	5.21 (1.82)	3.37 (1.57)	2.92 (1.30)	3.55 (1.51)	4.72 (1.44)	4.35 (1.59)
4: HNBN	4.72 (0.23)	5.24 (1.68)	4.77 (1.72)	5.23 (1.54)	4.49 (1.50)	3.18 (1.48)	4.27 (1.51)	4.31 (1.74)	4.28 (1.71)
5: HNBYRY	4.59 (0.22)	5.32 (1.29)	4.55 (1.64)	5.32 (1.30)	4.43 (1.51)	3.09 (1.24)	4.04 (1.40)	4.14 (1.77)	4.39 (1.51)
6: HNBYRN	3.29 (0.16)	5.16 (1.34)	1.33 (0.79)	5.06 (1.41)	3.00 (1.10)	2.57 (1.27)	3.19 (1.21)	3.97 (1.48)	4.06 (1.65)

In Figure 2a, the notable drops in TG (dotted lines) and, to a larger extent, in TT2 for Group 3: HYBYRN (green) and Group 6: HNBYRN (orange) were consistent with the expectation that no repair for the breakdown undermined the participants' trust in the chatbot.

Nonetheless, the trust for TT1 and TT3 seemed comparable across the six groups; there seemed no "spill over" effect from the negative experience and impaired trust by no repair. Participants summarized the overall trust (TG) after completing the three tasks with the chatbot, and they could remain task-focused when evaluating trust based on the respective experience for the individual task. A similar but less contrasting trend could be observed for TBR, which were also lowest in Group 3 and Group 6 (green and orange, respectively, Figure 2b).



(a)



(b)

Fig. 2. (a) Task-specific trust and general trust per group; (b) Measures per group: trusting belief reliability (TBR), perceived anthropomorphism (PA), social presence (SP), past chatbots usage experience (EXP) and general satisfaction (GSAT) with chatbots in the past.

As explained in Section 4.1, both parametric and non-parametric tests were performed. As the results of both types of tests converged to the same conclusion to accept/reject the null hypotheses, we do not report non-parametric statistics here. Specifically, two-way ANOVA was applied for verifying H1-H5, H8 and H9 (Table 5). Post hoc Tukey's tests were run for IV2; only those levels (Level 1: No Breakdown; Level 2: Breakdown with Repair; Level 3: Breakdown without Repair) with significant differences are reported here for brevity's sake.

Table 5. Inferential statistical results of two-way ANOVA

		TG	TT1	TT2	TT3	TBR	PA	SP
Humanlikeness	F(1,245)	**4.617**	**8.802**	2.902	1.149	3.629	**14.781**	**6.740**
	p	**.033**	**.003**	.096	.285	.058	**<.001**	**.01**
	η	**.018**	**.035**	.025	.015	.015	**.057**	**.027**
Conversational performance	F(1,245)	**36.98**	0.87	**210.91**	0.840	**31.89**	**11.87**	**17.05**
	p	**<.001**	.422	**<.001**	.433	**<.001**	**<.001**	**<.001**
	η	**.23**	.007	**.63**	.007	**.21**	**.09**	**.12**
Level 1:2	t(160)	-	-	**2.12**	-	-	-	-
	p			**.042**				
Level 2:3	t(172)	**6.85**	-	**17.81**	-	**6.250**	-	**3.787**
	p	**<.001**		**<.001**		**<.001**		**<.001**
Level 1:3	t(164)	**7.71**	-	**19.22**	-	**7.387**	**4.61**	**5.674**
	p	**<.001**		**<.001**		**<.001**	**<.001**	**<.001**
Interaction effect	F(1,245)	.437	1.740	1.904	0.247	0.859	2.156	0.870
	p	.647	.178	.151	.781	.425	.118	.42
	η	.004	.014	.015	.002	.007	.017	.007
Hypotheses		**H1**	**H2**	**H3**	H4	**H5**	**H8**	**H9**

Significant statistics and (subsequently) rejected null hypotheses are in boldface

There was no significant interaction effect for any of the variables. IV1 and IV2 had significant effects on PA and SP; they shaped the perceived humanness and engagement of the chatbot, which in turn could contribute to the overall trust. For TT1, only IV1 had a significant effect. Perhaps since this task was simple and smooth for all participants, who might then focus on the appearance and conversational style rather than performance of the chatbot. In contrast, for TT2 with breakdown and repair for some participants, the performance issue became more salient and IV2's effect became significant. For TT3, after experiencing the two previous tasks, participants might become less sensitive to the chatbot's features; none of the main effects was significant.

