AIS Transactions on Human-Computer Interaction

Volume 13
Issue 1 *Design Science Research in Human-Computer Interaction*

Article 5

3-31-2021

Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment

Stephan Diederich
University of Göttingen, stephan.diederich@stud.uni-goettingen.de

Tim-Benjamin Lembcke
University of Göttingen, tim-benjamin.lembcke@uni-goettingen.de

Alfred Benedikt Brendel
University of Goettingen and Technische Universität Dresden, Alfred_benedikt.brendel@tu-dresden.de

Lutz M. Kolbe
University of Göttingen, lkolbe@uni-goettingen.de

Recommended Citation

Diederich, S., Lembcke, T., Brendel, A. B., & Kolbe, L. M. (2021). Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment. *AIS Transactions on Human-Computer Interaction, 13*(1), 82-103. https://doi.org/10.17705/1thci.00143
DOI: 10.17705/1thci.00143

Transactions on Human-Computer Interaction

Volume 13 Issue 1

3-2021

Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment

Stephan Diederich

Chair of Information Management, University of Göttingen, stephan.diederich@stud.uni-goettingen.de

Tim-Benjamin Lembcke

Chair of Information Management, University of Göttingen, tim-benjamin.lembcke@uni-goettingen.de

Alfred Benedikt Brendel

Chair of Business Informatics, esp. Intelligent Systems and Services, TU Dresden, alfred_benedikt.brendel@tu-dresden.de

Lutz M. Kolbe

Chair of Information Management, University of Göttingen, lkolbe@uni-goettingen.de

Recommended Citation

Diederich, S., Lembcke, T.-B., Brendel, A.B., & Kolbe, L. (2021). Understanding the impact that response failure has on how users perceive anthropomorphic conversational service agents: Insights from an online experiment. *AIS Transactions on Human-Computer Interaction*, *13*(1), pp. 82-103.

DOI: 10.17705/1thci.00143

Available at http://aisel.aisnet.org/thci/vol13/iss1/5

Transactions on Human-Computer Interaction

Research Paper DOI: 10.17705/1thci.00143 ISSN: 1944-3900

Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment

Stephan Diederich
Chair of Information Management,
University of Göttingen

Tim-Benjamin Lembcke
Chair of Information Management,
University of Göttingen

Alfred Benedikt Brendel
Chair of Business Informatics, esp. Intelligent
Systems and Services, TU Dresden

Lutz M. Kolbe
Chair of Information Management,
University of Göttingen

Abstract:

Conversational agents (CAs) have attracted the interest from organizations due to their potential to provide automated services and the feeling of humanlike interaction. Emerging studies on CAs have found that humanness has a positive impact on customer perception and explored approaches for their anthropomorphic design, which comprises both their appearance and behavior. While these studies provide valuable knowledge on how to design humanlike CAs, we still do not sufficiently understand this technology’s limited conversational capabilities and their potentially detrimental impact on user perception. These limitations often lead to frustrated users and discontinued CAs in practice. We address this gap by investigating the impact of response failure, which we understand a CA’s inability to provide a meaningful reply, in a service context. To do so, we draw on the computers are social actors paradigm and the theory of the uncanny valley. Via an experiment with 169 participants, we found that 1) response failure harmed the extent to which people perceived CAs as human and increased their feelings of uncanniness, 2) humanness (uncanniness) positively (negatively) influenced familiarity and service satisfaction, and 3) the response failure had a significant negative impact on user perception yet did not lead to a sharp drop as the uncanny valley theory posits. Thus, our study contributes to better explaining the impact that text-based CAs’ failure to respond has on customer perception and satisfaction in a service context in relation to the agents’ design.

Keywords: Conversational Agent, Anthropomorphic Design, Computers Are Social Actors, Theory of the Uncanny Valley.

Stefan Morana was the accepting senior editor for this paper.

1 Introduction

Conversational agents (CAs), technological artifacts with which users interact through natural language (McTear, Callejas, & Griol, 2016), continue to gain interest in research (Maedche et al., 2019) and practice (Oracle, 2016) alike. Praised for their potential to provide a humanlike interaction experience, CAs have seen increasing use in both private and professional life. From a theoretical perspective, such agents represent a particular interesting phenomenon as humans show social responses to these agents (Diederich, Brendel, & Kolbe, 2020; Pfeuffer, Benlian, & Gimpel, 2019). As the computers are social actors (CASA) paradigm posits (Nass & Moon, 2000; Reeves & Nass, 1996), CAs' manifold social cues, such as their interaction via natural language, their name and (humanlike) avatar, and the way they express emotions through verbal and non-verbal communication, trigger social responses and lead users to anthropomorphize CAs (Seeger, Pfeiffer, & Heinzl, 2018). Emerging design-oriented studies on humanlike CAs provide valuable knowledge on the impact that social cues have on humanness, which we understand as the degree to which users attribute actual human properties (e.g., thoughtfulness) to these agents. Moreover, different research suggests anthropomorphism has further effects, such as on service satisfaction (Gnewuch, Morana, Adam, & Maedche, 2018), likability (Bickmore & Picard, 2005), or familiarity (de Visser et al., 2016). Thus, how people perceive anthropomorphic CAs can contribute to relevant context-specific variables. The knowledge base for anthropomorphic CA design offers various social cues that one can incorporate into these artifacts' design to make them appear like humans (Feine, Gnewuch, Morana, & Maedche, 2019).

While these studies provide valuable knowledge for crafting CAs with a humanlike appearance and behavior, the current debate for anthropomorphic design neglects their limited conversational capabilities. As researchers primarily conducted these studies via experimental research (Diederich, Brendel, & Kolbe, 2019a), the CAs provided relevant responses to users' requests. In practice, however, designing agents that continuously offer meaningful responses in an evolving dialogue represents a major challenge (Følstad & Brandtzæg, 2017). In fact, many CA creators discontinued their creations particularly due to their inability to adequately respond to varying user input (Ben Mimoun, Poncin, & Garnier, 2012). As anticipating user requests for natural language software represents a challenging endeavor due to such interactions' unpredictability, situations where a CA needs to provide some kind of fallback response (e.g., "Unfortunately, I did not understand your request. Can you please rephrase?") will likely occur and can remind users that they interact with a machine that has limited capabilities (Ashktorab, Jain, Liao, & Weisz, 2019). Such failure to provide a meaningful reply might prevent humans from perceiving CAs as humanlike and reduce further positive effects, thus, diminish the impact that the social cues in the agent's design have on user perception. In short, the impact that response failure has on how users perceive anthropomorphic CAs represents a substantial practical design problem for which we yet lack a solid understanding.

In our study, we seek to address this research gap by investigating the impact that response failure, which we understand as a CA's inability to meaningfully respond to a valid user's request, has on user perception with the following research question:

> **RQ:** How does failure to provide a meaningful response influence how users perceive anthropomorphic CAs in a service encounter?

Drawing on extant studies on anthropomorphic CA design and on the CASA paradigm and the theory of the uncanny valley as two key theories on how humans perceive and interact with humanlike artifacts, we develop a research model that comprises eight hypotheses and test it in a two by two experiment 169 participants. With this research, we make three main contributions. First, we better explain the influence that agents' response failure due to limited conversational capabilities has on how humans perceive their humanness and uncanniness (i.e., the feeling of strangeness that arises when one feels anthropomorphic artifacts have inhuman qualities). Second, we demonstrate the positive (negative) impact that humanness (uncanniness) has on familiarity (i.e., the degree to which users feel acquainted with agents) and service satisfaction (i.e., users' satisfaction with encountering and interacting with agents). Third, we better explain the magnitude of the effect that modest response failure has on how users perceive agents, which depends in particular on the agent's machine- or humanlike design.

