Is Making Mistakes Human? On the Perception of Typing Errors in Chatbot Communication

Bührke, Johannes
University of Goettingen, Germany
johannes.buehrke@stud.uni-goettingen.de

Brendel, Alfred Benedikt
Technische Universität Dresden, Germany
alfred_benedikt.brendel@tu-dresden.de

Lichtenberg, Sascha
University of Goettingen, Germany
sascha.lichtenberg@uni-goettingen.de

Greve, Maike
University of Goettingen, Germany
maike.greve@uni-goettingen.de

Mirbabaie, Milad
University of Bremen, Germany
milad.mirbabaie@uni-bremen.de

Abstract

The increasing application of Conversational Agents (CAs) changes the way customers and businesses interact during a service encounter. Research has shown that CA equipped with social cues (e.g., having a name, greeting users) stimulates the user to perceive the interaction as human-like, which can positively influence the overall experience. Specifically, social cues have shown to lead to increased customer satisfaction, perceived service quality, and trustworthiness in service encounters.

However, many CAs are discontinued because of their limited conversational ability, which can lead to customer dissatisfaction. Nevertheless, making errors and mistakes can also be seen as a human characteristic (e.g., typing errors). Existing research on human-computer interfaces lacks in the area of CAs producing human-like errors and their perception in a service encounter situation. Therefore, we conducted a 2x2 online experiment with 228 participants on how CAs typing errors and CAs human-like behavior treatments influence user's perception, including perceived service quality.

1. Introduction

Companies are increasingly investing in new technologies to increase effectiveness and efficiency in service encounters while maintaining high customer satisfaction [1]. One current and frequently applied technology for optimizing service encounters are Conversational Agents (CAs), which replace calling a service number, searching in FAQs, or writing a detailed e-mail with a natural language interface (e.g., a chatbot chatting with customers via written language) [2], [3]. In general, CAs are defined as "*software-based systems designed to interact with humans using natural language*" [4]. The capabilities of CAs are increasing, driven by improvements in machine learning and natural language processing. Overall, the technology for CA development is now widely available (e.g., Google Dialogflow, IBM Watson, Tensorflow), which has led to a widespread application of CAs in practice [5], improving the way customers and information systems (IS) interact [4]. CAs shift the service encounter interaction from interpersonal human-to-human interaction to a computer-to-human interaction, addressing most requests independently and automatically [6]. While human service encounters have a limited time availability and capacity, CAs can support customers at any time, at any place, and can provide a comfortable and convenient user experience [7].

However, many CAs have been discontinued because of their lack of providing meaningful responses and engagement in an interactive dialogue [8], rendering the pursue to understand CA design to be highly relevant for practice and research [4], [9]. Following Gnewuch et al. [3], CAs can be designed with a variety of social cues, ranging from visual (e.g., emoticons) to verbal (e.g., greetings), to auditory (e.g., the gender of voice), and invisible cues (e.g., response time) [4]. Research has shown that the interaction with CAs equipped with social cues leads to social responses by users (i.e., users are mindlessly responding to the CA as if it was human) [9]. Based on this effect, various studies reported that the human-like design of a CA (e.g., communicating in natural language, having a name) could lead to increases in perceived service quality [3], enjoyment [10], and perceived trustworthiness [11]. Overall, implementing a successful CA depends on the appropriate design of the human-like features [12].

Against this background, human-like conversation features have yet to be investigated. Lortie and Guitton [13], whose research focused on conversational agents language, note that the quality of writing (e.g., grammar) affects judgement to perceived humanness,

URI: https://hdl.handle.net/10125/71158
978-0-9981331-4-0

while mistakes as a contributor to perceived humanness were not be identified significant. Westerman et al. [14] attribute chatbots that make typographical mistakes with less perception of humanness in the context of information privacy. However, regarding aspects of service encounters, it remains unclear how CAs are perceived when their messages contain typing errors. Against this background, we formulate the following research question:

How do typing errors of a CA influence the user's perception of the CA in a service encounter?

We applied a 2x2 (human-like design x making typing errors) online experiment with 228 participants to address this research question. In our experiment, we investigated how the overall human-like design interacts with messages, including typing errors. Specifically, we analyze the effects on perceived social presence, perceived humanness, and service quality. Our results revealed that applying typing errors (based on common human errors [15]) had no positive impact on perceived humanness, perceived social response, and service satisfaction, contrary to the expected outcomes. Based on the participants' comments, the typing errors were attributed to the developer rather than the CA.

2. Research Background

2.1. Conversational Agents in the Context of Service Systems

Overall, the capabilities of CAs to interact via written or verbal language with customers has improved significantly in recent years [11], [16]. To this end, CAs use machine learning mechanisms and algorithms to improve the human-computer interface [17]. Due to the enhancement of technological capabilities in the past years, CAs have a huge potential to increase customer satisfaction and service quality [4], [11].