For H6, we applied mixed factorial ANOVA with Task being within-subject factor and ExpGroup (experimental group) being between-subject factor. Results show that the main effect of Task ($F(2,490)=240.49$, $p<.001$, $\eta=.495$) and of ExpGroup ($F(5,245)=17.31$, $p<.001$, $\eta=.26$) and the interaction effect ($F(10,490)=50.80$, $p<.001$, $\eta=.51$) were significant. The trust ratings TT2 varied more notably with groups than the other trust ratings did (Figure 2a).

For H7, bivariate correlations were computed. All five trust ratings (TG, TT1, TT2, TT3 and TBR) were highly correlated to each other for the whole sample ($N = 251$, $p<.001$) with the range of Pearson coefficient $r = 0.32$ (TT1 -TT2) to $r = 0.88$ (TG-TBR), $p<.01$. However, when breaking down in groups, for both Group 3 and Group 6, the correlations between TT1 and TT2 ($r = .21$; $r = .11$) and of between TT2 and TT3 ($r =.28$, $r =-.02$) were insignificant ($p>.05$). These findings suggest that participants of the two groups rated the trust for the three tasks independently. As the experiences with the tasks were contrasting, the participants were able to distinguish and rate them accordingly.

For H10, we applied linear regression with EXP and GSAT being predictors and TG the criterion. Results show that EXP was not a significant predictor but GSAT was ($F(1, 249) = 132.6$, $p<.001$, $R2 = 0.374$, $ß = .574$), suggesting the experiential transfer across context.

In summary, nine of the ten null hypotheses were rejected: fully (both main effects were significant for H1, H6, H8 and H9) or partially (one main effect was significant in the case of H2, H3, H5; most correlations were significant for H6; one predictor was significant for H10). H4 was the only null hypothesis that was not rejected.

4.2 Qualitative Data Analysis

All 251 participants answered the open-ended question on the factors influencing their trust in chatbots for customer service (Section 3.1.2). The length varied from 2 to 116 words (Mean = 30.94, SD = 18.55). The answers were analysed with the thematic analysis approach [6]. A hierarchical coding scheme (Appendix A) was developed by two researchers, who were experienced in qualitative data analysis and chatbot research. At the highest level, the answers are clustered into two types of quality: Informational and Interactional. The former is divided into two themes - the Process of generating responses and the Attribute of such responses as well as the Style of delivering them, whereas the latter is divided into three themes - Affective (i.e. user emotional experience), Cognitive (i.e. system performance) and Privacy. The themes Process and Style & Attribute are further divided into four and two subthemes, respectively. No subthemes are under the three Interactional themes.

Each theme or subtheme is instantiated with a set of representative concepts (NB: 'Nature of Information' and 'Cognitive' have more than one set of concepts, which are placed within a blue box). Each set is tallied per experimental group. Taking the theme 'Underlying Functionality' as an example, concepts related to 'automation algorithms' are mentioned 4, 3, 14, 1, 5, and 25 times in Group 1-6, respectively, amounting to a total frequency of 54. An intriguing pattern is observed here: Group 3 and Group 6, who experienced breakdowns without repair and seemed stimulated to comment on the mechanisms, had a higher frequency of 14 and 25 as compared to only once in Group 4 (no breakdown). A similar pattern is detected in 'Nature of Task' (2, 1, 10, 5, 6,11); Group 3 and Group 6 commented on 'scope of enquiry' more often than the other groups. These two groups also tended to express less positive affect (e.g. zero for Group 3) and more negative ones (e.g. six for Group 6), albeit the contrasts were less sharp; the total frequency of the theme 'Affective' was modest. For the other themes/subthemes, no distinct patterns could be identified. In Table 6, each (sub)theme is described and illustrated with sample quotes.

5 DISCUSSION

5.1 Research Questions Revisited

As mentioned in Introduction (Section 1), the research work presented in this paper has been motivated by several open questions, which we revisit here based on the analysis results of the empirical data collected (Section 4).