This paper proceeds as follows: in Section 2, we present related work and the theoretical background for our work. In Section 3, we derive eight hypotheses and introduce our research model. In Section 4, we describe how we designed our experiment. In Section 5, we present our results. In Section 6, we discuss our results' implications for designing humanlike CAs, highlight our study's limitations, and discuss directions for future research. Finally, in Section 7, we conclude the paper.

2 Related Work and Theoretical Background

Today, conversational agents, which we define as software with which users interact through natural language (McTear et al., 2016), permeate our private and professional lives (Maedche et al., 2019) in various areas, such as customer service (Diederich, Janßen-Müller, Brendel, & Morana, 2019b; Hu et al., 2018), marketing and sales (Vaccaro, Agarwalla, Shivakumar, & Kumar, 2018), human resources (Liao et al., 2018), finance (Dolata, Kilic, & Schwabe, 2019), and education (Crockett, Latham, & Whitton, 2017).

In addition to these application areas, one can distinguish different forms of CA by their primary communication and embodiment mode. In general, technology interaction through natural language can occur in spoken form, such as with Amazon's Alexa or Apple's Siri, or via written text, such as with chatbots on company websites or social media (Gnewuch, Morana, & Maedchem 2017). Furthermore, CAs can be physically embodied such as service robots (Stock & Merkle, 2018; Stock et al., 2019), have a virtual static avatar (Wünderlich & Paluch, 2017) or virtual interactive avatar (Beer, Smarr, Fisk, & Rogers, 2015), or be disembodied (i.e., without any form of avatar at all) (Araujo, 2018). In this study, we focus on a CA with which users communicate via written text (chatbot) and a static virtual avatar (image) in a customer service context.

2.1 Conversational Service Agents and their Responsiveness

Customer service currently represents a popular application area for CAs in enterprises where such agents can fulfill requests such as handling complaints or providing product information (Gnewuch et al., 2017; Pfeuffer et al., 2019). While current CAs primarily cover rather simple, frequent, and repetitive service requests, they may support or even fully assume increasingly complex tasks that human service personnel currently perform (Verhagen, van Nes, Feldberg, & van Dolen, 2014; Marinova, de Ruyter, Huang, Meuter, & Challagalla, 2017). As technological components of service systems, CAs exist between current service technology that one can always access but lacks the feeling of human interaction (e.g., online portals for self-service) and human-provided services that offer personal contact but have limited availability. In practice, one can find different CAs in a service context across industries (Oracle, 2016). For example, the American railroad company Amtrak uses the virtual agent "Julie" that answers more than five million customer requests per year (NextIT, 2018). Similarly, the clothing brand H&M offers an artificial sales agent that provides individual product recommendations that consumers can directly purchase from the company's online store (Morana, Friemel, Gnewuch, Maedche, & Pfeiffer, 2017). Finally, Facebook deployed more than 100,000 agents in the first year after it opened the Messenger platform (Johnson, 2018).

Despite their popularity and success stories, many CAs have failed to reach the expectations set by their owners(Luger & Sellen, 2016), who subsequently discontinued them due to flaws related to their design (Ben Mimoun et al., 2012). In assessing 80 conversational agents on French commercial websites, Ben Mimoun et al. (2012) identified inadequate appearance, insufficient interactivity, and intelligence as reasons for CA failure. They argued that a mismatch between CAs' humanlike appearance and their actual service possibilities/competence leads to negative customer reactions due to unfulfilled, high expectations. Likewise, Luger and Sellen (2016) found "user expectations dramatically out of step with the operation of the systems, particular in terms of known machine intelligence, system capabilities and goals" (p. 5286). Hence, both studies suggested a gap between user expectations and technical capabilities as a main reason for why users negatively perceive CAs and suggested different design approaches to manage user expectations more adequately.

Against this background, Følstad and Brandtzæg (2017) emphasized that a natural language interface resembles a blank canvas that mostly hides a system's capabilities from users and that designers need to anticipate a much larger variety of input compared to graphical user interfaces. Consequently, they argued that fallbacks in a conversation will likely occur (Følstad & Brandtzæg, 2017). Similarly, Go and Sundar (2019) highlighted that equipping a CA with the ability to provide meaningful responses in a way that depends on previous content in a conversationrepresents a substantial design issue.

Overall, we observe that CAs have grown in popularity in a service context. However, we also note that the fact they lack sufficient conversational capabilities to respond to highly varying user input represents a major design challenge and leaves room for response failures in the interaction.

2.2 CASA Paradigm and the Theory of Uncanny Valley in the Context of CAs

A key paradigm that guides efforts to design and interact with CAs focuses on understanding such artifacts as social actors (Nass & Moon, 2000; Reeves & Nass, 1996). The CASA paradigm posits that humans apply social and expectations to technology that exhibits traits or behavior usually associated with humans (Nass & Moon, 2000). In various experiments, Nass and Moon (2000) discovered that humans overuse social categories, such as gender, and show social behaviors, such as reciprocity, in an interaction with a humanlike artifact. Furthermore, the researchers found that humans show premature cognitive commitments to computers, such as when they label artifacts with a specific social role (e.g., member of the customer service team). According to the researchers, the more human characteristics a technological artifact features, the stronger it leads to social responses (Nass and Moon, 2000). As CAs typically exhibit various social cues (Feine et al., 2019) that range from basic cues, such as the interaction via natural language and turn-taking in a conversation, to more complex ones, such as understanding and expressing emotions, they trigger humans to make substantial social responses. As Seeger et al. (2018) suggested, CAs' anthropomorphic design comprises social cues in three different dimensions: 1) human identity (e.g., age, gender, ethnicity), 2) verbal cues (e.g., syntax and word variability or self-references in a conversation), and 3) non-verbal cues (e.g., response delays to indicate thinking or emoticons to express emotions). In short, designers have various social cues at their disposal to make CAs seem as similar to humans as possible (Feine et al., 2019).

Figure 1. CASA Paradigm in the Context of Conversational Agents

Recent experiments on how people perceive anthropomorphic CAs have discovered that a humanlike design has mostly positive effects on aspects such as social presence (Pereira, Prada, & Paiva, 2014), trustworthiness (Benlian, Klumpe, & Hinz, 2019), persuasiveness (Diederich, Lichtenberg, Brendel, & Trang, 2019c; Hanus & Fox, 2015), enjoyment (Liao et al., 2018), willingness to pay (Yuan & Dennis, 2019), and service satisfaction (Gnewuch et al., 2018). However, some studies have found unintended adverse effects. For example, Wünderlich and Paluch (2017) described a risk of perceived uncertainty as whether users interact with a machine or an actual human. Furthermore, Sohn (2019) discovered increased privacy concerns due to the mere presence of an anthropomorphic agent on an e-commerce website. In addition, Seeger et al. (2018) assumed that a CA that included social cues from all three aforementioned dimensions had a negative effect on whether people perceive CAs as anthropomorphic . Thus, they suggested ways to find an appealing combination of social cues rather than follow a "more is more" approach.

In this context, Mori (1970) and Mori et al. (2012) hypothesized about the relationship between humanlike objects and affinity (or familiarity depending on who translates the original Japanese) in the theory of the uncanny valley. The theory, originally from the robotics field, posits that a nonlinear relationship between the degree to which an object resembles a human and humans' emotional responses to the object exists. The "valley" refers to a sharp drop in affinity or familiarity for the object before it fully resembles a human. MacDorman, Green, Ho, and Koch (2009) described the valley as a shift in human attention from an object's anthropomorphic qualities to the aspects that seem inhuman by stating that "as something looks more human it looks also more agreeable, until it comes to look so human that we start to find its nonhuman imperfections unsettling" (p. 2). This negative reaction comprises strong feelings of uncanniness due to an object's nonhuman imperfections (MacDorman et al., 2009).

