Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

Human Technology Interaction

Dec 11th, 12:00 AM

New Bots – The Influence of a Conversational Agent’s Rookie Personality on Users’ Satisfaction

Fabian Hildebrandt
Technische Universität Dresden, fabian.hildebrandt@tu-dresden.de

Benedikt Brendel
Technisch Universität Dresden, Alfred_benedikt.brendel@tu-dresden.de

Alan Dennis
Indiana University, ardennis@indiana.edu

Agrim Sachdeva
Indiana University, agsach@iu.edu

New Bots – The Influence of a Conversational Agent's Rookie Personality on Users' Satisfaction

Completed Research Paper

Fabian Hildebrandt
TUD Dresden University of Technology
Dresden, Germany
fabian.hildebrandt@tu-dresden.de

Alfred Benedikt Brendel
TUD Dresden University of Technology
Dresden, Germany
alfred_benedikt.brendel@tu-dresden.de

Alan R. Dennis
Kelley School of Business,
Indiana University
Bloomington, IN, United States
ardennis@indiana.edu

Agrim Sachdeva
Kelley School of Business,
Indiana University
Bloomington, IN, United States
agsach@iu.edu

Abstract

Conversational agents (CAs) are not likely to be error-free, and efforts are being made by research and practice to mitigate the negative consequences of such errors (e.g., reduced service satisfaction). In this context, our study examines the impact of a CA's rookie personality (i.e., the CA expresses that it is new and still learning) on users. Our findings reveal that the rookie personality is a double-edged sword: while it increases users' perception of humanness, which increases the perception of reliability, it also directly reduces perceived reliability, resulting in less service satisfaction. To explain these seemingly contradictory effects, we turn to the dual processing theory of cognition and propose that the rookie personality influences both automatic and deliberate thinking. Users actively and consciously contemplate the CA's messages, leading them to view the software artifact as "broken" and low-quality. Additionally, users' automatic thinking is influenced by the perception of humanness.

Keywords: Conversational Agents, Errors, Rookie Personality, Anthropomorphism, Perceived Humanness, Expectation Confirmation, Reliability, Service Satisfaction

Introduction

The abilities of conversational agents (CAs) have increased drastically in recent years (McTear et al., 2016), for instance, enabling them to be smart assistants at home (e.g., Amazon's Alexa and Apple's Siri) or customer interfaces for e-commerce (McTear, 2017). CAs are defined as "software-based systems designed to interact with humans using natural language" (Feine, Gnewuch, et al. 2019, p. 1.). CAs can automate various manual tasks, which were traditionally done by customer service employees, such as responding to customer requests and providing answers to FAQ inquiries (Gnewuch et al., 2017; Vu et al., 2021). These services are provided time and place independently and with a highly convenient user experience (Verhagen et al., 2014). To harness the benefits of CAs, many companies have increased their efforts in developing more effective and efficient CA-users interactions intended to increase customer satisfaction, cost savings, and revenue (McTear et al., 2016).

Nonetheless, despite great efforts, CAs will still produce errors because of the complexity of natural interactions (Brandtzæg & Følstad, 2018) (Christiansen & Kirby, 2003). Subsequently, in the past, CAs have been discontinued because of their inability to engage in effective dialogue and provide consistently meaningful responses (Ben Mimoun et al., 2012). Errors can have detrimental effects on users' perception of the CA and its service. For instance, not understanding user inputs and responding with a fallback answer (e.g., "I did not understand that. Could you rephrase your request?") is a common error in CA-user interactions (Diederich et al., 2021). Studies have shown that such errors lead to reduced service satisfaction and intention to use (Brandtzæg & Følstad, 2018; Bührke et al., 2021; Diederich et al., 2021; Sheehan et al., 2020). In response, developers and designers have engaged in addressing this issue in two main ways (Diederich et al., 2021; Larivière et al., 2017): (1) by improving the technology behind CA (Lester et al., 2004), and (2) by finding ways to reduce the negative effects of errors (Benner et al., 2021).

Against this background, research has engaged in investigating when and how the humanlike design of a CA (i.e., equipping CAs with social cues, such as human name, avatar, and greeting users) can be a remedy for the negative effects of errors (Riquel, Brendel, Hildebrandt, Greve, & Kolbe, 2021). The humanlike design leads users to a perception of humanness in the CA (Gnewuch et al., 2017; Nass & Moon, 2000). This perception has been shown to counteract or lessen the negative effects of CAs errors (Riquel, Brendel, Hildebrandt, Greve, & Kolbe, 2021). However, it has also been pointed out that relying on the perception of humanness might not be the most effective or only way to remedy the negative effects of errors (Benner et al., 2021). For instance, one study found that the perception of humanness can lead to greater frustration with errors, potentially explained by users perceiving the CA to have caused the error intentionally (Riquel, Brendel, Hildebrandt, Greve, & Dennis, 2021).

Instead, research points to investigating how CA could portray certain human personalities that are specifically intended to counteract the effects of errors (Brendel et al., 2020; Pradhan & Lazar, 2021). Prominently, a rookie – i.e., someone that is new and still learning – could be a potential solution (Riquel, Brendel, Hildebrandt, Greve, & Dennis, 2021), which has also been applied in practice (e.g., ChatGPT stating its limitations and that it is still learning). At first glance, applying a rookie personality makes sense because it is associated with lower expectations and forgiveness of errors in human-to-human interactions (Boostrom, Jr., 2008). However, it remains unclear if these effects translate from human-to-human to human-to-CA interactions. Furthermore, based on the expectation-confirmation theory (Oliver, 1981), two contradicting effects can be derived. On the one hand, stating upfront that a CA is still learning and errors are likely to occur could be a means of expectation management, having a positive effect on service satisfaction (Ahmad et al., 2022; Oliver, 1981). On the other hand, stating that errors are likely to occur could lead to users perceiving the CA to be less "good" (i.e., of lesser quality, having worse performance, and not being reliable), which could lead to reduced service satisfaction (Antonio et al., 2022). In this study, to address this tension, we address the following research question:

> *RQ: How does portraying a rookie personality change users' perceptions of and satisfaction with a CA that produces errors?*

To answer this question, we conducted a two-conduction online experiment with 106 participants. For the experiment, we implemented two chatbots that produce an error during the interaction. Both were designed with a generic humanlike design but only one of them displayed a rookie personality by stating to the users that it is still learning, and errors might happen. Based on the data, we analyzed how the portrayal of a rookie influence the perceived reliability of the CA, the confirmation of expectations, the level of perceived humanness, and service satisfaction.

Based on existing theory and evidence from other studies, we deducted three distinct pathways for the effect of a CA's rookie personality on service satisfaction. Our results provide support for the first pathway: a CA's rookie personality increases users' perception of humanness, which leads to higher confirmation of expectations, perceived reliability, and subsequent service satisfaction. However, there is no indirect effect of the rookie personality on service satisfaction via this path. Further, we find no support for our second pathway: a rookie personality *does not* have a direct effect on users' confirmation of expectations. Lastly, our third pathway was supported by our data, showing that a CA's rookie personality has a negative effect on users' perceived reliability, which in turn reduces service satisfaction. Against this background, we would describe equipping a CA with a rookie personality as a double-edged sword. On the one side, it increases the perception of humanness, leading to positive effects (i.e., increases confirmation of expectations, perceived reliability, and service satisfaction). On the other side, it has negative effects (i.e., decreasing confirmation of expectations and service satisfaction).

Research Background and Related Work

CAs can communicate with users via verbal speech (often called voice assistants (Schuetzler et al., 2018)) or via written text (often called chatbots (Følstad & Brandtzaeg, 2017)). Since the very first CA named ELIZA (Weizenbaum, 1966) (Gnewuch et al., 2017), advancements in machine learning and language processing (i.e., in the area of Natural-Language-Processing) have drastically increased the abilities of CAs (McTear, 2017). Furthermore, the rapid increase in CAs in practice (McTear, 2017) is driven by the widespread availability of mature CA development technology (e.g., Google Dialogflow, ChatGPT). Now, CAs can replace human employees for various professional work and service interactions (McTear et al., 2016). They have been applied in various contexts, such as human resources (Liao et al., 2018), sales (Adam et al., 2022), and customer service (Araujo, 2018; Gnewuch et al., 2017). CAs are free of common limitations of human-based services, such as time and place restrictions (McTear et al., 2016). Against this background, in the following sections, we will present research on the humanlike design of CA and how it influences users. Lastly, we will outline the phenomenon of CAs producing errors and related research.