In practice, CAs offer convenient access to information or managing customer requests [16], [6]. However, with the increasing capabilities of CAs, customer expectations are also rising [16], [17]. Currently, CAs are still prone to produce errors (e.g., not understanding user input) and remind the users that they are still interacting with a machine [7], [18]. Failures are caused by natural language processing problems (e.g., limited vocabulary) but also by errors associated with the human-computer relation (e.g., inappropriate use of emoticons) [18]. Overall, limited CA capabilities or ineffective design lead to negative user perception, dissatisfaction, and a lack of utilization [8]. Thus, research and development efforts have to be made in two areas. First, improving the technical aspects of CAs, like architecture and algorithms [17]. Second, understanding how the human-like design of CA (e.g., application of social cues, such as having a name and an avatar) shapes affective, cognitive, and behavioral responses of users [4], [6], [9], [19].

2.2. Social Responses to Conversational Agents

Since CAs use natural speech and social cues, the communication contains behavioral and social characteristics that can strongly influence the conversational interaction. In this way, the interaction between CA and human begins to feel similar to a communication between two real people in real life [3]. In this context, the Computers Are Social Actors (CASA) paradigm presents a framework to study CAs and their human-like design (i.e., social cues) [1], [9], [11].

According to Nass and Moon [9], the paradigm implies that people thoughtlessly apply social rules and expectations to everything that has human-like traits or behavior, including computers [9], [20]. Hence to improve CA design, current research addresses various social cues of CAs, such as virtual characters [21], emoticons [7], typefaces [22], degree of interactivity [23], communication style [7], or assumed agency [24]. However, some experiments report users perceiving CAs as uncanny [16], when their expectations and the human-like design are in dissonance [25], [26]. Despite the richness and ever-increasing body of current research, not all available social cues have been studied. In this study, we would like to focus on the social cues of typing errors, which we will further conceptualize in the following.

2.3. Typing Errors as a Social Cue

The conversational abilities of CAs are continuously improving, leading to nearly no language errors in the future. However, real human communication is not always flawless [27], making mistakes and errors a human characteristic. For instance, making typing errors can create a sense of connection with another individual through the interface [28]. During a written conversation, errors can occur when manually entering text via keyboard, so-called typographical, or typing errors [15], [28]. MacNeilage [15] describes typing errors, "*as Clues to Serial Ordering Mechanisms in Language Behavior."* In the study of McNeilage, five different categories of typing are identified (Table 1.) [15]. The most common are spatial and temporal errors. Spatial errors result from typing a letter immediately adjacent to its keyboard neighbor (e.g., "g" for "h"), while temporal errors occurring when the order in which the required letters were typed was wrong (e.g., "th" for "ht") [15].

Table 1: Extract of typing error classification (MacNeilage) [15]

<table>
<tr><th colspan="4">Classification of typing errors</th></tr>
<tr><td colspan="4">Spatial errors - result from entering a letter directly next to the key that is required. They can be divided into:</td></tr>
<tr><td>Horizontal errors:</td><td>Vertical errors:</td><td colspan="2">Diagonal errors:</td></tr>
<tr><td>Consists of entering a letter immediately to the left or right of the intended keyboard letter, in the same row on the keyboard.</td><td>Consists of entering a letter just above or below the desired letter in the same column of the keyboard.</td><td colspan="2">A keyboard letter is entered in a line and a column connected to the letter but is not the intended one.</td></tr>
<tr><td>e.g., "e" for "r", or "d" for "f"</td><td>e.g., "f " for "r" or "e" for "d".</td><td colspan="2">e.g., "d" for "r".</td></tr>
<tr><td colspan="4">Temporal errors – are any errors that occur in the order in which the required letters were typed. These were divided into:</td></tr>
<tr><td>Reversal errors:</td><td>Omission errors:</td><td>Equivocal errors:</td><td>Anticipation errors:</td></tr>
<tr><td>Occur when two letters next to each other in the correct order are in reverse order.</td><td>Occurs when a letter in a sequence is omitted.</td><td>Occurs when the letter which is stroked, is one stroke ahead of the one required. Afterward, the context stops as the user becomes aware of the error.</td><td>Occurs when a letter is typed more than one stroke ahead of the required one.</td></tr>
<tr><td>e.g.,"ht" for "th"</td><td>e.g., "lenth" for "length"</td><td>e.g., “stiml-” for “stimulus”</td><td>e.g., "ext-" for "expected"</td></tr>
<tr><td colspan="4">Miscellaneous errors - are a series of further specific types of errors that can be distinguished.</td></tr>
<tr><td>Interpola-tion errors:</td><td>Type errors:</td><td>Dynamic errors:</td><td rowspan="2">Contralat-eral errors</td></tr>
<tr><td rowspan="2">Consists of a letter that seems to have nothing to do with the correct order was inserted.</td><td rowspan="2">Occurs when a letter of a word is changed, making it similar to a similar word but not in context.</td><td rowspan="2">Occurs when the letter in the sequence to an adjacent double-typed letter is typed twice.</td></tr>
<tr><td rowspan="2">Phonemic errors</td></tr>
<tr><td>e.g., “formend” for “formed”</td><td>e.g., “that” for “than”</td><td>e.g., “eroors” for “errors”</td></tr>
<tr><td colspan="4">Multiple classification errors - can be placed in more than one category</td></tr>
<tr><td colspan="4">Unclassifiable errors - cannot be placed in any above categories</td></tr>
</table>

Regarding the influence of typing errors on the relation of CAs and users, recent studies reported that users were less likely to share private information [14] but also perceived less trust [29] when a CA makes typing errors. Furthermore, Westermann et al. [14] report that typing errors of CAs contribute to a lower level of perceived humanness. However, to the best of our knowledge, the effect of CAs with typing errors in a service encounter has yet to be investigated.