Researchers have conducted various empirical studies to investigate the propositions that the theory of the uncanny valley makes. Most studies have found that people react to different stimuli such as morphed images, human images, or interactive agents in line with Mori's propositions or at least found a significant effect related to uncanniness (e.g., Bartneck, Kulić, Croft, & Zoghbi, 2009; Riek, Rabinowitch, Chakrabartiz, & Robinson, 2009; Tinwell, Grimshaw, Nabi, & Williams, 2011; Mathur & Reichling 2016). However, Hanson (2006) and Hanson, Olney, Pereira, and Zielke (2005) did not find empirical support for the uncanny valley, which led researchers to propose that aesthetics represent a key factor that influences how people perceive anthropomorphic artifacts beside the extent to which it resembles humans. In short, many studies have found empirical evidence that a high degree of anthropomorphism related to the uncanny valley has detrimental effects on user perception; however, further aspects might also influence such effects to varying degrees. Figure 2 depicts the uncanny valley as Mori et al. (2012) conceptualized it.

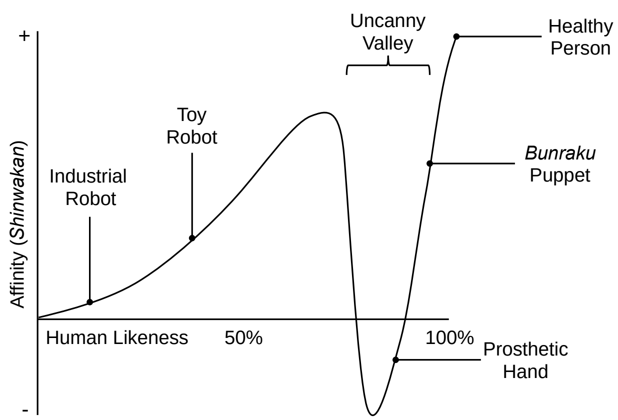

Figure 2. The Uncanny Valley (Based on Mori et al., 2012)

3 Hypotheses and Research Model

In this study, we examine the impact that a humanlike conversational agent's failure to respond in a natural language dialogue has on user perception. To do so, we propose a model with eight hypotheses.

As per the CASA paradigm (Nass & Moon, 2000; Reeves & Nass, 1996), a technological artifact that displays a humanlike appearance and behavior triggers social responses in humans. As Nass and Moon (2000) argued, the more technological artifacts, such as a computer, exhibit humanlike characteristics, the more likely they will be to trigger social reactions. In the CA context, emerging studies on anthropomorphic design indicate that social cues lead people to perceive them as human in the interaction (Feine et al., 2019; Wünderlich & Paluch, 2017). In line with these studies and the CASA paradigm, we hypothesize:

H1: Social cues have a positive impact on an agent's humanness.

Second, we propose that an appealing combination of different social cues that comprise an agent's human identity and verbal and non-verbal cues (Seeger et al., 2018) reduce feelings of uncanniness when interacting with the agent. Different studies on anthropomorphic design indicate that social cues have positive effects such as on likability (Bickmore & Picard, 2005; Cowell & Stanney, 2005), trust (de Visser et al., 2016; Nunamaker, Derrick, Elkins, Burgoon, & Patton, 2011), and enjoyment (Qiu & Benbasat, 2010). Thus, we suggest that social cues reduce feelings of uncanniness when interacting with an anthropomorphic agent and hypothesize:

H2: Social cues have a negative impact on an agent's uncanniness.

Against the background of the theory of the uncanny valley, Mori et al. (2012) hypothesized that a "person's response to a humanlike robot would abruptly shift from empathy to revulsion as it approached, but failed to attain, a lifelike appearance" (p. 98). While a CA may not be able to sustain a humanlike appearance in a dialogue for many reasons (e.g., due to imperfectly representing an interactive, humanlike avatar) (Seymour, Riemer, & Kay, 2018), we argue that an inability to provide a meaningful response due to user input's complexity and unpredictability in a natural language interaction constitutes one of the most likely reasons (Følstad & Brandtzæg, 2017). As a result, this failure to respond would abruptly shift users' attention to the fact that they are not interacting with an actual human and, thereby, decrease whether they perceive the CA as human. Thus, we hypothesize:

H3: Response failure has a negative impact on an agent's humanness.

Similarly, we expect an agent's inability to provide a meaningful response to induce feelings of uncanniness (Tinwell & Sloan, 2014) as it constitutes a strange situation that does not conform with users' expectations about a humanlike conversation (Luger & Sellen, 2016). Thus, we hypothesize:

H4: Response failure has a positive impact on an agent's uncanniness.

Furthermore, anthropomorphized artifacts an ability to induce feelings of familiarity (Epley, Waytz, & Cacioppo, 2007) because social cues make it easier for users to connect with them, potentially even on a personal level (Burgoon et al., 2000), and feel at ease with their form and function (Duffy, 2003). The theory of the uncanny valley further conceptualizes this relation (Mori, 1970; Mori et al., 2012) in positing that familiarity increases steadily until users reach the valley. Thus, we suggest perceiving humanness in an artifact positively impacts familiarity and hypothesize:

H5: Humanness has a positive effect on an agent's familiarity.

Next, we consider the relation between uncanniness and familiarity. Based on a similar reasoning for H5, we argue that feelings of uncanniness in the interaction harm familiarity in line with the theory of the uncanny valley (Mori, 1970; Mori et al., 2012). As uncanniness, manifested in feelings of for example strangeness or eeriness during the interaction with an anthropomorphic agent, contributes to a negative user perception (Tinwell & Sloan, 2014), it diminishes the perception of the agent as familiar:

H6: Uncanniness has a negative impact on an agent's familiarity.

Finally, different studies on CAs in a service context have argued that both perceiving humanness and related social responses (as understanding CAs as social actors suggests) (Nass & Moon, 2000) have a positive relationship with service satisfaction, which has particular importance in online service encounters. In this context, Gnewuch et al. (2018) found that dynamic response delays to indicate that an agent thinks and types before responding (Gnewuch et al., 2018) led to an increased feeling of humanness and service satisfaction. Similarly, Hu et al. (2018) suggested that a sentiment-adaptive design has a positive impact on the perceived empathy of CAs in interactive service encounters. Thus, we hypothesize that humanness in a service encounter positively impacts satisfaction and, similarly, that negative feelings of uncanniness negatively impact service satisfaction:

H7: Humanness has a positive impact on service satisfaction.

H8: Uncanniness has a negative impact on service satisfaction.

Figure 3 summarizes our eight hypotheses and visualizes the research model for our study.

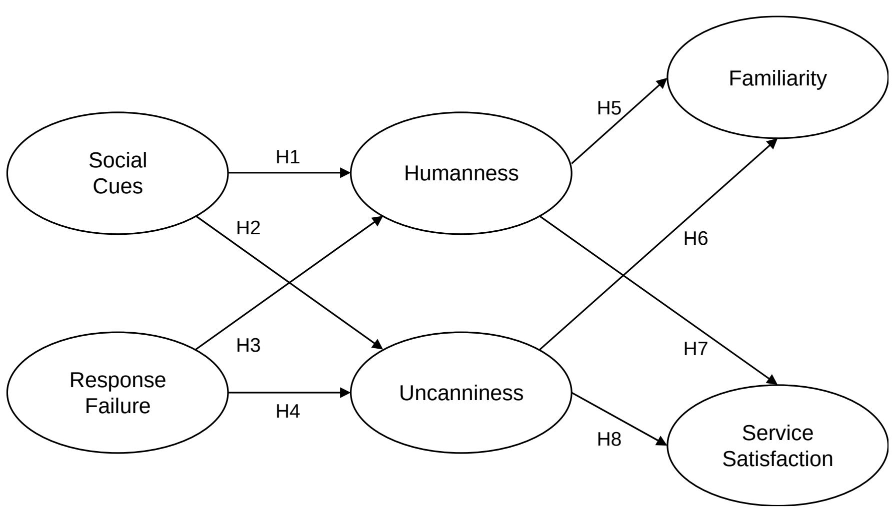

Figure 3. Research Model

4 Research Design

We tested our hypotheses regarding anthropomorphic design and response failure in an online experiment with a text-based CA. To provide a familiar and understandable context and task, we selected a customer service setting with an online retailer. In this section, we describe the data-collection procedure and sample, the four experimental conditions, the manipulation check, and the measures we used in the post-experimental survey.

