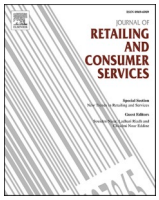
Journal of Retailing and Consumer Services 84 (2025) 104257

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser



Recovering customer satisfaction after a chatbot service failure – The effect of gender

Alexandra Rese *, Lennart Witthohn

Marketing & Innovation, University of Bayreuth, Universitätsstraße 30, 95447, Bayreuth, Germany

ARTICLE INFO

Handling editor: H. Timmermans

Keywords:
Chatbots
Service recovery
Gender effects
Gender matching
Perceived humanness

ABSTRACT

Chatbots in customer service often fail to meet customer expectations, largely because they are considered prone to comprehension errors. Service recovery can decisively restore perceived humanness and user satisfaction through perceived warmth and competence after a service failure. In this study, we investigate the effect of the chatbot's gender on the user in service recovery. The majority of chatbots in customer service display female characteristics. We use a pre-study (n = 30) to determine the perceived gender of several chatbot avatars and a scenario-based experiment (n = 300) in which the service recovery after an outcome failure and the gender of the chatbot are manipulated. The results show that the service recovery significantly improved user satisfaction with the chatbot. In addition, the chatbot was perceived as significantly warmer and more competent, which resulted in higher perceived humanness and increased user satisfaction. Male chatbots were perceived as less warm in failure situations when service recovery was not achieved. However, following service recovery, there are no differences in the perception of the chatbot's warmth and gender. Perceived warmth is correlated with perceived competence. Gender incongruence between the chatbot and the respondent resulted in a higher perceived humanness of the chatbot in service recovery. Therefore, firms should pay particular attention to the contexts in which chatbots are used and whether gender matching is appropriate.

1. Introduction

When, in November 2022, OpenAI publicized that ChatGPT would be made available to the general public for testing (Open AI, 2023), the announcement excited worldwide hype over the transformative potential of chatbots. Two months after the announcement, ChatGPT had already reached 100 million active users (Hu, 2023). However, users reported problems with ChatGPT – for example, the large language model was giving wrong, fabricated, and nonsensical answers (hallucination) or using parts or patterns of the training data (stochastic parrots) (Li, 2023; Shaier et al., 2023). Even the less intelligent chatbots in customer service were often unable to fulfil the expectations placed on them (Sheehan et al., 2020) and were considered prone to errors (Adam et al., 2021; Tran et al., 2021). They often did not understand customer inquiries (Huang and Dootson, 2022), made incoherent statements (Coniam, 2014), and did not follow the logic of human conversations (Caldarini et al., 2022). Consequently, of 103 chatbots in practice from different application areas and various countries representing 10% of the database "chatbots.org", 53 were discontinued after 15 months in a period from May 2019 until September 2020 (Janssen et al., 2021).

On the other hand, chatbots have great *potential.* The value of the chatbot market is forecast to be worth 20.81 billion dollars by 2029 (Mordor Intelligence, 2024). Companies can increase their market capitalisation by implementing a chatbot in customer service (Fotheringham and Wiles, 2023). In particular, generative AI chatbots are expected to increase productivity by about 15–40 percent (Chui et al., 2023). Many companies have already made use of chatbots to increase efficiency in their customer service (Shin et al., 2023). They help companies to reduce costs (Adamopoulou and Moussiades, 2020) and relieve employees of routine inquiries (Kaczorowska-Spychalska, 2019). Users of chatbots benefit from the fact that they can be reached at any time (Chung et al., 2020), an immediate response can be expected (Tran et al., 2021), and cognitive effort can be saved (Kaczorowska-Spychalska, 2019). Furthermore, chatbots are more entertaining (Adamopoulou and Moussiades, 2020) and significantly more interactive than FAQ pages (Caldarini et al., 2022). A successful chatbot service can increase customer satisfaction and loyalty to the company (Jenneboer et al., 2022). However, if the performance of the

This article is part of a special issue entitled: Technology in Retailing published in Journal of Retailing and Consumer Services.

* Corresponding author.
E-mail address: alexandra.rese@uni-bayreuth.de (A. Rese).

https://doi.org/10.1016/j.jretconser.2025.104257
Received 6 November 2024; Received in revised form 6 February 2025; Accepted 7 February 2025
Available online 13 February 2025

chatbot falls short of the user's expectations, this leads to a reduced willingness to use the chatbot again (Sheehan et al., 2020), a loss of perceived competence (Toader et al., 2020) and perceived humanity, and a greater feeling of discomfort among users (Diederich et al., 2021). In addition, poor service from a chatbot can hurt the company due to lower purchase intentions, lower trust, and lower service satisfaction (Toader et al., 2020). This points to the important need for companies to enable their chatbots to restore satisfaction following poor service. Existing studies have shown that service recovery, contrary to service failure, increases the perceived warmth (Gelbrich et al., 2021; Huang and Ha, 2020) and competence of chatbots (Han et al., 2022; Toader et al., 2020; Liu et al., 2022; Yang et al., 2023), as well as customer satisfaction (Zhu et al., 2023). However, research on restoring user satisfaction following a chatbot's poor performance is still limited (Zhu et al., 2023), and researchers have been called on to transfer established theories from service research to the chatbot context (Blut et al., 2021; Grégoire and Mattila, 2021).