3. Research Model and Hypotheses

Our research seeks to contribute to a better understanding of the effect of CAs with typing errors on the perception of users. For this purpose, we developed an online experiment, modeled after an e-bike rental service. As illustrated in Table 2, we use human-like design and typing errors as dimensions for our 2x2 experiment design, leading to the development and application of four chatbot instances.

Table 2: Initial Setup of CA Instances

Setup of CAs Instances	No-Typing Errors	Typing Errors
Non-Human-Like Design	Non-Human-like and No-Typing Errors (R)	Non-Human-like and Typing Errors (TE)
Human-Like Design	Human-like and No-Typing Errors (H)	Human-like and Typing Errors (HTE)

Besides applying social cues [4], [11] to induce a human-like perception (e.g., trustworthy, politeness), we utilize typing error patterns based on MacNeilage [15] in our service encounter situation. Based on the findings of MacNeilages' [15] research, we design the CA in such a way that intentional typing errors mimic human typing errors. Applying both cues to our CAs, we hypothesize that human-like design [12] and typing errors [15] contribute to perceived social presence, perceived humanness, and service quality. Figure 1 summarizes the hypotheses.

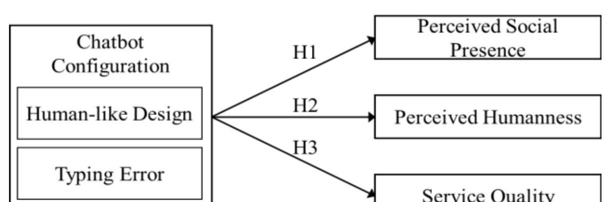

Figure 1. Research Model

3.1 Social Presence

Social cues in the form of avatars or emotions have positively stimulated social presence when an individual is interacting with CAs [7]. Social presence is understood as the degree of salience of the other person in a mediated communication and the consequent salience of their interpersonal interactions and has been shown to likewise exist without actual human contact [28]. Investigated by Mirning et al. [30], faulty CAs were significantly rated more likable than flawless CAs. Consequently, we assume that CAs designed with social cues (in this study called human-like design), even considered by making errors, are expected to yield a higher level of perceived social presence. Thus, we postulate the following hypotheses:

H1a: A human-like design of a CA leads to a higher perceived social presence.
H1b: Typing errors of a CA lead to a higher perceived social presence.
H1c: The combination of human-like design and typing errors of a CA leads to higher perceived social presence.

3.2. Humanness

Social cues play an essential factor in designing a human-like CA [4], [11], [31]. For instance, CAs with a name and their own unique customized behavior patterns can positively contribute to the perceived humanness [11], [19]. However, human-like behavior is not always flawless, as making errors and mistakes can occur (e.g., typing errors) [15].

Consequently, we assume that users will perceive different levels of humanness based on different CA treatments. Subsequently, CAs designed with social cues are expected to yield a higher level of humanness. Furthermore, we expect that human-like designed CA with human-like classified typing errors [15] also yields a higher level of perceived humanness. Follows, we set up the following hypotheses:

H2a: A human-like design of a CA leads to higher perceived humanness.
H2b: Typing errors of a CA lead to higher perceived humanness.
H2c: The combination of human-like design and typing errors of a CA leads to higher perceived humanness.

3.3. Service Quality

Service Quality is one of the crucial indicators for successful customer bonding and satisfaction [1], [32], [33]. In essence, service quality results from a comparison between the expectation and outcome of a service [32]. Following Parasuraman, service quality comprises service reliability, assurance, empathy, responsiveness, and tangible aspects [32]. Research by Yan, Solomon, and Mirchandani et al. [34] identified that human agents provide a higher level of service quality than CAs. Since we do attest that human-like classified typing errors contribute to perceived humanness positively and human-like design contributes positively to service quality [32]–[34], we expect that human-like design and human-like design combined with typing errors will positively influence service quality. Against this background, we propose the following hypotheses:

H3a: A human-like design of a CA leads to higher service satisfaction.
H3b: Typing errors of a CA lead to lower service satisfaction.
H3c: The combination of human-like design and typing errors of a CA leads to higher service satisfaction.

4. Research Design

4.1. Data Collection Procedure and Sample

Before the experiment's actual beginning, all participants were provided with a preliminary introduction document in which the context and structure of the following experiment were explained. Subsequently, the tasks within the service encounter were described. We provided each participant with the same document to ensure that the participants had the same information relevant to the experiment [35]. To check if the experiment and the processes were understood correctly, comprehension questions had to be answered to proceed and be assigned one of the chatbot instances. The document contained a link that randomly assigned the participants to the experiments to apply a non-biased assignment to the participants. Hence, one of the four chatbot configurations (see Table 2) was randomly assigned. After the interaction, the participants were forwarded to a survey considering quantitative but also qualitative feedback.