Are chatbots' humanlike features as important for trust as their ability to reliably provide support? Is trust in chatbots for customer service dominated by their conversational performance rather than their humanlikeness? No. We only saw significantly higher trust levels for humanlike chatbots (N=121, Mean = 5.73, SD = 1.16) compared to non-humanlike chatbots (N = 130, Mean = 5.24, SD = 1.42) ($t(349) = 2.97$, $p<.05$) in TT1: participants' first encounter with the chatbot without any conversational performance issue. However, as soon as participants ran into conversational performance issues in TT2, humanlikeness became nonsignificant (Humanlike: N=121, Mean = 3.92, SD=2.23; Non-humanlike: N = 130, Mean = 3.45, SD=2.13; $t(349) = 1.704$, $p>.05$). Even when the conversational performance was restored in TT3, the humanlikeness remained nonsignificant, which might be attributed to the users' reduced sensitivity to the chatbot's look and tone, even after only two short interaction episodes.

Table 6. Descriptions of nine (sub)themes with the total frequency of representative concepts in brackets and sample quotes cited with the participant identifier: G(one-digit group number, 2-digit serial number)

(Sub)Theme	Description	Sample quotes
Underlying functionality (54)	Analysing the chatbot performance with regard to algorithms and other technical mechanisms.	Automated features should be able to pick up what I want from a variety of input. (G629) I don't trust chatbot because if you enter something not inside his algorithm the chatbot doesn't understand. (G623)
(Mis-) Understanding (67)	The capacity of the chatbot to understand the requests, and the effect on trust if it doesn't.	I tried phrasing it many different ways and it just didn't understand the request. A better-functioning chatbot would gain more trust. (G608) I think the most important is the chatbot being able to understand the task I am asking for (G407)
Legitimacy (47)	The accuracy of the chatbot responses is verifiable via external resources, including people.	I would trust a chatbot like this if i knew it was from the legitimate website of my bank (G642) Accurate and descriptive information with the option to speak to a real person to confirm and clarify details (G418)
Nature of Task (35)	Trust is impacted by the complexity of the task to be handled by the chatbot.	I trust the answers to simple questions, such as the ones that were asked, but am hesitant about more complex questions. (G223) The chatbot can give me basics like rates, and I would trust that info, but I still felt I needed to talk to a real person to apply. To apply by chatbot makes me a bit uncomfortable (G126)
Conversational Style (50)	The way the chatbot delivers information, such as follow-up questions, grammar, the overall tone.	The information being clear and concise. The tone being helpful and friendly inspires confidence. (G117) For me to trust its information and advice I would need it to speak in a professional, although ideally not robotic, manner. (G603)
Nature of Information (147)	The characteristics of the information given by the chatbot (e.g. clear, accurate).	It is important to me that a chatbot is straight to the point and provides detailed information, this makes it seem more reliable and trustworthy. (G104) That the chatbot provides reliable, accurate, precise information. (G229)
Affective (34)	Users' positive or negative experiential responses to chatbot interactions.	The most important thing would be for the chatbot to be more warm and friendly. (G412) It was a fast chat service, I'm happy.(G114)
Cognitive (56)	User perception of the chatbot performance (e.g. speed, ease of use).	Using a simple and effective way to communicate with me. also understanding my language in a fast and effective manner. (G339) It was simple and easy to use. The chatbot seemed genuine and friendly. (G535)
Privacy (13)	Whether the chat content is secured and private.	It's important to me to know the chat is encrypted. (G417) I would like to know that my private information is protected.(G606)

Moreover, as can be seen in Table 5, the main effect of Humanlikeness was significant for four of the seven variables whereas the main effect of Conversational Performance was significant for five of them. Additionally, in the cases when both IVs were significant, the effect size of Conversational Performance was consistently higher than that of Humanlikeness. Furthermore, it was found that the perceived reliability (TBR) of the chatbot was not influenced significantly by the chatbot's humanlikeness (Humanlike: N=121, Mean = 4.29, SD = 1.55; Non-humanlike: N=130, Mean =3.93, SD = 1.53; $t(349) = 1.81$, p>,05).

This suggests that the conversational performance seems the decisive factor for both trust and TBR, contrasting earlier findings [23, 41, 45, 55]. We speculate that our operationalisation of humanlikeness may be one reason for this difference. For example, our humanlike interactions were short and not embodied, in contrast to the setup in [23], potentially weakening the impact of humanlikeness. Additionally, our participants' orientation towards the chatbot may have played a role. People typically have a more relational or utilitarian orientation towards any entity: essentially relating to whether you care mostly for the service provider or rather the service itself [46].