4.1 Data Collection Procedure and Sample

We asked our participants to interact with a fictitious online retailer's virtual service agent to track and cancel an existing order and to ask for confirmation that it had done so. Every participant received a link to a briefing website in which we described the experiment context (online retailer), the structure (interaction with a virtual customer service agent followed by a questionnaire), and the participants' tasks. The tasks comprised contacting the service agent and finding out the current order status for a given identification number, authenticating with the agent, asking for order cancellation, and requesting a confirmation via email. Similar to recent studies on CAs (e.g., Cho, 2019; Gnewuch et al., 2018), we selected a rather specific set of tasks to enable a structured, comparable dialogue across the conditions and contribute to the agent's responsiveness in the interaction. After participants successfully completed the last task, the CA provided a link to the questionnaire. Overall, the participants took around nine minutes each to complete the experiment. In total, 169 people successfully completed the experiment. They ranged in age from 19 to 59 years (mean: 27.8 years), and 40.6 percent were female. We removed four participants who provided straight-line answers, which decreased the final sample size to 165 participants. We did not provide monetary compensation for participating the experiment. We recruited the participants were from our personal networks and comprised mainly students from a German university.

4.2 Experimental Conditions

For the experimental conditions, we designed four CA instances using Google's natural language platform Dialogflow (see https://dialogflow.com). Dialogflow provides the technical capabilities to detect a user's intent from a natural language statement and formulate a response. All instances received the same training phrases. We varied the design and responsiveness as we visualize in Table 1.

Table 1. Experimental Conditions

		Social cues	
		Few	**Many**
Response failure	With	**Condition 1** Agent with few social cues and with response failure	**Condition 2** Agent with many social cues and with response failure
	Without	**Condition 3** Agent with few social cues and without response failure	**Condition 4** Agent with many social cues and without response failure

The two instances with a humanlike design received social cues to make the agents appear humanlike (see Figure 4). Based on the three anthropomorphic design dimensions that Seeger et al. (2018) proposed, we provided the agent with a comic-like avatar of a female customer service employee (Gong, 2008), a human name (Cowell & Stanney, 2005), and a gender (Nunamaker et al., 2011) to establish a human identity.

We further integrated self-references (Sah & Peng, 2015), self-disclosure (Schuetzler, Grimes, & Giboney, 2018), a personal introduction and greeting (Cafaro, Vilhjalmsson, & Bickmore, 2016), and variability in syntax and word choice for the agent's responses (Seeger et al., 2018) in terms of verbal cues. With regard to non-verbal cues, we added dynamic response delays to indicate thinking and typing (Gnewuch et al., 2018) in combination with blinking dots (de Visser et al., 2016) and emoticons to express emotions (Wang et al., 2008).

With regard to the second dimension, response failure, we designed the agent in the first and second conditions to indicate that it did not understand at one point in the interaction. When the participants in those conditions requested to cancel the given order, the agent politely responded that it did not understand the user's input and asked the user to reformulate the request two times. After the participant entered the request for order cancellation a third time, the agent provided a meaningful response and confirmed the cancellation. Table 1 shows the agent's responses in the two conditions with low responsiveness.

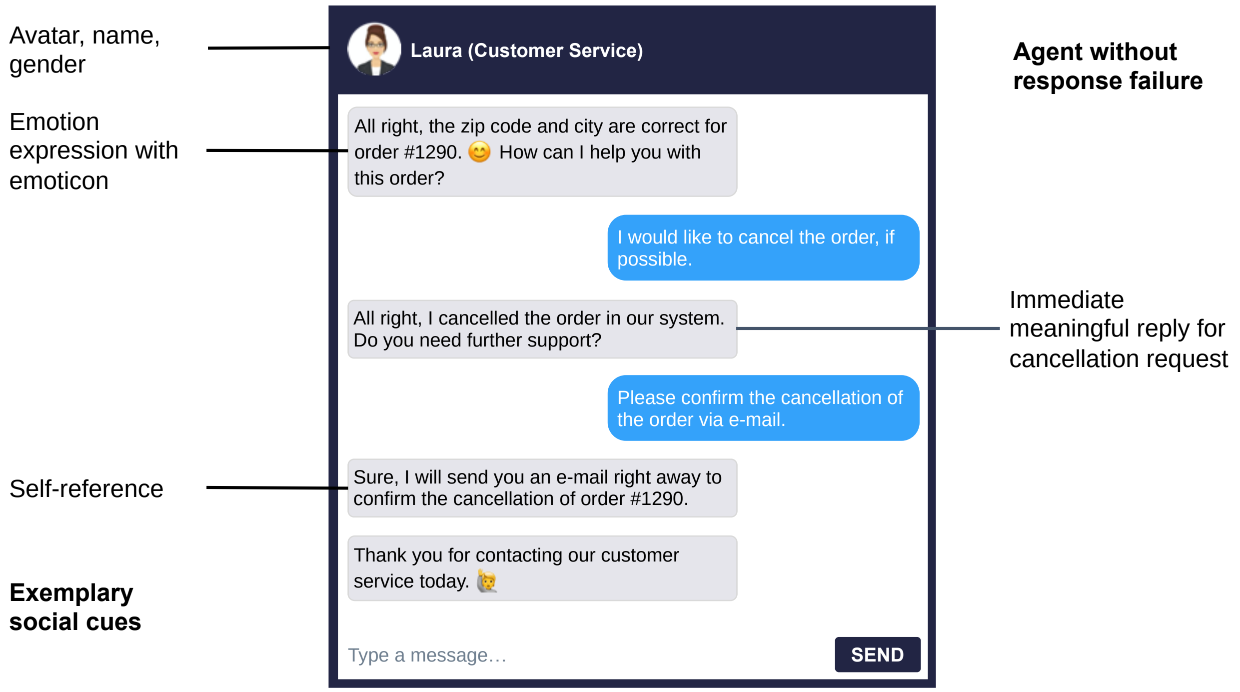

Figure 4. CA with Humanlike Design and Without Response Failure (Condition 4)

Table 2. Agent Statements with Response Failure (Translated to English)

Iteration	Condition 1 (few social cues)	Condition 2 (many social cues)
1	"Unfortunately, I do not understand your request."	"Unfortunately, I do not understand your request. Can you please reformulate it?"
2	"Unfortunately, I do not understand your request."	"I am so sorry, but I do not understand what you are saying. Can you please formulate it differently?"
3	"Your order is now cancelled."	"All right, I cancelled the order in our system. Do you need further support?

4.3 Manipulation Check

To check whether we successfully manipulated the responsiveness dimension in the sense that participants received only meaningless responses as intended without further fallback replies, we analyzed the conversation data that Google Dialogflow provided. Reviewing the interactions with the agents in the four conditions showed that, in most cases, the agents demonstrated the communication behavior as intended (i.e., on average, they made 1.5 fallback responses overall per interaction).

All conditions showed similar average fallbacks between 1.7 and 1.4 messages that occurred when the agent did not understand a user's intent. Thus, only a minimal number of fallbacks, similar across all groups, existed in the interactions.

4.4 Measures

We asked every participant to complete a survey to measure their perceptions of humanness, uncanniness, familiarity, and service satisfaction. We used established measurement instruments for the four constructs. Table 2 shows the constructs, items, factor loadings, Cronbach's α, composite reliability (CR), and average variance extracted (AVE).