In the experiment, each participant was supposed to make an e-bike rental booking via a chatbot interface, where different locations and types of bikes were offered to be selected. The conversation was divided into six steps building upon each other: (1) Introduction and clarification of needs, (2) indication of the desired booking day for the e-bike, (3) indication of the location, (4) selection of a bicycle-type depending on the availability at the location, (5) indication of the name for the booking registration, and (6) confirmation of the appointment considering an e-mail address to be entered. On average, the participation time per experiment took around five minutes. The study conducted has a sample size of n = 228 participants, ranging from 17 to 64 years of age (M=24,54 years, SD=

5,76). Of the participants attending the experiment, a share of 46% was female and 52% male. 2% of the participants made no statement regarding their gender. The participants have been acquired through personal networks and social media. In addition, financial compensation was offered for participation in the form of a raffle. Among all participants, a total of three 10€ online shopping vouchers were raffled.

4.2. Configurations

For our 2x2 experiment, we developed our chatbots instance via Google's "Dialogflow" framework and a custom-made web interface (see Figure 2). All CAs received the same set of training phrases. Hence, each chatbot understood the statements and intentions of the customer entries made during the service encounter to respond to the users' input. They could extract parameters, such as bicycle types and storage locations, and use them paraphrased in the subsequent

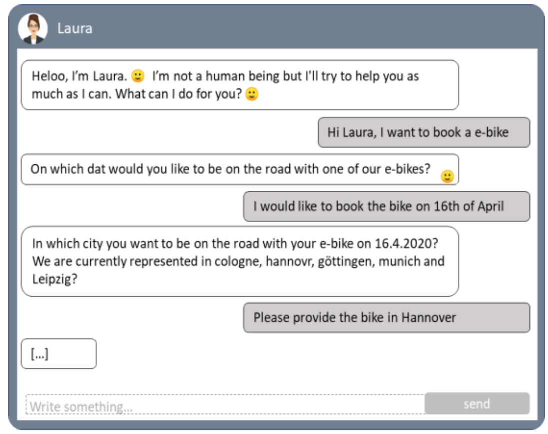

Figure 2: Web interface of a human-like chatbot with typing errors (translated to English from German)

Table 3: Response examples

CA Instance Setup			Error Classification (MacNeilage 1964)		
Chatbot Setup	**Typing Error**	**Response Examples**	**Typing Error Example**	**Classification**	**Subcategory**
Non-human-like design	No typing errors	*"Welcome to the e-bike rental"*			
		"Enter the date:"			
		"In which city should the e-bike be on 16.4.2020? (Cologne, Hannover, Göttingen, Munich, Leipzig)"			
	Typing errors	*"Welcome to the e-bike erntal."*	*"erntal" instead of "rental"*	Spatial error	Horizontal error
		"Enter the dat:"	*"dat" instead of "date"*	Temporal error	Omission error
		"In which city should the e-bike be on 16.4.2020? (cologne, hannover, göttingen, munich, leibzig)"	*„Leibzig" instead of „Leipzig"*	Spatial Error	Vertical error
Human-like design	No typing errors	*"Hello, I'm Laura. [emoticon] I'm not a human being but I'll try to help you as much as I can. What can I do for you?[emoticon]."*			
		"On which date would you like to be on the road with one of our e-bikes? [emoticon]"			
		"In which city you want to be on the road with your e-bike on 16.4.2020? We are currently represented in Cologne, Hannover, Göttingen, Munich and Leipzig. [emoticon]"			
	Typing errors	*"Heloo, I'm Laura. [emoticon] I'm not a human being but I'll try to help you as much as I can. What can I do for you?[emoticon]."*	*"Heloo" instead of "Hello"*	Dynamic Error	Miscellaneous error
		"On which dat would you like to be on the road with one of our e-bikes? [emoticon]"	*"dat" instead of "date"*	Temporal error	Omission error
		"In which city you want to be on the road with your e-bike on 16.4.2020? We are currently represented in cologne, hannovr, göttingen, munich and leipzig.[emoticon]"	*"hannovr" instead of "hannover"*	Spatial Error	Vertical error

Note, all responses are translated from German. Examples of typing errors are adapted to the language of study.

conversation. The non-human-like designed CAs were not with additional social cues, while the human-like CAs were (e.g., greeting and using emoticons). Per the suggestions of Seeger et al. [12], the human-like CAs were designed with human identity, verbal, and non-verbal capabilities, as illustrated in Figure 2.

The human-like chatbots were stating a personal introduction at the beginning of the conversation, for example, "Hello, I'm Laura. I'm not a human being but I'll try to help you as much as I can. What can I do for you?". Additionally, the chatbots were applying emoticons within the customer interaction.

In regard to non-verbal human-like, they were using blinking dots and dynamic response delay. Depending on the response texts' length, as suggested by Gnewuch [16], a process response time deviation simulates how a service employee thinks and type the response text-message. Furthermore, the chatbots were able to understand different variations of sentences and elicit the intended meaning.

Regarding the typing errors, both chatbots (TE and H+TE) were equipped with typing errors following the categories of spatial, temporal, and miscellaneous errors, as identified by MacNeilage [15]. Thus, the instances designed with typing errors are used, such as "Heloo" instead of "Hello" which can be identified as a miscellaneous error. Furthermore, spatial errors (e.g., vertical error, "e-maijl" instead of "e-mail"), as well as temporal errors (e.g., omission error, "reserve" instead of "reserved"), have been used (see Table 3 for examples).