Humanlike Design and Personalities of Conversational Agents

The tendency to ascribe humanlike characteristics to non-human entities (e.g., animals or cartoon characters (Epley et al., 2007)) is deeply ingrained in human nature (Kunda, 1999). To provide an example, Yuan & Dennis (2019) edited a picture of a tablet to have a cartoonish face and hands, leading to onlookers reporting a perception of humanness. Similarly, this tendency is also present when people interact with CAs (e.g., Alexa or Siri) (Araujo, 2018). In this context, the “Computers are Social Actors” (CASA) paradigm (Nass et al., 1994) and the Social Response Theory (Nass & Moon, 2000) explain how the perception of humanness takes effect in human-to-computer interaction.

CASA argues that users attribute some degree of humanness to a computer, despite knowing that it is a machine and not humane (Nass et al., 1994). The degree of the perceived humanness depends on the humanlike features – so-called social cues (e.g., having a name or gender) they perceive. Social cues are “multimodal verbal and nonverbal characteristics usually associated with humans” (Feine, Gnewuch, et al. 2019, p. 1) and the term “humanlike design” refers to a computer (e.g., CAs) that is equipped with social cues to appear more similar to a human (Araujo, 2018; Feine et al., 2019; Seeger et al., 2018). Because of the perceived humanness caused by social cues, users apply social norms (e.g., gender or gender stereotypes) to the interaction with the computer (Lang et al., 2013; Nass et al., 1994; Nass & Moon, 2000). Building upon CASA, the social response theory (Nass & Moon, 2000) adds that the humanlike design can trigger automatic responses (Feine et al., 2019; Nass & Moon, 2000), which leads to the interaction with the computer to feel similar to the interaction with a human (Gnewuch et al., 2017). This phenomenon can be further explained by the dual processing theory, which proposes that humans have two modes of cognition: automatic and deliberate (Kahnemann, 2011). Automatic cognition is fast and instinctive, while deliberate cognition is slower and effortful (Kahnemann, 2011). Automatic cognition controls most of our attitudes and behaviors and we only invoke deliberate cognition when we are motivated to expend effort, typically when something unexpected occurs (Kahnemann, 2011). Equipping a CA with social cues leads to an automatic cognition followed by a social response, which type and strength depend on the perceived humanness (Gong, 2008) – i.e., a high degree of perceived humanness makes it more likely that users react with social behavior (Nass & Moon, 2000). For instance, users respond politely and express gratitude (i.e., say thank you) when they perceive higher degrees of humanness in CAs (Wang et al., 2008).

Besides implementing a generic human-like design, a recent trend is to implement personalities (Pradhan & Lazar, 2021). In the context of CAs, the term “personality” is used to describe a CA’s stable traits, which guide the way a CA interacts across contexts and time (Lessio & Morris, 2020). One common example is the comparison of Alexa and Siri (Pradhan & Lazar, 2021). Alexa is designed to be smart, approachable, humble, enthusiastic, helpful, and friendly (Kim et al., 2019) and Siri to be friendly and humble, but also with some sassiness (Kim, 2011; Mardsen, 2015). The personalities of Alexa and Siri are similar (friendly and humble) but not identical (e.g., sassiness), leading to different interactions for users. Against this background, research on CAs that are designed to portray a personality has shown various effects on users. On the one hand, a friendly and social personality of a CA (e.g., Alexa or XiaoIce) can lead to a long-term friendship (Shum et al., 2018). On the other hand, a CA with a persecutor personality got users' to show aggressive behavior towards the CA (Brendel et al., 2020). Nonetheless, empirical evidence remains somewhat scarce because of the novelty of the research area (Pradhan & Lazar, 2021; Shum et al., 2018; Sonlu et al., 2021).

Errors of Conversational Agents

Despite the sophistication and increased maturity of the technology underlying CAs, CAs are and will probably never be perfect; they are prone to produce errors because they cannot use natural language at the same level as a human can (Brandtzæg & Følstad, 2018). A CA's ability to interact with users is depended on the abilities of the developers and the technology used (Brandtzæg & Følstad, 2018; Verhagen et al., 2014). For instance, errors commonly occur because of limitations in natural language processing, such as a limited vocabulary, inappropriate choice of words, or limited training data (Zemčík, 2021). Subsequently, CAs with inadequate development and training cannot understand all user requests (Brandtzæg & Følstad, 2018). For instance, there are various ways of stating an input (i.e., the agreement can be expressed as "yes" and also as "sure") and human language is ever-evolving, leading to new words and phrases (Christiansen & Kirby, 2003).

These errors of CA (either in language or content) are detrimental to the user experience (Ben Mimoun et al., 2012; Brandtzæg & Følstad, 2018; Bührke et al., 2021). To address this issue (besides improving natural language processing technology) research has engaged in findings ways to design the interaction so that the negative effects of errors are reduced (Diederich et al., 2021; Gnewuch, Morana, et al., 2018; Larivière et al., 2017). However, research on this topic is still new and upcoming. Extending on a recent comprehensive review of publications in IS and HCI outlets by Diederich et al. (2022), we were able to identify a very limited set of studies on the topics of CA's errors.

Sheehan et al. (2020) showed that interacting with a flawed (i.e., error-producing) CA leads users to perceive the interaction very negatively, despite similar errors also occurring in human-to-human interactions. In their study, De Angeli & Brahnam (2008) analyzed 146 conversations of users with the Jabberwacky chatbot. They focused on the reasons for users to behave aggressively and found that one of the main reasons was the occurrence of errors. Similarly, Seering et al. (2020) found that the occurrence of errors is also the driver of aggression toward the chatbot named "Babybot." The results of Weiler et al. (2021) suggest that informing users before the interaction with a CA about the potential for error (i.e., inoculation messages) reduced the probability of users discontinuing the interaction when such an error occurred. Riquel, Brendel, Hildebrandt, Greve, & Kolbe (2021) found that the human-like design of CA can preserve service satisfaction when errors occur because of the related pleasant emotions. Regarding the relation of error, frustration, and aggression, Riquel, Brendel, Hildebrandt, Greve, & Dennis (2021) showed that the perception of humanness in CA has contradicting effects, increasing frustration directly and reducing it indirectly via reduced dissatisfaction. In summary, some recent studies have engaged the topic of errors by CAs. Most recently, the topic of humanlike design of CAs has garnered increasing intention, highlighting that it has positive (increasing positive emotions (Riquel, Brendel, Hildebrandt, Greve, & Kolbe, 2021)) but also negative (increasing frustration (Riquel, Brendel, Hildebrandt, Greve, & Dennis, 2021)) effects.

In this context, designing CA to portray a specific personality has been highlighted to be a valuable avenue for research but has yet to be engaged extensively (Pradhan & Lazar, 2021; Shum et al., 2018; Sonlu et al., 2021). In the context of errors, we were only able to identify one study addressing this topic by studying a personality that blames others for mistakes (Brendel et al., 2020). Research has engaged with the study of human-CA interaction breakdown recovery strategies (Benner et al., 2021). For instance, the usage of emojis by a CA increases consumers' willingness to continue using them after service failures (Liu et al., 2023). Furthermore, Song et al. (2023) found that a good human-CA relationship is more effective than admitting the CAs limited competence. In this context, implementing a rookie CA links the portraying of a personality with the topic of error recovery strategies, adding a new aspect for service recovery.

Research Model and Hypotheses Development

Our research investigates the influence of a CA portraying a rookie personality on service satisfaction. Our research model (see Figure 1) is based on the CASA paradigm (Nass et al., 1994) (Nass & Moon, 2000).