In this paper, we focus on the effect of chatbot gender on service failure and recovery. Chatbots are often equipped with anthropomorphic design features, such as a name (Zheng et al., 2023), an avatar (Pizzi et al., 2023), and a human communication style (Go and Sundar, 2019; Lu et al., 2024), so that users are able to humanise the chatbots (Seeger et al., 2018). Recently, new generative AI tools have been introduced that can design even more human-like chatbots (Ma and Huo, 2024). The appearance and behaviour of chatbots can increase their perceived competence and social presence (Xie et al., 2024). On the other hand, increased warmth can evoke negative attitudes towards chatbots (Kim et al., 2019). In this respect, a closer examination of gender in service recovery is required for its effects on the two dimensions of social cognition but also perceived humanness and satisfaction. This is particularly important because the majority of chatbots in customer service are equipped with female characteristics (Feine et al., 2020). Female chatbots are perceived as warmer (Borau et al., 2021) and more authentic (Esmark Jones et al., 2022) and are more likely to be forgiven after a mistake (Toader et al., 2020). Therefore, adopting a female chatbot as a baseline appearance is a reasonable strategy for firms. However, firms can easily and cheaply alternate between male, gender-neutral, and female chatbot representations. Against the current gender norms in society and depending on consumers' gender, there may be configurations in which the female chatbot is outperformed. For example, the use of gender-neutral chatbots might be a strategic choice to circumvent gender preferences in times of increasing criticism of feminized service stereotypes (Aumüller et al., 2024) and socio-political movements that raise awareness of gender issues (Cammarota et al., 2023). The following two research questions can be derived from this: *(1) What part does chatbot gender play following poor service performance and service recovery? (2) Is gender matching or mismatching of user and chatbot more important?*

To answer the research questions, we conducted a scenario-based experiment in which we manipulated both the service outcome (recovery–failure) and the gender of the chatbot (perceived female, neutral, male). For this purpose, we use social cognition (van Doorn et al., 2017) and the stereotype content model (SCM) (Fiske et al., 2002). We draw on a sample of 300 German respondents to investigate the effect of a chatbot service failure on the consumer's satisfaction with the chatbot and the effect of chatbot gender and its (mis)match with the user on the failure and recovery outcome.

Our results show the importance of a successful service recovery over anthropomorphic design elements. Perceived competence of the chatbot is more important than perceived warmth and, similarly, it is high for the three chatbot gender types with no effects of gender (mis)matching. A gender mismatch can increase perceived humanness and can point to a preference for a mismatch in service recovery. Our study adds to the literature on the role of gender in chatbot service failure (Liang et al., 2024; Toader et al., 2020). It also clarifies the issues of service failure and recovery through online chat in general (Esmark Jones et al., 2022; Huang et al., 2024) and with reference to satisfaction in particular (Hsu and Lin, 2023). We expand existing research with new insights into the importance of the two dimensions of social cognition – gender matching and the use of gender-neutral chatbots.

2. Theoretical background

2.1. Chatbots in service

Chatbots communicate with their users using text or voice to solve queries (Crolic et al., 2022). They simulate human conversations (Luo et al., 2019), which they can conduct with thousands of users simultaneously (Caldarini et al., 2022). They are, therefore, particularly suitable for customer service (Sheehan et al., 2020). Chatbots are based on artificial intelligence (AI) and use natural language processing (NLP) algorithms to understand text input from users and respond to them (Hoyer et al., 2020). NLP algorithms are used to identify the request and derive a task for the chatbot from the user's input. The response of the chatbot is based on a rule-based, retrieval-based, or generative model (Adamopoulou and Moussiades, 2020). An interaction strategy is then selected based on the user's input as to how the chatbot should respond and process possible follow-up questions from the user (Suta et al., 2020). Strategies include determining the conversation leader and error handling and confirmation (Cahn, 2017). The chatbot uses a knowledge database or the internet to answer the inquiry. With the help of natural language generation (NLG), the chatbot responds to the user in natural human language (Adamopoulou and Moussiades, 2020). The algorithm is trained and continuously improved through interaction with users (Hoyer et al., 2020). Thanks to deep learning algorithms, the chatbot can learn to adapt its language to emotional customers (Suta et al., 2020). Recent improvements in NLP have made chatbots increasingly easy to implement and maintain and even better at imitating human conversations (Caldarini et al., 2022).