4.3. Measures and Descriptive Statistics

An online survey was conducted after the service encounter interaction between the human participants, and the CA was completed. The survey measured three different constructs (perceived humanness, perceived social presence, and service quality) by asking various items. The items were measured on a scale from 1 (not applicable at all) to 7 (is very accurate). The conducted survey design was based on established constructs used in previous studies [31]–[33], [36]. In order to check the attention of the participants, we have integrated control questions into the questionnaire. Therefore, the participants had to select a certain number on a scale

Table 4. Measurement of latent variables

Constructs and items	Loadings	Source
Perceived humanness (α = .832, CR = .835, AVE = .561)		
The CA seemed to be human-like.	.796	Holtgraves and Han [31]
The CA seemed to be well competent.	.616	
The CA seemed to be well conscientious.	.796	
~~The CA responded well to my answers.~~	~~.431~~	
The CA seemed to be well committed to my questions.	.776	
Perceived social presence (α = .920, CR = .921, AVE = .746)		
I felt a sense of human contact with the system.	.863	Gefen and Straub [28]
I felt a sense of personalness with the system.	.847	
I felt a sense of sociability with the system.	.853	
I felt a sense of human warmth with the system.	.893	
Service quality (α = .920, CR = .922, AVE = .500)		
~~*[R]* The CA provides services as promised.~~	~~.570~~	Parasuraman [32], Jiang [33]
[R] The CA is reliable in dealing with service problems of customers.	.645	
~~*[R]* The CA performs services correctly the first time.~~	~~.503~~	
[R] The CA delivers services within the promised time.	.600	
~~*[RE]* The CA keeps customers informed when services are running.~~	~~.560~~	
~~*[RE]* The CA provides customers with speedy service.~~	~~.584~~	
[RE] The CA is ready to help customers.	.645	
[RE] The CA is ready to respond to customer requests.	.637	
[A] The CA increases the confidence of customers.	.786	
[A] The CA makes customers feel secure in their transaction.	.779	
[A] The CA is polite throughout.	.692	
[A] The CA has the knowledge to answer customer questions.	.745	
[E] The CA gives customers individual attention.	.640	
[E] The CA treats customers with care.	.726	
[E] The CA works in the best interest of the customers.	.757	
[E] The CA understands the needs of its customers.	.712	
[T] The CA is modern.	.623	
~~*[T]* The CA is visually appealing.~~	~~.520~~	
[T] The CA has an elegant and professional appearance.	.563	
[R]= Reliability; [RE]= Responsiveness; [A]= Assurance; [E]= Empathy; [T]=Tangibles		

Note that all items were translated to German for the survey.

twice. To verify the factor loadings of the items for each construct, we conducted a confirmatory factor analysis CFA. Subsequently, only elements with a factor loading above the threshold value of .60 have been considered. We have further evaluated the constructs supported by Cronbach's Alpha (a) and the Composite Reliability (CR).

Both require a value larger than 0.80. In addition, the average variance extracted (AVE) requires at least a value of 0.50 [37]. Table 4 summarizes the constructs perceived social presence, statistically significant difference for perceived humanness (F(3,224)=0.38, p=.771) and service quality (F(3,224)=0.40, p=.754), while evidence for variance heterogeneity was found for perceived social presence (F(3,224)=16.34, p<.001). As there is no equivalent non-parametric test, we lowered the required perceived humanness, and service quality with its corresponding items and factor loadings. Weighted sum scores have been calculated, as suggested by DiStefano et al. [38], to create one metric variable for each construct.

5. Results

The survey data collected for the CA service encounter were analyzed using descriptive statistics and variance analysis to compare the three groups' mean values. Statistical software R was used for the analysis. We first considered the assumptions for variance analysis before we conducted a two-way ANOVA for each dependent variable. Due to the data measurement procedure, we ensured the sample's independence, as each participant only received one treatment and only conducted the survey once. Furthermore, the groups have similar sample sizes, as shown in Table 5. We checked that the residuals are approximately normally distributed by visualizing the residuals through a qq-plot for each group. Hence, we validated the approximate normal distribution. To validate variance homogeneity, we conducted the Levene's test. The test showed no evidence that suggests the variance across groups is a significance level for the two-way ANOVA from 5% to 1%. The two-way ANOVAs illustrate that for perceived social presence, the CA configuration typing errors (F(1, 224) = 1.13) is not significant in comparison to the others, while human-like design (F(1, 224) = 119.57, p<.001) and the interaction of human-like design combined with typing errors (F(1, 224) = 6.351, p=.012) show statistically significant effects. The ANOVA for the dependent variable perceived humanness reveals that configurations of typing errors (F(1, 224) = 7.39, p=.007) and human-like design (F(1, 224) = 111.55, p<.001) significantly influence the perceived humanness, while the interaction of both (F(1, 224) = 0.51) shows no significant effect. In contrast, the dependent variable of service quality shows only significant differences in human-like design (F(1, 224) = 55.66, p<.001). Since the overall results show significant main and interaction effects, the Tukey HSD post-hoc test was applied to provide detailed insights to the hypotheses through pairwise comparisons of the individual groups. With respect to our hypotheses, the dependent variables significantly vary in their mean values when comparing the human-like configuration with the control configuration. As illustrated in Table 6, we can support **H1a** because the **perceived social**