Table 3. Constructs, Items, and Factor Loadings

Constructs and items	Loadings	Scale and source
Humanness (α = 0.904, CR = 0.927, AVE = 0.680)		Nine-point semantic differential scale (Holtgraves & Han, 2007)
Extremely inhumanlike – extremely humanlike	0.887	
Extremely unskilled – extremely skilled	0.882	
Extremely unthoughtful – extremely thoughtful	0.853	
Extremely impolite – extremely polite	0.671	
Extremely unresponsive – extremely responsive	0.821	
Extremely unengaging – extremely engaging	0.816	
Uncanniness (α = 0.911, CR = 0.932, AVE = 0.698)		Seven-point Likert scale (MacDorman et al., 2009; Tinwell and Sloan, 2014)
I perceived the agent as eerie.	0.665	
I perceived the agent as inhumanlike.	0.784	
I perceived the agent as strange.	0.880	
I perceived the agent as unappealing.	0.909	
I perceived the agent as inclement.	0.853	
I perceived the agent as unpleasant.	0.895	
Familiarity		Nine-point semantic differential scale (MacDorman, 2006)
Extremely strange – extremely familiar	n/a	
Service satisfaction (α = 0.888, CR = 0.931, AVE = 0.819)		Seven-point Likert scale (Verhagen et al. 2014)
How satisfied are you with the agent's advice?	0.914	
How satisfied are you with the way the agent treated you?	0.854	
How satisfied are you with the overall interaction with the agent?	0.944	

We measured both humanness and familiarity on a nine-point semantic differential scale with items from Holtgraves and Han (2007) and MacDorman (2006). To measure feelings of uncanniness, we adapted a seven-point Likert scale from MacDorman et al. (2009) and Tinwell and Sloan (2014). Similarly, we measured service satisfaction on a seven-point Likert scale using items from Verhagen et al. (2014). Furthermore, we collected demographic information (age, gender) and information on the frequency with which participants used digital assistants (e.g., Siri, Alexa, and chatbots). Finally, we asked for free form feedback on how participants perceived the agent. We added attention checks by inverting two items in the survey.

Following Gefen and Straub's (2005) suggestions, we used items with loadings larger than 0.60 in the analysis. Humanness, uncanniness, and service satisfaction showed sufficient values for CR (larger than 0.80), Cronbach's alpha (larger than .80) and AVE (larger than 0.50) considering the levels by Urbach and Ahlemann (2010). We include a table with the cross loadings in the appendix (see Table 8). One can see that the constructs loaded highest on the constructs we intended them to measure. Thus, we found support for discriminant validity.

5 Results

We analyzed the experimental data via partial least squares structural equation modeling (PLS-SEM). In this section, we present our results, assess direct effect sizes, and exploratively compare means for the four constructs depending on the agent's machine- or humanlike design.

5.1 Hypotheses and Research Model

We tested our hypotheses using partial least squares (PLS) and SmartPLS (version 3). We calculated the significance of the path coefficients with a bootstrapping resampling approach with 5,000 samples (Chin, 1998). We show the resulting path coefficients, R^2 values for the dependent variables, and significance levels in Figure 5.

The paths between social cues and humanness and between social cues and uncanniness show significant relationships. In line with the CASA paradigm, we found evidence that social cues positively impact a CA's humanness (social cues → humanness, β = 0.462, p ≤ .001), which supports H1.

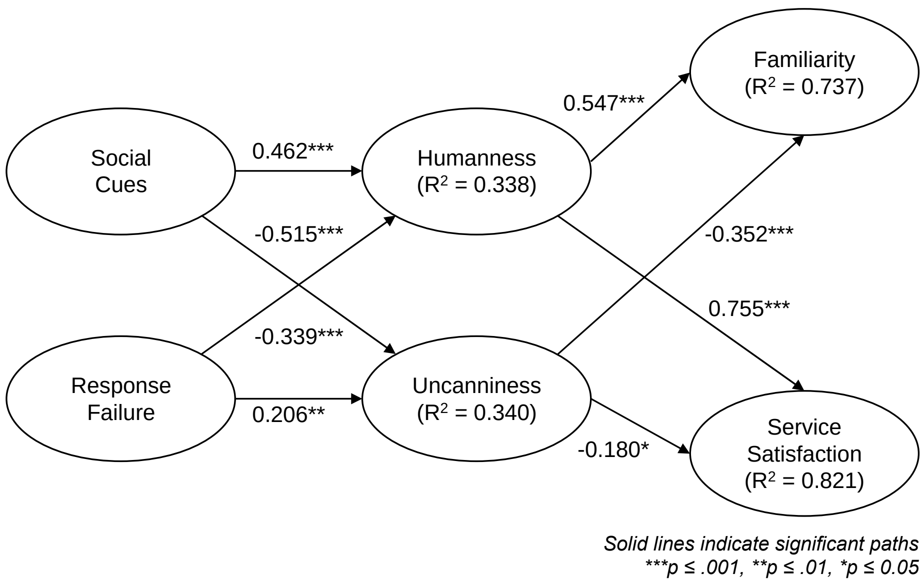

Figure 5. PLS Structural Model (N = 165)

Furthermore, we found that evidence that a humanlike design with social cues has a negative impact on uncanniness, which supports H2 (social cues → uncanniness, β = -0.515, p ≤ 0.001). With regard to response failure, we observed a negative impact on humanness (response failure → humanness, β = -0.339, p ≤ 0.001) and a positive impact on uncanniness (response failure → uncanniness, β = 0.206, p = 0.006), which supports H3 and H4.

We further found evidence that humanness has a positive (humanness → familiarity, β = 0.547, p ≤ 0.001) and uncanniness has a negative (uncanniness → familiarity, β = -0.352, p ≤ 0.001) impact on familiarity with an agent, which supports H5 and H6, respectively. Finally, we found evidence that humanness positively contributes to satisfaction (humanness → service satisfaction, β = 0.755, p ≤ 0.001) and that uncanniness has a detrimental influence on service satisfaction (uncanniness → service satisfaction, β = -0.180, p = 0.006), which supports H7 and H8, respectively.

5.2 Supplementary Analyses

To complement our model results, we analyzed the effect that the control variables (age, gender, prior experience with digital assistants and chatbots) had on our results. However, we found that only participants' experience with assistants had a significant effect on service satisfaction (experience with Assistants → service satisfaction, β = 0.087, p = 0.029). Furthermore, we conducted an analysis to examine moderating effects. We found that participants' experience with assistants to moderated the relationship between uncanniness and service satisfaction (β = 0.152, p = 0.014) and between humanness and service satisfaction (β = 0.123, p = 0.046). Thus, our results suggest that individual experience with CAs in a customer service context influences participant satisfaction with the service encounter and the impact of humanness and uncanniness in the interaction.

Furthermore, we assessed the size of the direct effects with f^2 values. Using the levels that Cohen (1988) provided, we interpreted the values 0.02, 0.15, and 0.35 as small, medium, and large, respectively. All significant relationships had effect sizes that exceeded the small effect threshold.

Table 4. Effect Sizes for Significant Paths according to Cohen (1988)

Small effect ($f^2 \geq 0.02$)	**Medium effect** ($f^2 \geq 0.15$)	**Large effect** ($f^2 \geq 0.35$)
Response failure → uncanniness Uncanniness → Service satisfaction	Response failure → humanness Uncanniness → familiarity Social Cues → humanness	Humanness → familiarity Humanness → service satisfaction Social Cues → uncanniness

The paths from response failure to uncanniness (f^2 = 0.054) and from uncanniness to service satisfaction (f^2 = 0.036) exhibited small effect sizes. The paths from social cues to humanness (f^2 = 0.0313), from response failure to humanness (f^2 = 0.175), and from uncanniness to familiarity (f^2 = 0.155) exhibited medium effect sizes. The paths from social cues to uncanniness (f^2 = 0.407), from humanness to familiarity (f^2 = 0.453), and from humanness to service satisfaction (f^2 = 1.24) exhibited large effect sizes.

Finally, we indicatively compared the differences in means for the latent variables to investigate whether we could observe a strong negative adverse response as the theory of the uncanny valley postulates (see Figure 6).