Furthermore, we combine it with the expectation-confirmation theory (Oliver, 1981). We theorize that the rookie personality influences users' service satisfaction via three distinct pathways. First, a rookie personality adds to the humanlike design of a CA, increasing the perceived level of humanness. This increase in perceived humanness also influences the user's evaluation of the CA, resulting in an increased confirmation of expectations and perceived reliability, which are both drivers of service satisfaction. Second, the message of the CA is still learning, and errors can happen and can be understood as a means of expectations management, which should lower the expectations of users, making it more likely that the CA can confirm these lower expectations, which results in higher service satisfaction. Third, as part of the rookie personality, the CA highlights its flaw to the users, which reduces the users' perceived reliability of it, which reduces users' service satisfaction. In the following, we will describe our hypotheses in more detail.

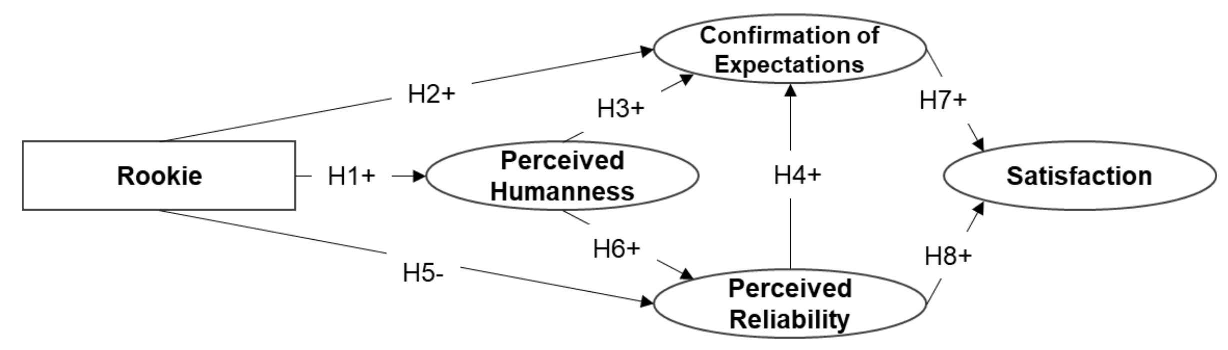


Figure 1. Research Model

Perceived Humanness

As we noted above, CAs can be designed to be more humanlike by equipping them with social cues (Seeger et al., 2018). These social cues can trigger anthropomorphism in users – i.e., perception of humanness (Gnewuch, Morana, et al., 2018). This perception of humanness can be understood as the degree to which a person believes that a CA might be human (Kirakowski et al., 2007). From a CASA perspective, ascribing humanness to computers equipped with social cues (e.g., having a name) is an automatic behavior (Nass & Moon, 2000). Users know that computers, including CAs, are not human, but this does not prevent them from perceiving with some degree of humanness (Nass & Moon, 2000). In this context, the portrayal of a personality (i.e., traits that are consistent during interactions and across time) adds to the humanlike design and increases the perceived humanness because having a personality is a human trait (Ahmad et al., 2022; Pradhan & Lazar, 2021). In this study, we focus on the personality of a rookie (Boostrom, Jr., 2008), which consists of the trait to express that it is still new, learning, and errors can happen (i.e., disclosure and social strategy (Benner et al., 2021)). This display of "weakness" and anticipating the expectations of others (i.e., the chatbot knows that users will not expect errors, therefore, disclosing that they will probably happen) increases users perceived humanness because it adds additional dimensions to its humanlike design (Seeger et al., 2018). Support can be found in the literature. For instance, Wagner & Schramm-Klein (2019) showed that the personality of Amazon's Alexa increased the perception of humanness. Moreover, Zhu et al. (2019) showed that anthropomorphized objects are perceived as more humanlike if they displayed weakness (e.g., through social roles - a child vs. a mother). Thus, we hypothesize:

H1: Displaying a rookie personality increases perceived humanness.

Confirmation of Expectation

Humans are constantly forming expectations – i.e., a consideration of was is most likely to happen or what attributes and characteristics an entity (e.g., product or service) will have (Zeithaml et al., 1993) – and evaluating to which degree their expectations are confirmed by future occurrences (Coye, 2004; Zeithaml et al., 1993). In general, humans strive for and prefer confirmation of their expectations – i.e., their predictions were correct (Oliver, 1981). In terms of CAs, expectations are higher towards a machinelike CA than a humanlike CA (Mirnig et al., 2017). In this context, the message of the rookie personality can be expected to influence the expectations of the users (Ahmad et al., 2022; Kim et al., 2019; Sonlu et al., 2021). Because of the messages, users are aware of the ability of the CAs (i.e., still learning and errors can happen) and can, therefore, adjust their expectations accordingly – i.e., normally errors are unexpected and perceived as negative, but because of the rookie messages errors are no longer unexpected (Benner et al., 2021; Pradhan & Lazar, 2021; Zeithaml et al., 1993). In literature, for instance, Mayhew et al. (2003) showed that more mistakes or accidents are expected from a novice driver and are therefore more likely to be forgiven. Similarly, Newell (1983) showed, that a class is more likely to forgive their rookie law teacher for his mistakes. However, there are no studies on this matter in the context of CAs. Nonetheless, based on the presented evidence, we postulate the following hypothesis:

H2: Displaying a rookie personality increases confirmation of expectations.

The process of evaluating the confirmation or disconfirmation of expectations is a subjective one and depended on the available information (Boulding et al., 1993; Coye, 2004). Any thinking, including perceiving one's surroundings and processing them, is subjective, meaning that emotions and biases are highly influential (Lerner et al., 2015; Levinson, 1995). In this context, the human tendency to seek human likeness in non-human entities (e.g., nature and objects) (Epley et al., 2007) – i.e., anthropomorphism – and the influence the perception of humanness has on one's thinking, can be expected to also influence the evaluation of expectation confirmation (Grimes et al., 2021; Oliver, 1981; Zeithaml et al., 1993). In general, humans prefer social interaction (Levinson, 1995) which should have a positive influence on the evaluation confirmation of expectations because, in general, a favorable state of mind leads to more positive perceptions and evaluations (Blanchette & Richards, 2010). Support can be found in current publications. For instance, Babel et al. (2021) showed that users tend to trust a robot more when it is designed humanlike. Hence, the humanlike design influences users' evaluation of the trustworthiness of the CA, despite no logical relation between humanlike design and trust. Another example is that Pak et al. (2012) found an increase in users' perceived performance if a CA is designed humanlike. Lastly, Grimes et al. (2021) found that higher conversational skills influence users' evaluation of expectations. Against this background, we hypothesize:

H3: Perceived humanness increases confirmation of expectations.

Perceived Reliability

In the context of software development, reliability can be understood as the ability of the software, such as CAs, to perform consistently well under the intended conditions (Jiang et al., 2002; Kettinger & Lee, 1994). For users, this perception of reliability is based on their experience with the CA as it performs (i.e., provides its services) (Meyer-Waarden et al., 2020). In the context of the expectation confirmation theory, the performance of a product or service has a great influence on individuals' confirmation of expectations (Oliver, 1981). Thus, because reliability is a characteristic of the performance of a system (Jiang et al., 2002; Meyer-Waarden et al., 2020), it influences users' confirmation of expectations (Boulding et al., 1993; Oliver, 1981) – i.e., is the CA as reliable as expected. Recent research provides empirical support for these considerations. For instance, in the context of hotel ratings, it has been found that reliable service leads to people expressing greater satisfaction because their expectations have been met (Nam et al., 2020). Similarly, in context of clinical information systems, the perceived performance is positively related to clinician expectations congruency (Karimi et al., 2015). However, no such research has been conducted in the context of CAs. Nonetheless, based on the evidence in other contexts, we hypothesize:

H4: Perceived reliability increases confirmation of expectations.