Table 5. Descriptive statistics and variance analysis with post-hoc comparison

Dependent variable *(Scale)*	**All**		**CA Treatment**				**Variance Analysis**	**Post-hoc comparison between groups**		
	(N=228)		**R** (N=59)	**TE** (N=50)	**H** (N=59)	**H+TE** (N=60)	Two-way ANOVAs	Comparison	Significance	Hypotheses
Perceived Social Presence *(Metric)*	Mean SD	2.61 1.49	1.63 0.74	1.80 0.92	3.76 1.58	3.13 1.35	TE: F(1, 224) = 1.13, n.s. H: F(1, 224) = 119.57, p<.001*** H+TE: F(1, 224) = 6.351, p=.012**	TE – R H – R H+TE – R H – TE H+TE – TE H+TE – H	p= .878 n.s. p < .001 *** p < .001 *** p < .001 *** p < .001 *** p =.023 *	H1b H1a H1a – – H1b; H1c
Perceived Humanness *(Metric)*	Mean SD	5.16 1.79	4.33 1.55	3.85 1.50	6.50 1.41	5.76 1.36	TE: F(1, 224) = 7.39, p=.007** H: F(1, 224) = 111.55, p<.001*** H+TE: F(1, 224) = 0.51, n.s.	TE – R H – R H+TE – R H – TE H+TE – TE H+TE – H	p= .309 n.s. p < .001 *** p < .001 *** p < .001 *** p < .001 *** p= .031 *	H2b H2a H2a – – H2b; H2c
Service Quality *(Metric)*	Mean SD	4.71 1.21	4.25 1.13	4.04 1.20	5.41 0.96	5.02 1.06	TE: F(1, 224) = 3.27, n.s. H: F(1, 224) = 55.66, p<.001*** H+TE: F(1, 224) = 0.55, n.s.	TE – R H – R H+TE – R H – TE H+TE – TE H+TE – H	p= .720 n.s. p < .001 *** p < .001 *** p < .001 *** p < .001 *** p= .205 n.s.	H3b H3a H3a – – H3b; H3c

SD = Standard Deviation, p = p-value, R = No-human-like and no-typing-error (control), TE = No-human-like and typing-error,
H = Human-like and no-typing-error, H+TE = Human-like and typing-error
Significance level: * 0.05; ** 0.01; *** 0.001; n.s. = not significant

presence is identified higher for human-like configurations H (M=3.76, SD=1.58) and H+TE (M=3.13, SD=1.35) in comparison to the control group R (M=1.63, SD=0.74). Also, **H2a** can be supported

because **perceived humanness** yields a higher mean value for human-like treatments as H (M=6.50, SD=1.41) and H+TE (M=5.76, SD=1.36) in contrast to the control group (M=4.33, SD = 1.55) illustrate. In addition, **H3a** connected to human-like design influencing the **service quality** can also be supported, as the treatments H (M=5.41, SD=0.96) and H+TE (M=5.02, SD=1.06) yield higher mean values than the control group R (M=4.25, SD=1.13). Conducting the configuration of typing errors on perceived social presence, perceived humanness, and service quality, we can identify that the treatment only affects **perceived humanness (H2b),** while **H1b** and **H3b** cannot be supported.

However, as H+TE (M=5.76, SD=1.13) yields a lower mean value in perceived humanness as H (M=6.50, SD=1.41) and the comparison between the mean values of perceived humanness of the group's TE and the control group R is not significant, **H2b** is not supported. Applying post-hoc comparison to CAs with human-like design combined with typing errors compared to a CA with the human-like design only, we cannot support **H1c**. Furthermore, since the mean of H+TE (M=3.13, SD=1.35) compared to H (M=3,76, SD=1.35) shows a higher value for perceived social presence, **H1c** has been contradicted. Validating the perceived humanness of H+TE (M=5.76, SD=1.36) compared to H (M=6.50, SD=1.41), we state **H2c** as contradicted as well. In addition, we cannot support the hypothesis **H3c**. As proposed already by our research design and our quantitative conducted statistical survey analysis, we allowed qualitative feedback in our survey. Our analysis found that participants reacted to our CA with the typing error configuration with direct feedback addressed to us as a developer. The participants stated that the developer should *"improve the CA's language skills"* and that *"typing errors are a sign of unprofessionalism"* in CA development.

6. Discussion and Implications

Our research contributes to the improvement of human-computer interaction in the context of CA design. Concerning the theoretical implications, our study supports existing research on the positive influence of human-like design [4], [11], perceived social presence, perceived humanness, and service quality. In previous research, the design of typing errors was influenced by randomness [14], while we considered a well-founded human-like typing error classification [15]. According to MacNeilage [15], in free-flowing writing situations, typing errors are typical. In this context, therefore, we consider chatbot typos plausible for perceiving errors as human-like. In another experimental context of a human-embodied system, Mirning et al. [30] found that error designed CAs were rated more likable than flawless CAs. However, we had to discover that typing errors in the human-computer chatbot interface did not increase the perceived humanness. Furthermore, our results indicate that typing errors lead to a negative effect on the perception of humanness and social presence. Based on our results, we would like to offer the following explanation. Users are not connecting typing errors with the human-like behavior of a CA. They assume that CA typing errors are a lack of developer competence. Participants called for an improvement of the chatbot before releasing it, stating *"the chatbot is not ready for society until the developer corrects the mistakes."*