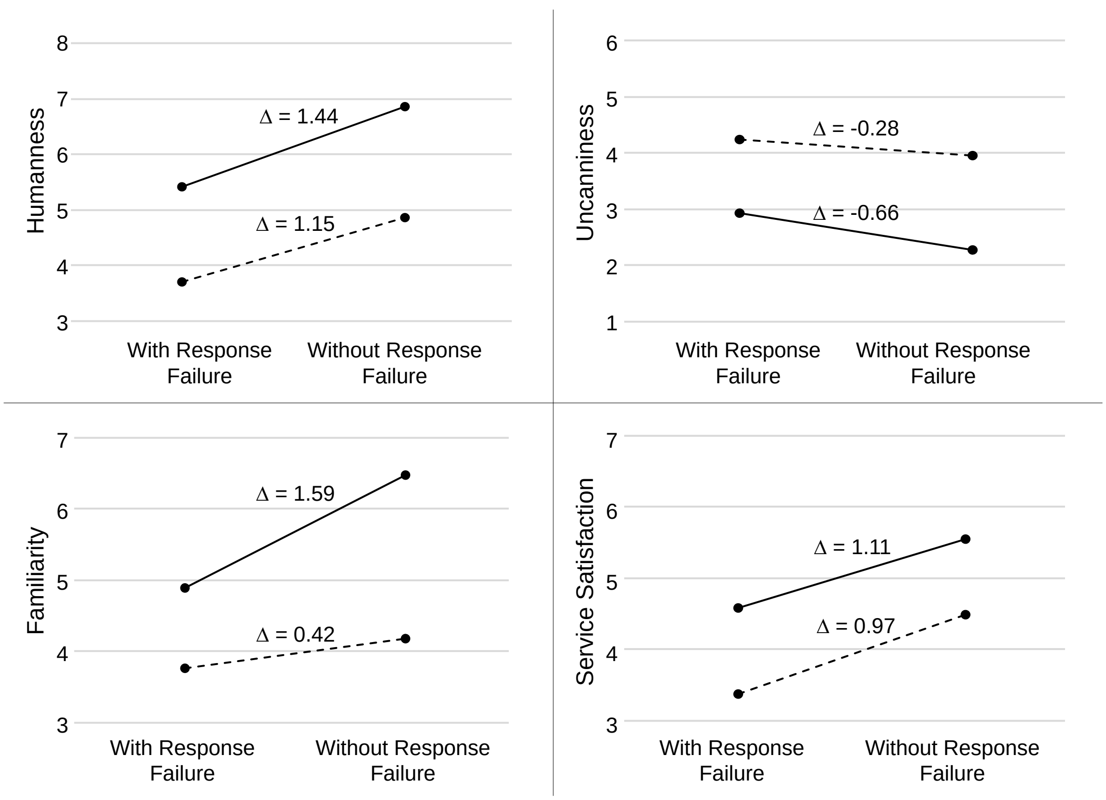

Figure 6. Mean Values differentiated by Social Cues (Dotted Lines Indicate Conditions with Few Cues)

Our dataset indicates that participants who received a design with many social cues showed substantially stronger negative reactions (Δ = 1.59) to response failure with regard to familiarity than participants who interacted with a machine-like CA with few cues (Δ = 0.42). However, the differences between humanness (Δ = 1.44, Δ = 1.15), uncanniness (Δ = -0.66, Δ = -0.28) service satisfaction (Δ = 1.11, Δ = 1.97) were comparatively smaller.

6 Discussion

We found empirical evidence that CA response failure has a negative impact on how people perceive a CA and a positive impact on an agent's unintended uncanniness in a service context. Furthermore, we found that humanness has a positive impact on familiarity and satisfaction and that uncanniness has a detrimental impact on these variables in a service context. In this section, we discuss our results' implications for research on anthropomorphic CA design and CA design in practice, indicate limitations, and suggest opportunities for future research.

6.1 Implications for Research on Anthropomorphic CAs

We found that CA response failure has a substantial negative impact on user perception in a service encounter context. As such, our results concur with recent research that has highlighted the need for CAs to have sufficient conversational capabilities (Følstad & Brandtzæg, 2017; Gnewuch et al., 2017; Schuetzler et al., 2018). Participants who interacted with a CA that demonstrated response failure indicated they perceived it to have less humanness, familiarity, and service satisfaction and more uncanniness. Even if one considered response failure in the first two conditions rather modest (i.e., the agent failed to provide a meaningful reply two times) and the CA in all cases ultimately completed the user's service request, the response failure had a substantial negative impact on user perception. The qualitative, free-form feedback on the CA's design supports this effect. For example, participants who interacted with a CA with response failure stated that they perceived it as "incomplete" or criticized that they had to "ask for order cancellation a thousand times". Furthermore, one participant commented that "What is the purpose of the nice design if the computer does not understand me?". Thus, the participants immediately recognized the agent's failure to respond meaningfully in the respective experimental conditions, and it did not conform with their expectations about essential conversational capabilities.

Considering the results from analyzing direct effect sizes (see Table 4), even a rather modest failure to respond in the conversation led to a medium-sized detrimental effect on the extent to which participants viewed the agent as humanlike. Interestingly, the large variety of social cues incorporated in the humanlike agent's design (second and fourth conditions) exhibited a comparable effect size on humanness. Our experimental data indicates that even a small response failure led to a substantial negative effect on humanness with an effect size comparable to the impact that rich social cues had on the agent's humanness. Therefore, our experiment emphasizes the importance of managing and matching user expectations when designing (anthropomorphic) CAs as Luger and Sellen (2016) or Følstad and Brandtzæg (2017) have suggested.

With regard to a potential uncanny valley effect (see Figure 2), we did not find a sharp drop in familiarity or increase in uncanniness as the theory suggests. While we observed that response failure had a substantial negative impact on user perception, the mean values for familiarity (uncanniness) were still higher (lower) in the conditions with a humanlike design with many social cues than in those with a machine-like design with few social cues (see Figure 6). Interestingly, however, the difference in familiarity depending on response failure seemed to be larger for the humanlike CAs ($\Delta = 1.59$) than the difference for the CAs with a machine-like design ($\Delta = 0.42$). In our view, we can explain this observation in three possible ways. First, our agent's anthropomorphic design with different social cues may achieve a level of humanlikeness close to the beginning of the uncanny valley but not reach it. If true, further increasing the agent's anthropomorphism would mean we could observe effects related to the uncanny valley due to its limited conversational capabilities. Second, a differently structured curve could describe the relationship between familiarity and humanlikeness for CAs as, for example, MacDorman (2006) reported in analyzing human reactions to robot video clips. Third, different aspects of the agent's design that do not directly relate to the degree of anthropomorphism, such as aesthetics, might alleviate the detrimental effect of the modest response failure as, for example, Hanson (2006) and Hanson et al. (2005) have suggested.

6.2 Implications for CA Design in Practice

We draw three practical implications for designing anthropomorphic CAs, particularly in a service context, from our results. First, a rich combination of social cues enables people to perceive a conversational service agent as human, which, in turn, positively contributes to familiarity and, in particular, service satisfaction. These results concur with Gnewuch et al.'s (2018) and Araujo's (2018) results. Thus, according to our data, designers should generally seek to craft CAs with an appealing humanlike representation and behavior in a service context.

Second, designers should consider equipping a CA with sufficient conversational capabilities to mitigate and adequately handle response failures when designing it due to the substantial negative impact that arises from even modest response failures. While designers admittedly face a challenging task in anticipating user input (Følstad & Brandtzæg, 2017), they need to treat conversations as the core object in a CA's design to build agents that fulfill user expectations and can maintain humanlike behavior (Clark et al., 2019). Furthermore, designers should carefully reflect on and select coping strategies to handle unanticipated situations in a conversation (i.e., when agents fail to successfully detect a user's intent). For example, designers could design CAs to provide users with different options for interpretation from which they could choose the intended request or the agents could highlight the words they cannot process and ask users to reformulate or clarify their input. In this context, Ashktorab et al. (2019) have summarized and evaluated different techniques to address response failure in dialogues.