In our experiment, participants engaged with an unknown chatbot working for a fictitious bank, doing little to stimulate a relational orientation. From this stance, it makes sense that our participants cared little about the way that the chatbot approached them and that it mattered much more whether it was able to help them complete their assigned tasks, matching other work relating people's orientation to the way they experience interactions with a robot [32].

Does chatbot humanlikeness strengthen trust resilience in the face of performance issues? No. There were significant differences in task-specific trust across the three tasks (H6). However, there was no interaction effect between the two IVs – Humanlikeness and Conversational Performance (Table 5). In other words, each IV exerted its effect on trust in the chatbot independently, and humanlikeness did not impact the effect of conversational performance on trust in case of a breakdown.

This becomes even clearer from the post hoc Tukey's tests: these showed that the difference in trust ratings between TT2 (breakdown) and TT3 (no breakdown) were nonsignificant between Group 3 (humanlike, breakdown with no repair) and Group 6 (non-humanlike, breakdown with no repair). Moreover, for both groups, there was a notable increase of trust from TT2 to TT3 (Figure 2a), but the magnitude was comparable, suggesting that the humanlikeness (Group 3) did not strengthen trust resilience. Furthermore, the nonsignificant pairwise correlations TT1-TT2 and TT2-TT3 for Group 3 and Group 6 (H7) suggest trust resilience, which seems independent of the effect of humanlikeness.

5.2 Practical Implications

From an industry point of view, these results carry some important practical implications. First, it provides solid evidence to look beyond humanlikeness when creating or improving trust in chatbots. It may feel like an intuitive step to "simply make the bot more humanlike" in order to improve trust, but our results show that this may be a misguided effort. There may be good reasons to dedicate time to make a chatbot more humanlike. For example, humanlike elements like pronouns, greetings or even jokes and emoji may be necessary to make sure that the conversations are consistent or that the replies fit in with the general personality that the chatbot is supposed to have [13]. Alternatively, if the goal is to improve hedonic experiences when people chat with the bot, adding humanlikeness may also be a good way to spend one's time [25]. However, it is crucial that teams working with chatbots understand that it will not necessarily lead to higher trust in end-users, nor that it will help their bot become more trust resilient in the face of conversational breakdown. This knowledge enables chatbot teams to be critical about when humanlikeness is beneficial to their problem or question at hand - and when it is not.

Conversations will inherently break down at one point or another, same with chatbots as it is for conversations between humans [22]. This means that the opportunity for repair will inevitably show itself for every chatbot out there. Based on our results, missing this opportunity will clearly be costly in terms of trust. Hence, chatbot teams need to spend time and resources implementing and thinking about conversational repair. In our setup, repair was enacted through a generic fallback message: the chatbot acknowledged its inability to understand and asked participants to please repeat their inquiry again, phrased differently. Therefore, thinking about the exact phrasing of these fallback messages is important: how do you want to ask people to repeat themselves? How often will you allow the chatbot to fail before it "gives up"? What options are available in case the chatbot really is unable to help the user?

5.3 Limitations and Future Studies

From our results, it would seem that any repair is better than no repair. However, it needs to be noted that our repair was always successful: the chatbot's replies were programmed in such a way that it would give the appropriate response after users tried reformulating once. In real life situations, it is likely that repair will at times be unsuccessful. Future work should investigate the impact on trust of unsuccessful repair and potentially find ways to mitigate it.

Second, there is likely to be a trade-off between trying to prevent breakdown and conversational efficiency [19]. Our results may suggest that chatbot teams should aim to reduce the chance of breakdown as much as possible. Although one way to do that would be to improve the chatbot's model to increase prediction accuracy, another approach would be by only allowing button replies, limiting the user's free input and thereby reducing the chance of breakdown. Although there could be benefits in doing so [25], it may also lead to inefficient conversations where users need to click through a dialogue that is much more tedious than letting them phrase their inquiry directly. Knowing where on this trade-off the optimal point lies would be invaluable for any team working with chatbots.

Third, while our findings clearly indicate that chatbot humanlikeness is of lesser importance to trust than conversational performance and that is it does not entail significant trust resilience, it is important to be aware that humanlikeness may be obtained by a range of design cues. Possibly, other design cues, such as cues providing a stronger sense of conversational intelligence (e.g. [50]) or embodied emotions [23] could have led to a more substantial impact of humanlikeness. Furthermore, the relative simple tasks and interactions involved in the experiment may represent a limitation. While generic responses to user frequent questions, similar to those in the experiment, is a prevalent use-case in chatbot for customer service , tasks involving more complex goals and longer interactions may potentially yield different results. Future research is needed to investigate the possible impact of nuances in cues for humanlikeness as well as task and interaction complexity.