The rookie personality informs the user about its flaws (Ahmad et al., 2022; Benner et al., 2021) and is, therefore, likely to be perceived to be less reliable (Jiang et al., 2002; Meyer-Waarden et al., 2020). The actual performance is not influenced by this statement, but the perception of the CA is still changed and priming users' assessment (Buck & Dinev, 2020; Meyer-Waarden et al., 2020). Support can be found in the literature for this deduction. For instance, Miller & Peake (2010) found that a politician was perceived as less reliable because of her rookie image. Similarly, rookie basketball players are perceived as less durable and reliable by the viewer, independent of their actual performance (Berri et al., 2011; Solow & von Allmen, 2016). Against this background, we hypothesize:

H5: Displaying a rookie personality reduces perceived reliability.

Similar to the confirmation of expectations, a user's evaluation of a CA's reliability is subjective and bound by limited information (Zeithaml et al., 1988). Their affective reaction to the CAs can be expected to influence the users' assessment of the CA's reliability. As we already argued for the evaluation of expectation confirmation, the perceived humanness can be expected to lead to a positive mood that increases the perceived reliability. We can find support for this proposition in the literature. For instance, Daryanto et al. (2022) showed that an anthropomorphized brand logo has a positive impact on the perceived functional performance of the service. Similarly, a humanlike service robot in hospitality and tourism increases the users' perceived quality, despite providing the same service as a non-humanlike robot (Murphy et al., 2017). Therefore, we hypothesize:

H6: Perceived humanness increases perceived reliability.

Service Satisfaction

Service satisfaction is understood as a cumulative process that reflects the individual's affective and cognitive evaluative response toward a product, service, benefit, or reward (Millán & Esteban, 2004; Oliver, 1997). A high level of service satisfaction expresses an individual's perception of high quality and service experience (Jiang et al., 2002; Oliver, 1997). Following expectation confirmation theory, service satisfaction is driven by the evaluation to which degree expectations were met (Oliver, 1981). Support for the relationship between confirmation of expectations and service satisfaction can be found across disciplines. For instance, Wu et al. (2020) show that confirmation of customer expectoration in online shopping (i.e., through product descriptions) leads to higher satisfaction. In the context of CAs, Li et al. (2022) showed that patients' continuance intention toward CAs is driven by higher satisfaction, which is a result of confirmation of expectations. Thus:

H7: Confirmation of expectation increases service satisfaction.

Besides the confirmation of expectations, service satisfaction is highly influenced by perceived performance (Coye, 2004; Jiang et al., 2002). In this context, reliability is an aspect of performance (Meyer-Waarden et al., 2020) – i.e., a software's performance is partly evaluated based on its reliability (Jiang et al., 2002; Kettinger & Lee, 1994). Thus, the perceived reliability of CAs can be expected to also influence a user's satisfaction. In literature, support for this proposition can be, for instance, found in the study of Korda & Snoj (2010). In their study, they reported a relationship between perceived quality and perceived value of a bank service on overall satisfaction. Furthermore, Antonio et al. (2022) showed that chatbots in e-commerce customer service systems have a higher associated perceived service satisfaction, which is affected by its perceived reliability. Therefore, we postulate this hypothesis:

H8: Perceived reliability increases service satisfaction

Method

In the following sections, we will summarize our sample of participants, the implemented task and procedure for our experiment, the four different treatment designs, and the measures included in our survey.

Participants

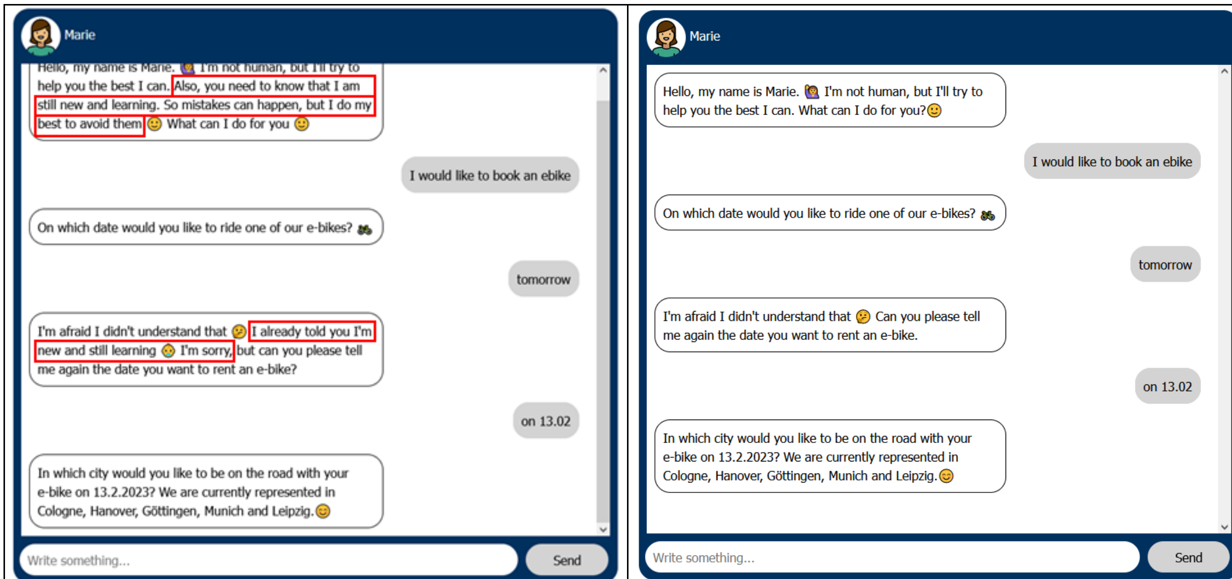
We recruited 106 students from a German university via email. For their participation, they had the chance to win one of five €10 online shopping vouchers. To ensure participation, the experiment was easily accessible via a web link and could be completed anywhere, at any time, and with any device that has an internet connection (the interface of the experiment was implemented with a responsive interface that is correctly displayed on any device, including tablets and smartphones). We had to remove eight responses because of failed attention checks or incomplete responses, resulting in a final sample of 98 valid responses. The participant's ages ranged from 18 to 32 (mean: 23 years), and 48% of them identified as female.

Task and Procedure

Based on the example of other experimental studies on CAs (e.g., Bührke et al. (2021), Diederich et al. (2021), and Gnewuch, Adam, et al. (2018)), we implemented a user-CA interaction that was for a specific task and had a clear dialogue structure. Every participant received the same information (Dennis & Valacich, 2001), describing the task: rent an e-bike via a chatbot. In this context, we explained that the chatbot was a machine, and participants would not interact with an actual human. The rental process consisted of 8 steps: (step 1) starting the e-Bike reservation process, (step 2) inserting the date, (step 3) inserting the city, (step 4) stating their reason for use, (step 5) choosing an e-Bike type, (step 6 & 7) providing a first and last name, and (step 8) providing an e-mail address. Afterwards, the CA provided a link to the survey. Overall, the experiment took about 10 minutes per participant.

For the error, the chatbots were programmed to misunderstand the first input of the date they want to rent the bike (see Figure 2). We implemented this type of error because of its high realism – dates can be stated in various ways, such as different phrasings ("tomorrow" versus "in one day") and formats ("9th December 2022" versus "9/12/2022"). Hence, a chatbot not understanding users' input of a date can be considered a common error (Diederich et al., 2021).

Treatments



Note: Messages related to the rookie personality are highlighted in red.

A. Rookie Personality	**B. No Rookie Personality**

Note: Dialogues translated from German to English

Figure 2. Treatment Designs

To avoid carryover effects. we implemented a between-subject design for our experiment (Boudreau et al., 2001). Every participant was randomly assigned to either the rookie or the non-rookie chatbot (see Figure 2). The chatbots in both treatments were implemented identically (e.g., the same interface, natural language processing engine 'Google's Dialogflow', and training phrases). The chatbots could process and understand different wordings and extract, validate, and repeat parameters from user inputs (e.g., use the user's entered name in responses).