Third, since we did not find very strong negative effects as the theory of the uncanny valley posits, designers should favor a humanlike design even if small response failure may occur. As we indicate in our exploratory comparison (see Figure 6), people still perceive an agent with many social cues and modest response failure as more humanlike, less uncanny, and more familiar and have a higher average service encounter satisfaction. Thus, according to our experimental data, equipping CAs with appealing social cues seems to be advantageous despite the detrimental impact of modest response failure due to the limited conversational capabilities in contemporary natural language technology.

6.3 Limitations and Opportunities for Future Research

Our research exhibits different limitations and offers opportunities for future research on anthropomorphic CA design. Our experimental setting meant we gained control and precision to test our hypotheses. However, the experiment had limited realism, and one needs to take care when generalizing the results (Dennis & Valacich, 2001; Karahanna, Benbasat, Bapna, & Rai, 2018).

Similar to other studies on CA design, we provided the participants with rather specific tasks. Hence, we could create a setting in which the agent consistently failed to provide a meaningful response around two times, which allowed us to better understand the impact on user perception of the agent by comparing the experimental groups. However, in a practical interaction with a CA, response failures will likely occur with different frequency depending on the agent's design.

In addition, other reasons for response failure exist, such as spelling errors or out-of-context questions, which we did not have the scope to consider in our experiment yet represent a worthwhile opportunity for future research. Further, we measured familiarity based on a single item as other researchers have done (e.g., MacDorman (2006). Moreover, we conducted our experiment in a specific context (customer service) with users expecting the agent to be able to fulfill their rather trivial service request (order cancellation). Therefore, we suggest that future studies explore CA response failure in different, potentially more complex (service) contexts and, in particular, in situations where the agent cannot ultimately fulfill a customer's request. In particular, varying the service context and investigating whether customer reactions might diverge with regard to how they perceive an agent and their satisfaction with the service encounter represents a worthwhile research opportunity. Furthermore, future studies can investigate how an agent's design influences users' expectations with regard to anthropomorphism and response failure in different service contexts and potentially over multiple interactions by drawing on, for example, expectation confirmation theory (Oliver, 1977, 1980) in order to advance our understanding regarding individual users and their expectations, CA design, and how people perceive agents, and service satisfaction.

Finally, our sample comprised mainly (IS) students from a German university and, thus, likely shared a similar cultural background and affinity toward technology. While these subjects constitute part of the population we sought to generalize to, we could better understand the impact of anthropomorphic CA response failure from research that replicates our experiment with samples that have different compositions, both with regard to demographics as well as affinity toward anthropomorphic agents. Furthermore, our findings concerning the rather strong detrimental impact of response failures in a conversation offer two directions for future studies on anthropomorphic CAs.

First, researchers could conduct design-oriented research to investigate how one could mitigate response failure, such as by providing a more transparent structure with a CA or having the CA suggest answers in the interaction as frequently done in practice via quick reply buttons, and, thereby, lead the conversation in a direction where the agent can provide relevant responses again.

Second, researchers could conceptualize and empirically test different approaches for CAs in a service context to react to unexpected input, such as polite and personal context-specific fallback responses or offering the possibility to contact a human service employee.

7 Conclusion

CAs in organizational contexts promise to provide always available automated service that resembles human interaction. However, current agents have limited capabilities that often lead to situations in which they fail to provide meaningful replies in a service encounter. As such, response failures may negatively impact how people perceive anthropomorphic CAs. Since research has largely neglected response failures, we conducted an experiment to better understand the relationship between response failure and how users perceive CAs.

We found that modest response failures have a detrimental effect on the extent to which people perceive CAs as human and positive effect on the extent to which they perceive them as uncanny and that that humanness (uncanniness) has a positive (negative) impact on familiarity and service satisfaction. Furthermore, we did not find a very strong negative emotional reaction to response failure as the theory of the uncanny valley posits but rather a comparatively moderate negative effect. The findings from our experiment have implications for research, especially regarding the uncanny valley effect in the anthropomorphic CA context. Our results further provide practical insights by confirming the positive impact of humanlike CA design for innovative service provision and by emphasizing the need to mitigate response failures in natural language interactions.

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183-189.

Ashktorab, Z., Jain, M., Liao, V. Q., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81.

Beer, J. M., Smarr, C. A., Fisk, A. D., & Rogers, W. A. (2015). Younger and older users' recognition of virtual agent facial expressions. *International Journal of Human Computer Studies*, *75*, 1-20.

Ben Mimoun, M. S., Poncin, I., & Garnier, M. (2012). Case study-Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services*, *19*(6), 605-612.

Benlian, A., Klumpe, J., & Hinz, O. (2019). Mitigating the intrusive effects of smart home assistants by using anthropomorphic design features: A multimethod investigation. *Information Systems Journal*, 30(6).

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293-327.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human-computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior*, *16*(6), 553-574.

Cafaro, A., Vilhjalmsson, H. H., & Bickmore, T. (2016). First impressions in human-agent virtual encounters. *ACM Transactions on Computer-Human Interaction*, *24*(4), 1-40.

Chin, W. W. (1998). The partial least squares approach for structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295-336). Mahwah, NJ: Lawrence Erlbaum Associates.

Cho, E. (2019). Hey Google, can I ask you something in private? The effects of modality and device in sensitive health information acquisition from voice assistants. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cowell, A. J., & Stanney, K. M. (2005). Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International Journal of Human Computer Studies*, *62*(2), 281-306.

Crockett, K., Latham, A., & Whitton, N. (2017). On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees. *International Journal of Human Computer Studies*, *97*, 98-115.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331-349.

Dennis, A. R., & Valacich, J. S. (2001). Conducting experimental research in information systems. *Communications of the Association for Information Systems*, *7*, 1-41.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019a). On conversational agents in information systems research: Analyzing the past to guide future work. In *Proceedings of the International Conference on Wirtschaftsinformatik*.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering*, *62*(3), 193-209.

Diederich, S., Janßen-Müller, M., Brendel, A. B., & Morana, S. (2019b). 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.*

Diederich, S., Lichtenberg, S., Brendel, A. B., & Trang, S. (2019c). Promoting sustainable mobility beliefs with persuasive and anthropomorphic design: Insights from an experiment with a conversational agent. In *Proceedings of the International Conference on Information Systems.*

Dolata, M., Kilic, M., & Schwabe, G. (2019). When a computer speaks institutional talk: Exploring challenges and potentials of virtual assistants in face-to-face advisory services. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)* (pp. 105-114). Grand Wailea, Maui.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177-190.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864-886.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138-161.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, *24*(4), 38-42.

Gefen, D., & Straub, D. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems*, *16*, 91-109.

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2018). Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. In *Proceedings of the European Conference on Information Systems.*

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *Proceedings of the International Conference on Information Systems.*

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304-316.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494-1509.

Hanson, D. (2006). Exploring the aesthetic range for humanoid robots. *Proceedings of the ICCS/CogSci-2006 Long Symposium*.

Hanson, D., Olney, A., Pereira, I. A., & Zielke, M. (2005). Upending the uncanny valley. In *Proceedings of the AAAI Workshop*.

Hanus, M. D., & Fox, J. (2015). Persuasive avatars: The effects of customizing a virtual salespersons appearance on brand liking and purchase intentions. *International Journal of Human Computer Studies*, *84*, 33-40.

Holtgraves, T., & Han, T. L. (2007). A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods*, *39*(1), 156-163.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Leo, J., Akkiraju, R. (2018). Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Johnson, K. (2018). Facebook Messenger hits 100,000 bots. *Venture Beat*. Retrieved from https://venturebeat.com/2017/04/18/facebook-messenger-hits-100000-bots/

Karahanna, E., Benbasat, I., Bapna, R., & Rai, A. (2018). Opportunities and challenges for different types of online experiments. *Management Information Systems Quarterly*, *42*(4), 3-11.