For practice, our customer service encounter results provide prescriptive knowledge regarding the application of typing errors. Leaving out typing errors

Table 6. Results of hypotheses

Items	Hypotheses		Results
Perceived Social Presence	H1a	A human-like design of a CA leads to a higher perceived social presence.	**Supported**
	H1b	*Typing errors of a CA lead to a higher perceived social presence.*	*Not Supported*
	H1c	The combination of human-like design and typing errors of a CA leads to a higher perceived social presence.	**Contradicted**
Perceived Humanness	H2a	*A human-like design of a CA leads to higher perceived humanness.*	**Supported**
	H2b	*Typing errors of a CA lead to higher perceived humanness.*	*Not Supported*
	H2c	*The combination of human-like design and typing errors of a CA leads to higher perceived humanness.*	**Contradicted**
Service Quality	H3a	*A human-like design of a CA leads to higher service satisfaction.*	**Supported**
	H3b	*Typing errors of a CA lead to lower service satisfaction.*	*Not supported*
	H3c	*The combination of human-like design and typing errors of a CA leads to higher service satisfaction.*	*Not supported*
R = No-human-like and no-typing-error (control), TE = No-human-like and typing-error, H = Human-like and no-typing-error, H+TE = Human-like and typing-error			

promises a more pleasant user experience regarding the perceived social presence and perceived humanness.

In the following, we will discuss the limitations of our research and avenues for future research. Our experiment is based on a potential service encounter scenario. However, the participants did not book an e-bike, but only executed the request in a realistic experiment. A transfer to a real situation could lead to different user responses, such as a change in behavior and service rating.

As our study highlights the interrelation of typing errors connected to human-like cues, this study's result is highly dependent on the application. To understand the influence of typing errors on the human-like design of CAs for the future better, a deeper investment into the frequency of typing errors within a CA service encounter seems to be evident. It will also be interesting to see how users perceive a CA that reacts to his own produced errors (e.g., excusing for an error). Also, while we have observed that user feedback was addressed directly to the developer, a more in-depth analysis of users' thoughts and perceptions in the context of human-like error cues seems promising. Furthermore, typing errors are not the only way humans make mistakes. Other error cues connected to human-like design exist and should be investigated, too (e.g., wrong grammar, context-independent emoticons, and incorrect user interpretation).

7. Conclusion

Our research contributes to the field of human-computer interaction by providing knowledge regarding the design of CAs. Specifically, we transferred the social cue of making typing errors from a human-to-human interaction to the context of CAs. Subsequently, our goal was to provide insights into typing errors on perceived humanness, perceived social response, and service quality. Contrary to our hypotheses, we found that social cue typing errors lead to lower perceived humanness and social presence. Thus, our research provides evidence for the challenge of transferring human-to-human communication properties to the context of users of CAs. Specifically, we hypothesize that users attribute errors of CAs (e.g., typing errors) as failures of the development team.

Acknowledgements

We would like to thank Fabian Hildebrandt for his support during this research project.

References

[1] J. Feine, S. Morana, and U. Gnewuch, "Measung Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis," in *Proceedings of the International Conference on Wirtschaftsinformatik*, 2019, pp. 0–11.

[2] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A New Chatbot for Customer Service on Social Media," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2017, pp. 3506–3510.

[3] U. Gnewuch, S. Morana, and A. Maedche, "Towards Designing Cooperative and Social Conversational Agents for Customer Service," in *Proceedings of the International Conference on Information Systems (ICIS)*, 2017, pp. 1–13.

[4] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "A Taxonomy of Social Cues for Conversational Agents," *Int. J. Hum. Comput. Stud.*, vol. 132, no. June, pp. 138–161, 2019, doi: 10.1016/j.ijhcs.2019.07.009.

[5] S. Diederich, A. B. Brendel, and L. M. Kolbe, "Towards a Taxonomy of Platforms for Conversational Agent Design," in *Proceedings of the International Conference on Wirtschaftsinformatik*, 2019.

[6] D. Marinova, K. de Ruyter, M. H. Huang, M. L. Meuter, and G. Challagalla, "Getting Smart: Learning From Technology-Empowered Frontline Interactions," *J. Serv. Res.*, vol. 20, no. 1, pp. 29–42, 2017.

[7] T. Verhagen, J. van Nes, F. Feldberg, and W. van Dolen, "Virtual customer service agents: Using social presence and personalization to shape online service encounters," *J. Comput. Commun.*, vol. 19, no. 3, pp. 529–545, 2014.

[8] M. S. Ben Mimoun, I. Poncin, and M. Garnier, "Case study-Embodied virtual agents: An analysis on reasons for failure," *J. Retail. Consum. Serv.*, vol. 19, no. 6, pp. 605–612, 2012, doi: 10.1016/j.jretconser.2012.07.006.

[9] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *J. Soc. Issues*, vol. 56, no. 1, pp. 81–103, 2000.

[10] K. Hassanein and M. Head, "Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping," *Int. J. Hum. Comput. Stud.*, vol. 65, no. 8, pp. 689–708, 2007.

[11] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions," *Comput. Human Behav.*, vol. 85, pp. 183–189, 2018.

[12] A.-M. Seeger, J. Pfeiffer, and A. Heinzl, "Designing Anthropomorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework," in *Proceedings of the ICIS*, 2018, pp. 1–17.