For the rookie personality, we implemented three additional messages. The first message was included in the greeting ("Also, you need to know that I am still new and learning. So, mistakes can happen, but I do my best to avoid them"). The second message was placed after the error occurred and included an apology ("I already told you I'm new and still learning. I'm sorry"). The last message came out after the service was completed and the interaction ended ("I apologize again if I didn't understand something the first time because I'm still learning […]"). See Figure 2 for a screenshot of each chatbot. In summary, for the rookie personality, we conceptualize a rookie personality to have three aspects: taking responsibility for errors, apologizing for the error (Song et al., 2023), and explaining that the errors might be caused by inexperience (i.e., being new and still learning).

Both chatbots were implemented with a humanlike design adopted from other studies on humanlike chatbot design (Araujo, 2018; Bührke et al., 2021; Gnewuch, Morana, et al., 2018), and based on Seeger et al.'s (Seeger et al., 2018) three dimensions (a human identity, verbal cues, and non-verbal cues). The first dimension is implemented in the form of a human name (Marie), avatar, and stereotypical gender (female). The second dimension (verbal cues) is implemented in the form of greeting ("Hello, my name is Marie."), self-reference ("... can I do..."), and politeness ("Can you please..."). Non-verbal cues (third dimension) are present in the form of the usage of emojis and dynamic response delays with associated blinking dots.

Measures

To test the research model and related hypotheses, our survey included items related to perceived humanness (Gefen & Straub, 1997), confirmation of expectations (Bhattacherjee, 2001), perceived reliability (based on Stone-Romero et al. (1997)), and service satisfaction (Verhagen et al., 2014) (Table 1).

Latent Variable	**Mean**	**SD**	**Loading**
Perceived Humanness (Cronbach's α = .866, CR = .902, AVE = .649)			
I felt a sense of human contact with the chatbot.	4.755	1.378	.811
I felt a sense of personalness with the chatbot.	3.796	1.635	.819
I felt a sense of sociability with the chatbot	5.418	1.212	.775
I felt a sense of human warmth with the chatbot.	4.082	1.550	.837
I felt a sense of human sensitivity with the chatbot.	3.745	1.631	.778
Confirmation of Expectation (Cronbach's α = .793, CR = .881, AVE = .714)			
My experience with the chatbot was better than what I had expected.	4.918	1.482	.855
The service provided by the chatbot was better than I expected.	4.867	1.419	.861
Overall, most of my expectations from using the chatbot were confirmed.	5.337	1.186	.770
~~The expectations that I had about the chatbot were correct.~~	~~5.000~~	~~1.245~~	~~.577~~
Perceived Reliability (Cronbach's α = .866, CR = .906, AVE = .708)			
Unreliable - Reliable	6.439	1.835	.878
Not durable - Durable	5.571	2.010	.754
Uncertain - Dependable	6.408	1.916	.880
Low quality - High quality	6.398	1.677	.847
Service Satisfaction (Cronbach's α = .832, CR = .899, AVE = .748)			
I was satisfied with the overall interaction with the chatbot.	5.418	1.169	.898
I was satisfied with the way the chatbot treated me.	6.061	0.935	.837
I was satisfied with the chatbot's response.	5.173	1.464	.859

Table 1. Measurement Validation of Constructs

	1.	**2.**	**3.**	**4.**	**5.**
1. Perceived Reliability	**.841**				
2. Confirmation of Expectation	.388	**.845**			
3. Perceived Humanness	.440	.439	**.806**		
4. Rookie Message	-.087	.077	.257	**n. a.**	
5. Service Satisfaction	.555	.678	.429	.080	**.865**

n. a. = not applicable

Table 2. Discriminant Validity

All items were measured on a seven-point Likert scale. All measured constructs and the associated factor loadings, Cronbach's α, composite reliability (CR), mean (M), and standard deviation (SD) are summarized in Table 1. Furthermore, we included questions regarding demographics (age, gender, and education), three attention checks, and a binary manipulation check ("The chatbot told me that it is new and still needs to learn"). As suggested by Gefen & Straub (2005), we only included items for analysis with a factor loading above the threshold value of .60. Therefore, one item of confirmation of expectations was dropped. All measured constructs exhibit sufficient reliability due to a CR > .80 (Nunally, 1970) and a Cronbach's α > .70 (Cortina, 1993). Furthermore, results of convergent and discriminant validity analyses also indicate sufficient validity due to AVEs >.50 (Hair et al., 2010) and a fulfilled Fornell-Larcker criterion (Table 2) (Fornell & Larcker, 1981).

Results

Manipulation Check and Descriptive Statistics

For checking if the participants recognized the implemented rookie personality, we included a manipulation check in our survey ("The chatbot told me that it is new and still needs to learn" – 0: No, 1: Yes). Due to the binary scale, we investigated the difference between the control group and the treatment with a Fisher's Exact Test. We found a significant effect (p <.001), indicating that the manipulation was perceived by the participants as intended. Furthermore, the descriptive statistics (mean and SD) for each construct are summarized in Table 3. An a priori power analysis using G*Power found that 90 participants are required to detect a medium effect with a power of .80 (Faul et al., 2009).

		No Rookie (N=50)	**Rookie (N=48)**
Manipulation Check	Mean Standard Deviation	0.080 0.274	1.000 0.000
Perceived Humanness	Mean Standard Deviation	4.068 1.237	4.663 1.106
Confirmation of Expectation	Mean Standard Deviation	4.953 1.176	5.132 1.152
Perceived Reliability	Mean Standard Deviation	6.355 1.600	6.047 1.561
Service Satisfaction	Mean Standard Deviation	5.460 1.026	5.646 1.053

Table 3. Descriptive Statistics

Results of Hypotheses Testing

To test our assumptions, we used partial least square (PLS) regression with SmartPLS 3.3.6. Due to the estimator's advantages in terms of restricted assumptions, PLS is frequently utilized in experimental research (Fombelle et al., 2016). We opted to use the structural equation model as the research design for our study because the possibility of measurement errors and the complex multidimensional nature of theoretical constructs are taken into account (Bagozzi & Yi, 1988). According to Chinn's (1998) suggestion, the significance of the route coefficients was determined using a bootstrapping resampling approach using 5,000 samples. Figure 3 shows all findings together with the path coefficients and significance levels.

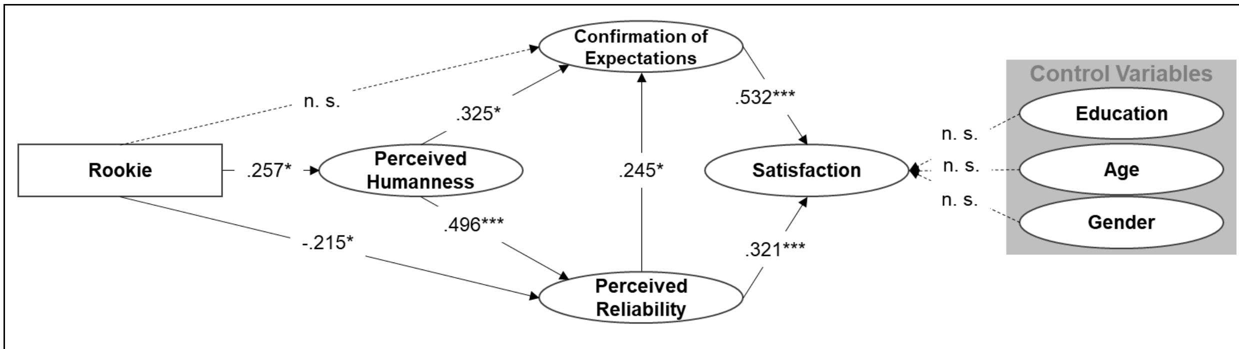


Figure 3. Structural Equation Model (n = 98)

A rookie personality significantly increases perceived humanness (β = .257, p = .012) (supporting **H1**) and significantly decreases perceived reliability (β = -.215, p = .017) (supporting **H5**). However, no significance was found from the rookie personality on confirmation of expectations (β = .014, p = .884), and thus **H2** is not supported. The relationship between perceived humanness and confirmation of expectation (β = .328, p = .013) and perceived reliability (β = .496, p <.001) reveal a significant impact (supporting **H3** and **H6**). Furthermore, we found support for **H4** because of a significant increase in confirmation of expectations through perceived reliability (β = .245, p = .021). The last hypotheses examined the relationship between confirmation of expectations and service satisfaction (**H7**) and perceived reliability and service satisfaction (**H8**). Both hypotheses are supported and showed a significant increase in satisfaction through the confirmation of expectations (β = .532, p <.001) and perceived reliability (β = .321, p <.001).