Liao, Q. V., Hussain, M. M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., Wang, D., Muller, M., Shami, N., & Geyer, W. (2018). All work and no play? Conversations with a question-and-answer chatbot in the wild. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Luger, E., & Sellen, A. (2016). “Like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

MacDorman, K. F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In *Proceedings of the ICCS/CogSci-2006 Long Symposium.*

MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior*, *25*(3), 695-710.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants. *Business & Information Systems Engineering*, *61(4)*, 535-544.

Marinova, D., de Ruyter, K., Huang, M. H., Meuter, M. L., & Challagalla, G. (2017). Getting smart: Learning from technology-empowered frontline interactions. *Journal of Service Research*, *20*(1), 29-42.

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, *146*, 22-32.

McTear, M., Callejas, Z., & Griol, D. (2016). *The conversational interface: Talking to smart devices*. Basel, Switzerland: Springer.

Morana, S., Friemel, C., Gnewuch, U., Maedche, A., & Pfeiffer, J. (2017). Interaktion mit smarten Systemen—aktueller stand und zukünftige entwicklungen im bereich der nutzerassistenz. *Wirtschaftsinformatik & Management*, *5*, 42-51.

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33-35.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics and Automation Magazine*, *19*(2), 98-100.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81-103.

NextIT. (2018). Helping a railroad service conduct business. Retrieved from https://cdn2.hubspot.net/hubfs/120925/Resources/Verint%20Next%20IT_Amtrak_Case-Study_Feb2019.pdf

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., & Patton, M. W. (2011). Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, *28*(1), 17-48.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, *62*(4), 480-486.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, *17*(4), 460-469.

Oracle. (2016). *Can virtual experiences replace reality? The future role for humans in delivering customer experience.* Retrieved from https://www.oracle.com/webfolder/s/delivery_production/docs/FY16h1/doc35/CXResearchVirtualExperiences.pdf

Pereira, A. T., Prada, R., & Paiva, A. (2014). Improving social presence in human-agent interaction. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic information systems. *Business & Information Systems Engineering*, *61*, 523-533.

Qiu, L., & Benbasat, I. (2010). A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human Computer Studies*, *68*(10), 669-688.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television and new media like real people and places*. Stanford, CA: Center for the Study of Language and Information Publications.

Riek, L. D., Rabinowitch, T. C., Chakrabartiz, B., & Robinson, P. (2009). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction*.

Sah, Y. J., & Peng, W. (2015). Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior*, *45*, 392-401.

Schuetzler, R. M., Grimes, G. M., & Giboney, J. S. (2018). An investigation of conversational agent relevance, presence, and engagement. In *Proceedings of the Americas Conference on Information Systems.*

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2018). Designing anthropomorphic conversational agents: Development and empirical evaluation of a design framework. In *Proceedings of the International Conference on Information Systems.*

Seymour, M., Riemer, K., & Kay, J. (2018). Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. *Journal of the Association for Information Systems*, *19*(10), 953-981.

Sohn, S. (2019). Can conversational user interfaces be harmful? The undesirable effects on privacy concern. In *Proceedings of the International Conference on Information Systems*.

Stock, R. M., & Merkle, M. (2018). Customer Responses to Robotic Innovative Behavior Cues During the Service Encounter. In *Proceedings of the International Conference on Information Systems (ICIS)* (pp. 1-17). San Francisco, USA.

Stock, R., Merkle, M., Eidens, D., Hannig, M., Heineck, P., Nguyen, M. A., & Völker, J. (2019). Understanding employee trust in assistive robots when robots enter our workplace: Understanding employee trust in assistive robots. In *Proceedings of the International Conference on Information Systems*.

Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human Behavior*, *27*(2), 741-749.

Tinwell, A., & Sloan, R. J. S. (2014). Children’s perception of uncanny human-like virtual characters. *Computers in Human Behavior*, *36*, 286-296.

Urbach, N., & Ahlemann, F. (2010). Structural equation modeling in information systems research using partial least squares. *Journal of Information Technology Theory and Application, 11*(2), 5-40.

Vaccaro, K., Agarwalla, T., Shivakumar, S., & Kumar, R. (2018). Designing the future of personal fashion experiences online. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

Verhagen, T., van Nes, J., Feldberg, F., & van Dolen, W. (2014). Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, *19*(3), 529-545.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human Computer Studies*, *66*(2), 98-112.

Wünderlich, N. V., & Paluch, S. (2017). A nice and friendly chat with a bot: User perceptions of ai-based service agents. In *Proceedings of the International Conference on Information Systems.*

Yuan, L. (Ivy), & Dennis, A. R. (2019). Acting like humans? Anthropomorphism and consumer’s willingness to pay in electronic commerce. *Journal of Management Information Systems*, *36*(2), 450-477.

Appendix: Cross Loadings for Items and Constructs

Table A1. Cross Loadings

	Familiarity	Humanness	Service satisfaction	Uncanniness
Familiarity (I1)	**1**	0.816	0.751	-0.769
Humanness (I1)	0.745	**0.882**	0.717	-0.685
Humanness (I2)	0.657	**0.887**	0.791	-0.598
Humanness (I3)	0.631	**0.854**	0.739	-0.581
Humanness (I4)	0.569	**0.663**	0.602	-0.725
Humanness (I5)	0.629	**0.829**	0.799	-0.561
Humanness (I6)	0.792	**0.815**	0.751	-0.706
Service satisfaction (I1)	0.609	0.804	**0.914**	-0.587
Service satisfaction (I2)	0.696	0.744	**0.854**	-0.774
Service satisfaction (I3)	0.733	0.868	**0.944**	-0.669
Uncanniness (I1)	-0.48	-0.405	-0.414	**0.663**
Uncanniness (I2)	-0.65	-0.684	-0.599	**0.781**
Uncanniness (I3)	-0.692	-0.676	-0.654	**0.88**
Uncanniness (I4)	-0.687	-0.722	-0.708	**0.912**
Uncanniness (I5)	-0.618	-0.645	-0.627	**0.853**
Uncanniness (I6)	-0.696	-0.703	-0.688	**0.897**

About the Authors

Stephan Diederich is a former research associate at the Chair of Information Management of the University of Göttingen and a senior consultant at a large management consultancy concentrating on the digital transformation of organizations. His research focuses on conversational agents, in particular chatbots, within an organizational context, and has been published in journals such as *Business and Information Systems Engineering* and IS conferences like the *International Conference on Information Systems* or the *European Conference on Information Systems*.

Tim-Benjamin Lembcke is head of the Smart Mobility Research Group, which is part of the Chair of Information Management at the University of Göttingen. His research focuses on smart mobility, conversational agents, design thinking, and digital nudging. He leads a team responsible for several national and international research projects within the area of smart mobility. His publications have appeared in several leading IS conferences, such as the *International Conference on Information Systems* or the *European Conference on Information Systems*.

Alfred Benedikt Brendel is the chairholder for business informatics, esp. intelligent systems and services, at TU Dresden, Germany, and holds a doctoral degree from the University of Göttingen. His research focuses on Design Science Research to develop novel design principles and theories in the areas digital health, smart mobility, and crowd working. Alfred's work has been published in IS conferences, such as the *International Conference on Information Systems* and *European Conference on Information Systems*, and journals like *Business & Information Systems Engineering* and *Information Systems Frontiers*.

Lutz Kolbe is a professor of Information Systems and leads the Chair of Information Management at the University of Göttingen. His research focuses on the management of information and information technology as a crucial factor for sustainable business success. Lutz manages the research and project activities of three groups covering the areas of digital transformation, digital health, and smart mobility. His publications have appeared in several IS journals and conferences, such as the *European Journal of Information Systems*, the *Journal of the Association for Information Systems*, and the *Information Systems Journal*.