[13] C. L. Lortie and M. J. Guitton, “Judgment of the humanness of an interlocutor is in the eye of the beholder,” *PLoS One*, vol. 6, no. 9, Sep. 2011.

[14] D. Westerman, A. C. Cross, and P. G. Lindmark, “I Believe in a Thing Called Bot: Perceptions of the Humanness of ‘Chatbots,’” *Commun. Stud.*, vol. 70, no. 3, 2019.

[15] P. F. MacNeilage, “Typing Errors as Clues to Serial Ordering Mechanisms in Language Behaviour,” *Lang. Speech*, vol. 7, no. 3, pp. 144–159, Jul. 1964.

[16] U. Gnewuch, S. Morana, M. T. P. Adam, and A. Maedche, “Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction,” in *Proceedings of the ECIS*, 2018, pp. 1–17.

[17] B. Larivière *et al.*, “‘Service Encounter 2.0’: An investigation into the roles of technology, employees and customers,” *J. Bus. Res.*, vol. 79, pp. 238–246, 2017.

[18] P. B. Brandtzæg and A. Følstad, “Chatbots: Changing User Needs and Motivations,” *Interactions*, 2018.

[19] S. Diederich, M. Janßen-Müller, A. B. Brendel, and S. Morana, “Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent,” in *Proceedings of the International Conference on Information Systems (ICIS)*, 2019, pp. 0–17.

[20] C. Nass, J. Steuer, and E. R. Tauber, “Computers are social actors,” in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1994, p. 204.

[21] A. M. Von Der Pütten, N. C. Krämer, J. Gratch, and S. H. Kang, “‘It doesn’t matter what you are!’ Explaining social effects of agents and avatars,” *Comput. Human Behav.*, 2010

[22] H. Candello, C. Pinhanez, and F. Figueiredo, “Typefaces and the perception of humanness in natural language chatbots,” *Conf. Hum. Factors Comput. Syst. - Proc.*, vol. 2017-May, pp. 3476–3487, 2017

[23] R. M. Schuetzler, G. M. Grimes, J. S. Giboney, and J. Buckman, “Facilitating Natural Conversational Agent Interactions: Lessons from a Deception Experiment,” *International Conference on Information Systems. 2014.*

[24] J. Appel, A. Von Der Pütten, N. C. Krämer, and J. Gratch, “Does humanity matter? Analyzing the importance of social cues and perceived agency of a computer system for the emergence of social reactions during human-computer interaction,” *Adv. Human-Computer Interact.*, p. 10, 2012.

[25] M. Mori, K. F. MacDorman, and N. Kageki, “The Uncanny Valley,” *IEEE Robot. Autom. Mag.* 2012.

[26] R. M. Stock and M. Merkle, “Customer Responses to Robotic Innovative Behavior Cues During the Service Encounter,” in *Proceedings of the International Conference on Information Systems (ICIS)*, 2018, pp. 1–17.

[27] J. Rasmussen, “Human error and the problem of causality in analysis of accidents.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 327, no. 1241, pp. 449–462, 1990..

[28] D. Gefen and D. Straub, “Gender Differences in the Perception and Use of E-Mail: An Extension to the Technology Acceptance Model,” *Manag. Inf. Syst. Q.*, vol. 21, no. 4, pp. 389–400, 1997.

[29] D. C. Toader *et al.*, “The effect of social presence and chatbot errors on trust,” *Sustain.*, 2020

[30] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, “To err is robot: How humans assess and act toward an erroneous social robot,” *Front. Robot. AI*, 2017, doi: 10.3389/frobt.2017.00021.

[31] T. M. Holtgraves, S. J. Ross, C. R. Weywadt, and T. L. Han, “Perceiving artificial social agents,” *Comput. Human Behav.*, vol. 23, no. 5, pp. 2163–2174, 2007.

[32] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, “SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality,” *J. Retail.*, vol. 64, no. 1, pp. 12–40, 1988.

[33] J. J. Jiang, G. Klein, and C. L. Carr, “Measuring Information System Service Quality: SERVQUAL from the Other Side,” *MIS Q.*, vol. 26, no. 2, pp. 145–166, May 2002, doi: 10.2307/4132324.

[34] A. Yan, S. Solomon, D. Mirchandani, M. Lacity, and J. Porra, “The role of service agent, service quality, and user satisfaction in self-service technology,” *ICIS*, pp. 1–21, 2013.

[35] A. R. Dennis and J. S. Valacich, “Conducting Experimental Research in Information Systems,” *Commun. Assoc. Inf. Syst.*, vol. 7, no. 5, pp. 1–41, 2001.

[36] C. Bartneck, T. Kanda, H. Ishiguro, and N. Hagita, “Is the uncanny valley an uncanny cliff?,” in *RO-MAN 2007-The 16th IEEE international symposium on robot and human interactive communication*, 2007, pp. 368–373.

[37] N. Urbach and F. Ahlemann, “Structural Equation Modeling in Information Systems Research Using Partial Least Squares,” *J. Inf. Technol. Theory Appl.*, 2010.

[38] C. DiStefano, M. Zhu, and D. Mîndrilă, “Understanding and using factor scores: Considerations for the applied researcher,” *Pract. Assessment, Res. Eval.*, vol. 14, no. 20, 2009.