H.	Relationship	β-value	t-value	p-value	Support
H1+	Rookie Personality → Perceived Humanness	.257	2.511	.012*	**Yes**
H2+	Rookie Personality → Confirmation of Expectations	.014	0.145	.884	No
H3+	Perceived Humanness → Confirmation of Expectations	.328	2.495	.013*	**Yes**
H4+	Perceived Reliability → Confirmation of Expectations	.245	2.312	.021*	**Yes**
H5-	Rookie Personality → Perceived Reliability	-.215	2.378	.017*	**Yes**
H6+	Perceived Humanness → Perceived Reliability	.496	6.719	<.001***	**Yes**
H7+	Confirmation of Expectations → Service Satisfaction	.532	7.005	<.001***	**Yes**
H8+	Perceived Reliability → Service Satisfaction	.321	4.480	<.001***	**Yes**
** p < 0.05, ** p < 0.01, *** p < 0.001*					

Table 4. Results of Hypothesis Tests

Next, we investigated the influence of the control variables age (birth year), education ((no) university graduation), and gender (0: male; 1 = female; Note diverse or no responses were not given) on service satisfaction. Thereby, all control variables show no significant impact on service satisfaction (age: β = .056, p = .385; education: β = -.133, p = .095; gender: β = .109, p = .088). Further, the R^2 values reveal, according to Cohen (1988), a small explanation power (.02 < x < .13) of perceived humanness (R^2 = .066), a medium power (< .26) for perceived reliability (R^2 = .237) and confirmation of expectations (R^2 = .240) and a large power (> .26) for service satisfaction (R^2 = .600). All results of the hypotheses testing are summarized in Table 4, including their β-value, t-value, p-value, and the derived support.

Finally, we analyzed specific indirect paths. The indirect path of the rookie personality over perceive humanness on perceived reliability is significant (β = .128, p = .022). Furthermore, the indirect path from the rookie personality over perceived reliability on service satisfaction was also significant (β = -.069, p = .018). All other specific indirect paths from rookie personality on service satisfaction were not significant. Furthermore, a mediation analysis regarding the relation of perceived humanness and service satisfaction revealed a full mediation due to a significant total effect (β = .425, p < .001) and a non-significant direct effect (β = .036, p = .682). Thereby, the specific indirect paths over confirmation of expectations (β = .171, p = .018), perceived reliability (β = .155, p < .001), and perceived reliability and confirmation of expectations (β = .063, p = .031) were significant.

Discussion

Developers and designers of CAs strive to develop and design "good" CAs that provide great interactions and customer service. However, CA will (probably) never be perfect because of the complexity of natural language processing and human communication (Brandtzæg & Følstad, 2018; Christiansen & Kirby, 2003; McTear et al., 2016). In this context, finding ways to reduce the negative effects of errors - besides technological advancements – is of high importance. Against this background, we investigated whether the portrayal of a rookie personality is a remedy for the negative effects of errors or not.

Initially, we derived three distinct pathways by which a CA's rookie personality influences users' service satisfaction. We find support for the first pathway – i.e., portraying a rookie personality increases perceived humanness, which increases confirmation of expectation, perceived reliability, and, subsequently, service satisfaction. Our results do not support our proposed second pathway, showing that a rookie personality does not influence users' confirmation of expectations. Lastly, our data supports that a rookie personality reduces perceived reliability, which, subsequently, also reduces service satisfaction.

Theoretical Implications and Future Research

First of all, our results indicate that portraying a rookie personality has no direct influence on users' confirmation of expectations. This goes against our theorizing that a CA expressing that it is new and still learning would function as a means of expectation management, reducing users' expectations. To explain this, we would like to offer the following explanation. In general, expectations are formed before an interaction (Oliver, 1981). Thus, the messages about the rookie personality are "too late" and are perceived as part of the interaction and service. The study of Weiler et al. (2021) provides evidence for this explanation. They were able to influence the formation of expectations before an interaction with a CA via inoculation messages that informed the users about potential errors before they interacted with the CA. In this context, the question arises of how the user's perception of a CA with a rookie personality changes over time. For instance, a second interaction with a rookie chatbot would probably be perceived differently by the users. From their prior interaction, they know that the chatbot is new and still learning, thus, their expectations should be different for the second interaction. However, the second interaction could also reinforce the perception of low reliability if errors happen because users are primed to look out for such errors. In the end, further research is needed.

Our results support that a rookie personality reduces users' perceived reliability. Thus, we provide the first evidence that informing users about problems and potential errors during their interaction leads users to perceive the CAs to be of lesser quality and providing bad performance. This is different from a human-to-human interaction, where rookies are met with a more forgiving mindset (Boostrom, Jr., 2008). Thus, a rookie personality has two somewhat paradoxical effects. On the one hand, we would attribute the reduced reliability to users thinking of the CA as a machine, which is "broken" and of lesser quality. On the other hand, the rookie personality directly increases perceived humanness. We would like to provide a potential explanation for these two effects based on the dual processing theory of cognition (Kahnemann, 2011). Against this background, the rookie messages appear to influence both modes of thinking. It influences automatic cognition, increasing perceived humanness and leading to related effects (e.g., perceived humanness influencing confirmation of expectations). At the same time, the occurrence of an error leads to deliberate thinking, which concludes that the rookie messages are an additional indication of a low-quality software artifact (i.e., there was only one error, but based on the messages of the CA, more are likely to happen). For future research, we see exploring this proposition to be a valuable avenue because it would provide further insights into the potential of personalities as a remedy for negative effects caused by errors. For instance, a study could investigate the interplay of perceived humanness and the effects of the rookie personality. We would expect that adding a rookie persona to a CA that is equipped with many social cues s leads to different outcomes compared to a CA with close to no added social cues. We would expect, at some point, the perceived humanness should "unlock" the positive users' mindset from the human-to-human for the CA-to-human interaction.

Practical Implications

For practice, we see two main implications. First, because the effect of the rookie personality is positive and negative at the same time, we would advise (for now) to avoid implementing such a personality. The danger of the negative effects outweighs the positives. Second, our results indicate that users' perception of humanness has great benefits – it increases the levels of perceived reliability and confirmation of expectations – despite the presence of an error. Hence, implementing a humanlike design similar to ours, except for the rookie personality, appears to lead to a desirable outcome. Therefore, we would advise CA designers to add social cues (similar to ours) to their chatbots.

Limitations

Our study is not free of the typical limitations of experiment-based research. First, the controlled setting constitutes a limitation. Participants did not have to complete a real-world service – i.e., they had no intention to rent an e-bike. Hence, the implemented error did not affect their real life. Overall, our experiments traits realism for controllability, as all experiments do. To address this limitation, future research could engage in analyzing communication logs of real-world CAs or interviewing/surveying customers that recently interacted with a CA that produced errors. Second, the implemented set of social cues constitutes a limitation. There are nearly endless ways to select and combine social cues. We adopted a rather generic and widely used set but adding or changing social cues could affect the results. Therefore, future research should investigate other sets of social cues for a CA with a rookie personality. Third, in this study, we only considered satisfaction and not dissatisfaction, which is defined as the sense of frustration and bitterness of users who have received less than promised (Buskirk & Rothe, 1970). Thus, dissatisfaction and satisfaction are similar but not the same and can also coexist (Chen et al., 2014). In this study, we focus on satisfaction and did not consider dissatisfaction, limiting our results. Fourth, there are many different types of error a CA can produce (de Sá Siqueira et al., 2023). In this study, we focused on one type of error (not understanding user inputs), which limits our results in their transferability. For instance, another type of error (e.g., hallucinations of large language models (Dziri et al., 2022)) could lead to different results. Lastly, our participants are from one source – students of a large German university. In general, student samples are acceptable for studying the behavior of humans when interacting with technology (Compeau et al., 2012). Despite seeing only limited theoretical reasons why other populations should behave vastly differently, we would like to suggest that future research should replicate our study with other populations (e.g., from other countries).