Fourth, the order of the tasks may have also influenced our results: there may be spill-over effects between the subsequent tasks that our setup does not allow us to isolate. To illustrate, participants in the no-repair condition are likely to have been influenced by the unsuccessful repair in Task 2 when they were asked to assess their trust in the chatbot to help them with Task 3. Despite our phrasing of the task specific trust measurement, which tried to be focused on the task at hand instead of the chatbot in general, it would be difficult for any participant to isolate their recent previous experience when answering. Interestingly, we observe that the task specific trust levels in Task 3 return to levels comparable to Task 1 - even in the no-repair condition, and regardless of whether the chatbot was humanlike or not. Future studies should aim to isolate and study these spill-over effects and the extent to which trust can "bounce back" after a conversational breakdown.

Fifth, there is little known work on the longitudinal development of trust in chatbots. Our results show that, by Task 3, none of our conditions had a significant effect any longer on the task-specific trust measurement. This could mean that users' trust in chatbots is settled within three short interactions.

Last, our qualitative results indicate that the nature of the task forms an important aspect in determining trust in the chatbot. Depending on the task, users may place different requirements on the chatbot to trust it. For example, we saw that users asked for ways to assess legitimacy of the answers when they asked for interest rates, or for confirmation when the chatbot made an appointment for them. Moreover, participants explicitly referred to some inquiries as simple and others as complex. This alludes to some kind of scale where simple questions are meant for chatbots and complex ones for humans, and where trust in chatbots is lost as soon as the chatbot cannot answer the simple kind.

As of now, there is no work that categorises the different (aspects of the) tasks that users may execute with chatbots. Future work should aim to close this gap in knowledge.

6 CONCLUSION

Chatbots are increasingly used in many different sectors of life. Online banking as a common and critical customer service can benefit from this burgeoning technology. While the bank used for our study is fictitious, the scenarios selected are representative of real-life use cases for chatbots in this context. Hence, insights gained from our empirical results can have practical implications for the design of such chatbots in industry (Section 5.2).

Among others, one intriguing implication is the utility and desirability of making autonomous AI-infused systems, including chatbots, humanlike. The chatbot's humanlike appearance and conversational style seem expendable; their absence may reduce trust but not to a profound extent. What determines a user to trust a chatbot seems essentially its demonstrated ability to give an accurate response to a request quickly, without requiring extra effort from the user (e.g. repeating or rephrasing the request), as indicated by our qualitative data (Section 4.2). In fact, our findings suggest that humanlikeness cannot enhance trust resilience (Section 4.1). Users seem able to 'forget' the damaged trust caused by breakdowns without repair and 'renew' their trust to the level close to that at the initial interaction with the chatbot when everything proceeds smoothly. These observations stimulate us to examine the issue about the relation between trust and emotion (positive and negative affective responses) (cf. Figure 3), which should not be conflated. There exist some empirical findings, albeit limited, which indicate that positive emotions can enhance trust whereas negative emotions undermine it (e.g. [15]). Future research on trust in chatbots should provide more empirical evidence to this issue.

Another challenge in the research area of trust in AI-infused systems is the measurement of trust. In our study, we adopted existing questionnaires on trust with established psychometric properties. Nonetheless, we are aware of the inherent limitations of using questionnaires as a research method. Ongoing work on developing viable and valid approaches to measuring trust (e.g. multisensory data) may advance our understanding of different factors shaping trust in chatbots (e.g. [7]). Overall, the research, design and application of chatbot are still facing a number of challenges [16]. The work presented in this paper contributes to the body of knowledge that helps resolve some of these challenges.

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 (2021), 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 (2018), 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 (2021), 632–658.

[5] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *International conference on internet science*. Springer, 377–392.

[6] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[7] 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 Publications Sage CA: Los Angeles, CA, 1595–1599.

[8] 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 (2019), 539–548.

[9] 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 (2003), 737–758.

[10] Cammy Crolic, Felipe Thomaz, Rhonda Hadi, and Andrew T Stephen. 2022. Blame the bot: anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing* 86, 1 (2022), 132–148.

[11] Robert Dale. 2016. The return of the chatbots. *Natural Language Engineering* 22, 5 (2016), 811–817.

[12] 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 (2016), 331.