Conclusion

CAs are probably never perfect and errors are likely to occur. In this context, research and practice try to find ways to mitigate the negative effects of these errors on users (e.g., reduced service satisfaction). Against this background, we investigate the effect a CA's rookie personality has on users' perception and service satisfaction. Our results indicate that a rookie personality is a double-edged sword. On the one hand, it increases users' perceived humanness, which leads to favorable effects, such as increased perception of reliability. On the other hand, it directly negatively influences users' evaluation of the CA's reliability, which reduces users* service satisfaction. To explain these somewhat paradoxical effects, were refer to the dual processing theory of cognition and propose that the rookie personality influences automatic and deliberate thinking. Users are actively and deliberately thinking about the CA's messages expressing the rookie personality ("I am new and still learning. I will do my best, but errors can happen"), which leads them to the perception of a "broken" and low-quality software artifact. Also, users' automatic thinking is influenced by the rookie personality, leading to higher levels of perceived humanness, which also influences users thinking. For future research, we see detangling these two effects as an important research area. Understanding how and when a CA's rookie personality influences users' automatic and deliberate thinking has the potential to bring forth a potent remedy for the negative effects of errors.

References

Adam, M., Roethke, K., & Benlian, A. (2022). Human Versus Automated Sales Agents: How and Why Customer Responses Shift Across Sales Stages. *Information Systems Research*, *Articles in Advance*, 1–21.

Ahmad, R., Siemon, D., Gnewuch, U., & Robra-Bissantz, S. (2022). A Framework of Personality Cues for Conversational Agents. *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS).*

Angeli, A. De, & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, *20*(3), 302–310.

Antonio, R., Tyandra, N., Nusantara, L. T., Anderies, & Agung Santoso Gunawan, A. (2022). Study Literature Review: Discovering the Effect of Chatbot Implementation in E-commerce Customer Service System Towards Customer Satisfaction. *International Seminar on Application for Technology of Information and Communication (ISemantic)*, 296–301.

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.

Babel, F., Kraus, J., Hock, P., Asenbauer, H., & Baumann, M. (2021). Investigating the validity of online robot evaluations: Comparison of findings from an one-sample online and laboratory study. *ACM/IEEE International Conference on Human-Robot Interaction*, 116–120.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*, 74–94.

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.
Benner, D., Elshan, E., Schöbel, S., & Janson, A. (2021). What do you mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents. *Proceedings of the 42nd International Conference on Information Systems (ICIS)*.
Berri, D. J., Brook, S. L., & Fenn, A. J. (2011). From college to the pros: Predicting the NBA amateur player draft. *Journal of Productivity Analysis*, *35*, 25–35.
Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, *25*(3), 351–370.
Blanchette, I., & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. *Cognition and Emotion*, *24*(4), 561–595.
Boostrom, Jr., R. E. (2008). The Social Construction of Virtual Reality and the Stigmatized Identity of the Newbie. *Journal For Virtual Worlds Research*, *1*(2).
Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly*, *25*(1), 1–16.
Boulding, W., Kalra, A., Staelin, R., & Zeithaml, V. A. (1993). A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions. *Journal of Marketing Research*, *30*, 7–27.
Brandtzæg, P. B., & Følstad, A. (2018). Chatbots: Changing User Needs and Motivations. *Interactions*, *25*(5), 38–43.
Brendel, A. B., Greve, M., Diederich, S., Bührke, J., & Kolbe, L. M. (2020). "You are an idiot!" - How conversational agent communication patterns influence frustration and harassment. *Proceedings of the 26th Americas Conference on Information Systems (AMCIS)*.
Buck, C., & Dinev, T. (2020). Low effort and privacy - How textual priming affects privacy concerns of email service users. *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*.
Bührke, J., Brendel, A. B., Lichtenberg, S., Greve, M., & Mirbabaie, M. (2021). Is Making Mistakes Human? On the Perception of Typing Errors in Chatbot Communication. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*.
Buskirk, R. H., & Rothe, J. T. (1970). Consumerism. An Interpretation. *Journal of Marketing*, *34*(4), 61–65.
Chen, A., Lu, Y., Gupta, S., & Xiaolin, Q. (2014). Can customer satisfaction and dissatisfaction coexist? An issue of telecommunication service in China. *Journal of Information Technology*, *29*(3), 187–267.
Chin, W. W. (1998). The Partial Least Squares Approach to Structural Equation Modelling. In G. A. Marcoulides (Ed.), *Modern Methods for Business Research* (pp. 295–336). Lawrence Erlbaum Associates Publishers.
Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, *7*(7), 300–307.
Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. In *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.
Compeau, D., Marcolin, B., Kelley, H., & Higgins, C. (2012). Generalizability of information systems research using student subjects A reflection on our practices and recommendations for future research. *Information Systems Research*, *23*(4), 1093–1109.
Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, *78*(1), 98–104.
Coye, R. W. (2004). Managing customer expectations in the service encounter. *International Journal of Service Industry Management*, *15*(1), 54–71.
Daryanto, A., Alexander, N., & Kartika, G. (2022). The anthropomorphic brand logo and its effect on perceived functional performance. *Journal of Brand Management*, *29*, 287–300.
de Sá Siqueira, M. A., Müller, B. C. N., & Bosse, T. (2023). When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *International Journal of Human-Computer Interaction*, *Online Fir*.
Dennis, A. R., & Valacich, J. S. (2001). Conducting Experimental Research in Information Systems. *Communications of the Association for Information Systems*, *7*(5).
Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *Journal of the Association for Information Systems*, *23*(1), 96–138.
Diederich, S., Lembcke, T.-B., 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.
Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.
Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies*, *132*(12), 138–161.
Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the New World of HCI. *Interactions*, *24*(4), 38–42.
Fombelle, P. W., Bone, S. A., & Lemon, K. N. (2016). Responding to the 98%: Face-enhancing strategies for dealing with rejected customer ideas. *Journal of the Academy of Marketing Science*, *44*(6), 685–706.
Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50.
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*(1), 91–109.
Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, *21*(4), 389–400.
Gnewuch, U., Adam, M. T. P., Morana, S., & Maedche, A. (2018). “The Chatbot is typing ...” - The Role of Typing Indicators in Human-Chatbot Interaction. *Proceedings of the 17th Annual Pre-ICIS Workshop on HCI Research in MIS*, 1–5.
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. *Proceedings of the 26th European Conference on Information Systems (ECIS)*, 1–17.
Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. *Proceedings of the 38th International Conference on Information Systems (ICIS)*.
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.
Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, *144*, 113515.
Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Education.
Jiang, J. J., Klein, G., & Carr, C. L. (2002). Measuring Information System Service Quality: SERVQUAL from the Other Side. *MIS Quarterly*, *26*(2), 145–166.
Kahnemann, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
Karimi, F., Poo, D. C. C., & Tan, Y. M. (2015). Clinical information systems end user satisfaction: The expectations and needs congruencies effects. *Journal of Biomedical Informatics*, *53*, 342–354.
Kettinger, W. J., & Lee, C. C. (1994). Perceived Service Quality and User Satisfaction with the Information Services Function. *Decision Sciences*, *25*(5–6), 737–766.
Kim, A. (2011). *How Apple Approached Developing Siri’s Personality*. https://www.macrumors.com/2011/10/15/how-apple-approached-developing-siris-personality/
Kim, H., Lee, G., Lim, Y. K., Koh, D. Y., & Park, J. M. (2019). Designing personalities of conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, 1–6.
Kirakowski, J., O’Donnell, P., & Yiu, A. (2007). The perception of artificial intelligence as “human” by computer users. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 376–384.
Korda, A. P., & Snoj, B. (2010). Development, Validity and Reliability of Perceived Service Quality in Retail Banking and its Relationship With Perceived Value and Customer Satisfaction. *Managing Global Transitions*, *8*(2), 187–205.
Kunda, Z. (1999). *Social Cognition: Making Sense of People*. The MIT Press.
Lang, H., Seufert, T., Klepsch, M., Minker, W., & Nothdurft, F. (2013). Are Computers Still Social Actors? *Conference on Human Factors in Computing Systems - Proceedings*, 859–864.
Larivière, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., Wünderlich, N. V., & De

Keyser, A. (2017). “Service Encounter 2.0”: An investigation into the roles of technology, employees and customers. *Journal of Business Research*, *79*, 238–246.
Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823.
Lessio, N., & Morris, A. (2020). Toward Design Archetypes for Conversational Agent Personality. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*.
Lester, J., Branting, K., & Mott, B. (2004). Conversational agents. In *The Practical Handbook of Internet Computing* (pp. 1–17).
Levinson, S. C. (1995). Interactional biases in human thinking. In E. N. Goody (Ed.), *Social Intelligence and Interaction* (pp. 221–260). Cambridge University Press.
Li, X., Xie, S., Ye, Z., Ma, S., & Yu, G. (2022). Investigating Patients’ Continuance Intention Toward Conversational Agents in Outpatient Departments: Cross-sectional Field Survey. *Journal of Medical Internet Research*, *24*(11).
Liao, Q. V., Hussain, M. M. U., Chandar, P., Davis, M., Khazaen, Y., Crasso, M. P., Wang, D., Muller, M., Shami, N. S., & Geyer, W. (2018). All work and no play? Conversations with a question-and-answer chatbot in the wild. *Conference on Human Factors in Computing Systems - Proceedings*.
Liu, D., Lv, Y., & Huang, W. (2023). How do consumers react to chatbots’ humorous emojis in service failures. *Technology in Society*, *73*, 102244.
Mardsen, R. (2015). *When Siri is spot on with cultural references – who makes it happen, man or machine?* https://www.independent.co.uk/tech/when-siri-says-something-amusing-who-makes-it-happen-man-or-machine-10394914.html
Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. *Accident Analysis and Prevention*, *35*(5), 683–691.
McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *International Workshop on Future and Emerging Trends in Language Technology*, 38–49.
McTear, M. F., Callejas, Z., & Griol, D. (2016). Conversational Interfaces: Past and Present. In *The Conversational Interface* (pp. 51–72). Springer.
Meyer-Waarden, L., Pavone, G., Poocharoentou, T., Prayatsup, P., Ratinaud, M., Tison, A., & Torné, S. (2020). How Service Quality Influences Customer Acceptance and Usage of Chatbots? *Journal of Service Management Research*, *4*(1), 35–51.
Millán, Á., & Esteban, Á. (2004). Development of a multiple-item scale for measuring customer satisfaction in travel agencies services. *Tourism Management*, *25*(5), 533–546.
Miller, M. K., & Peake, J. S. (2010). Rookie or Rock Star? Newspaper Coverage of Sarah Palin’s Vice Presidential Campaign. *APSA 2010 Annual Meeting Paper*.
Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers Robotics AI*, *4*(21), 1–23.
Murphy, J., Glatzel, U., & Hofacker, C. (2017). Service robots in hospitality and tourism: investigating anthropomorphism. *15th APacCHRIE Conference (Vol. 31)*.
Nam, K., Baker, J., Ahmad, N., & Goo, J. (2020). Determinants of writing positive and negative electronic word-of-mouth: Empirical evidence for two types of expectation confirmation. *Decision Support Systems*, *129*, 113168.
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues: A Journal of the Society for the Psychological Studies of Social Issues*, *56*(1), 81–103.
Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 72–78.
Newell, D. K. (1983). Ten Survival Suggestions For Rookie Law Teachers. *Journal of Legal Education*, *33*(4), 393–703.
Nunally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill.
Oliver, R. L. (1981). Measurement and Evaluation of Satisfaction Processes in Retail Settings. *Journal of Retailing*, *57*(3), 25–48.
Oliver, R. L. (1997). Satisfaction: A Behavioral Perspective on the Consumer. In *McGraw-Hill series*.
Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072.
Pradhan, A., & Lazar, A. (2021). Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. *CUI 2021 - 3rd Conference on Conversational User Interfaces (CUI ’21)*.

Riquel, J., Brendel, A. B., Hildebrandt, F., Greve, M., & Dennis, A. R. (2021). "F*** You !" – An Investigation of Humanness, Frustration, and Aggression in Conversational Agent Communication. *Proceedings of the 42nd International Conference on Information Systems (ICIS)*, 1–16.

Riquel, J., Brendel, A. B., Hildebrandt, F., Greve, M., & Kolbe, L. M. (2021). "Even the Wisest Machine Makes Errors" – An Experimental Investigation of Human-like Designed and Flawed Conversational Agents. *Proceedings of the 42nd International Conference on Information Systems (ICIS)*, 1–16.

Schuetzler, R. M., Grimes, G. M., & Giboney, J. S. (2018). An investigation of conversational agent relevance, presence, and engagement. *Proceedings of the 24th Americas Conference on Information Systems (AMCIS)*, 1–12.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2018). Designing Anthropomorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework. *Proceedings of the 39th International Conference on Information Systems (ICIS)*, 1–17.

Seering, J., Luria, M., Ye, C., Kaufman, G., & Hammer, J. (2020). It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. *Conference on Human Factors in Computing Systems - Proceedings*, 1–13.

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, *115*, 14–24.

Shum, H. yeung, He, X. dong, & Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. In *Frontiers of Information Technology and Electronic Engineering*.

Solow, J., & von Allmen, P. (2016). Performance expectations, contracts and job security. In *Research Handbook of Employment Relations in Sport* (pp. 46–68). Business 2016.

Song, M., Zhang, H., Xing, X., & Duan, Y. (2023). Appreciation vs. apology: Research on the influence mechanism of chatbot service recovery based on politeness theory. *Journal of Retailing and Consumer Services*, *73*, 103323.

Sonlu, S., Güdükbay, U., & Durupinar, F. (2021). A Conversational Agent Framework with Multi-modal Personality Expression. *ACM Transactions on Graphics*, *40*(1), 1–16.

Stone-Romero, E. F., Stone, D. L., & Grewal, D. (1997). Development of a multidimensional measure of perceived product quality. *Journal of Quality Management*, *2*(1), 87–111.

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.

Vu, T. L., Tun, K. Z., Eng-Siong, C., & Banchs, R. E. (2021). Online FAQ Chatbot for Customer Support. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction* (pp. 251–259).

Wagner, K., & Schramm-Klein, H. (2019). Alexa, are you human? Investigating the anthropomorphism of digital voice assistants - A qualitative approach. *Proceedings of the 40th International Conference on Information Systems (ICIS)*.

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.

Weiler, S., Matt, C., & Hess, T. (2021). Immunizing with information–Inoculation messages against conversational agents' response failures. *Electronic Markets*, *32*(2), 239–258.

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Wu, I. L., Chiu, M. L., & Chen, K. W. (2020). Defining the determinants of online impulse buying through a shopping process of integrating perceived risk, expectation-confirmation model, and flow theory issues. *International Journal of Information Management*, *52*, 102099.

Yuan, L. I., & Dennis, A. R. (2019). Acting Like Humans? Anthropomorphism and Consumer's Willingness to Pay in Electronic Commerce. *Journal of Management Information Systems*, *36*(2), 217–246.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1988). Communication and Control Processes in the Delivery of Service Quality. *Journal of Marketing*, *52*(2), 35–48.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1993). The nature and determinants of customer expectations of service. *Journal of the Academy of Marketing Science*, *21*(1), 1–12.

Zemčík, T. (2021). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI and Society*, *36*, 361–367.

Zhu, H., Wong, N., & Huang, M. (2019). Does relationship matter? How social distance influences perceptions of responsibility on anthropomorphized environmental objects and conservation intentions. *Journal of Business Research*, *95*, 62–70.