[13] Diana Deibel and Rebecca Evanhoe. 2021. *Conversations with Things: UX design for Chat and Voice.* Rosenfeld Media.

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

[15] Jennifer R Dunn and Maurice E Schweitzer. 2005. Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology* 88, 5 (2005), 736.

[16] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* 103, 12 (2021), 2915–2942.

[17] Asbjørn Følstad and Anders Mærøe. 2022. The Ethics of Chatbots in Public Sector Service Provision. In *CUI@CHI2022: Ethics of Conversational User Interfaces*. ACM, New York, NY, USA.

[18] 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.

[19] 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 (2021), 1–17.

[20] 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 (2019), 304–316.

[21] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11.

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

[23] 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 *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 493–500.

[24] 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 (2021), 1196–1229.

[25] 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 (2022), 102788.

[26] 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 (2022), 212–229.

[27] Leslie Joseph. 2018. *The Six Factors That Separate Hype From Hope In Your Conversational AI Journey.* Technical Report. Forrester. 10 pages. https://www.forrester.com/report/The+Six+Factors+That+Separate+Hype+From+Hope+In+Your+Conversational+AI+Journey/-/E-RES143773

[28] Knut Kvale, Eleonora Freddi, Stig Hodnebrog, Olav Alexander Sell, and Asbjørn Følstad. 2020. Understanding the user experience of customer service chatbots: what can we learn from customer satisfaction surveys?. In *International Workshop on Chatbot Research and Design*. Springer, 205–218.

[29] 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.

[30] 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 (2015), 1.

[31] Effie Lai-Chong Law, Asbjørn Følstad, Jonathan Grudin, and Björn Schuller. 2022. Conversational Agent as Trustworthy Autonomous System (Trust-CA)(Dagstuhl Seminar 21381). In *Dagstuhl Reports*, Vol. 11. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[32] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

[33] Sangwon Lee, Naeun Lee, and Young June Sah. 2020. Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. *International Journal of Human–Computer Interaction* 36, 10 (2020), 930–940.

[34] 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 (2011), e25085.

[35] Brian Manusama, Bern Elliot, Magnus Revang, and Anthony Mullen. 2019. *Market Guide for Virtual Customer Assistants.* Technical Report ID G00349067. Gartner. 29 pages. https://www.gartner.com/en/documents/3947357/market-guide-for-virtual-customerassistants

[36] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[37] 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 (2011), 1–25.

[38] Michael McTear. 2020. Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies* 13, 3 (2020), 1–251.

[39] Robert J Moore. 2018. A natural conversation framework for conversational UX design. In *Studies in Conversational UX Design*. Springer, 181–204.

[40] 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 (2019), 317–335.

[41] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797.

[42] 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 (2021), 102630.

[43] Grand View Research. 2022. *Chatbot Market Size, Share & Growth Report 2022-2030*. Technical Report GVR-1-68038-598-4. Grand View Research. 132 pages. https://www.grandviewresearch.com/industry-analysis/chatbot-market

[44] Magnus Revang and Anthony Mullen. 2022. *Magic Quadrant for Enterprise Conversational AI Platforms*. Technical Report ID G00748698. Gartner. 33 pages. https://www.gartner.com/doc/reprints?id=1-28HJXCBH&ct=211221&st=sb

[45] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96.

[46] Torsten Ringberg, Gaby Odekerken-Schröder, and Glenn L Christensen. 2007. A cultural models approach to service recovery. *Journal of Marketing* 71, 3 (2007), 194–214.

[47] 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 (1998), 393–404.

[48] Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research* 32, 3 (2021), 736–751.

[49] 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.

[50] 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 (2018), 94–102.

[51] Amir Shevat. 2017. *Designing bots: Creating conversational experiences*. O'Reilly Media, Inc., Boston, US.

[52] Statista. 2019. *Share of consumers who have used chatbots to engage with companies in the United States as of 2019, by industry*. Technical Report. Statista. https://www.statista.com/statistics/1042604/united-stated-share-internet-users-who-used-chatbots-industry/

[53] 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.

[54] Joseph Weizenbaum. 1976. *Computer power and human reason: From judgment to calculation*. WH Freeman & Co, New York, NY, USA.

[55] 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 (2017), 1–20.

[56] Juliana JY Zhang, Asbjørn Følstad, and Cato A Bjørkli. 2021. Organizational factors affecting successful implementation of chatbots for customer service. *Journal of internet commerce* (2021), 1–35.

A HIERARCHICAL CODING SCHEME