2.2. Service failure and service recovery – basic definitions

A *service failure* is a service performance that falls short of the customer's expectations or the acceptable customer service level (Holloway and Beatty, 2003). A perceived performance that is below expectations leads to lower customer satisfaction based on the confirmation–disconfirmation paradigm and is referred to as negative disconfirmation (Oliver, 1977). This is in line with the theory of expectancy violation (e.g., Crolic et al., 2022) or similar to negative expectations disconfirmation (e.g., Morgeson et al., 2020; Smith et al., 1999). The lower customer satisfaction caused by service failure results in lower customer loyalty (van Vaerenbergh et al., 2014) and can lead to customers churning and speaking negatively about the company (Bitner et al., 2000). Consequently, poor service can have a significant negative financial impact (Holloway and Beatty, 2003). Service failure can also lead to further significant costs because the service provider has to redo the service or compensate the customer (Bitner et al., 2000). In addition, disappointment in customer expectations can lead to emotional reactions, such as anger (Bougie et al., 2003) and aggression (Huang and Dootson, 2022), resulting in a poorer evaluation of the company (Mattila and Enz, 2002). These negative emotions can only be offset by monetary compensation (Valentini et al., 2020). Consequently, it is necessary to develop strategies to restore customer satisfaction after a service failure (Kelley et al., 1993).

An organization's response to a customer's perceived lack of service is referred to as *service recovery* (Holloway and Beatty, 2003; Kelley and Davis, 1994). Customers expect an effective response to an unsatisfactory state of affairs (Holloway and Beatty, 2003). These expectations depend on the severity of the failure (Hess Jr. et al., 2003; Miller et al., 2000). The expectations of service recovery are higher if the service quality is rated as high, if the customers are loyal to the service provider (Kelley and Davis, 1994), or if there is a service guarantee (Miller et al.,

2000). If the company removes the faulty servicing satisfactorily, the probability of retaining the customer increases, and customer loyalty and customer satisfaction can be restored (Miller et al., 2000). Paradoxically, customer satisfaction can be higher than before the service failure (Matos et al., 2007). This phenomenon is known as the service recovery paradox (Magnini et al., 2007) and leads to the conclusion that failures in service delivery present an opportunity for companies to build long-term customer relationships (Kelley et al., 1993). However, the meta-analysis by Matos et al. (2007) shows that the service recovery paradox is not transferable to the repurchase intention. It is more likely to occur in the case of errors with a low degree of severity, and it is significantly less likely following a second error (Magnini et al., 2007). Furthermore, the majority of dissatisfied customers do not complain (McCollough et al., 2000). Service recovery is, therefore, an economic necessity for companies to retain customers dissatisfied on the first occasion (Morgeson et al., 2020).

2.3. Chatbots and service failure and recovery

If the service is poor, chatbots appear less human. This leads to lower satisfaction with the chatbot and reduces customer willingness to use the chatbot again (Diederich et al., 2021; Sheehan et al., 2020). Moreover, they can appear more uncanny (Diederich et al., 2021), which may trigger negative emotions in the user, reduce trust in the chatbot, and lead to diminished loyalty towards the chatbot (Rajaobelina et al., 2021).

The negative emotions caused by expectancy violations, such as anger (Crolic et al., 2022) and aggression (Huang and Dootson, 2022) can be avoided by the skilful design of the chatbot. The pre-encounter expectations with the chatbot should be low key. Even a design that incorporates a few anthropomorphic features can help (Crolic et al., 2022). Similarly, an early disclosure of the availability of a human employee in the case of a chatbot service failure leads to a less emotional reaction than a late disclosure (Huang and Dootson, 2022). If the chatbot annoys customers, their satisfaction decreases, and company evaluation and purchase intentions go into decline (Belanche et al., 2020; Crolic et al., 2022). Due to these negative consequences of service failure with chatbots, good service recovery is necessary.

For chatbots, classic service recovery strategies, such as an apology or compensation, work less well (Mattila et al., 2011). Instead, users expect an immediate solution to the problem as they engage with the chatbot (Fiore et al., 2019). Repair strategies for chatbots are based on communication theories (Ashktorab et al., 2019) – in particular, the framework grounding in communication (Clark and Brennan, 2004). A conversation is regarded as a collective action to build a shared understanding. If this common understanding cannot be established due to incoherent statements, the conversation partner tries to repair the conversation. These repair strategies support the chatbot in task completion (Ashktorab et al., 2019). Several repair strategies have been identified in the literature, such as repeating the request and asking users to rephrase their question (out-of-vocabulary explanation). These are often used in combination (Ashktorab et al., 2019; Benner et al., 2021).

2.4. Literature review of service recovery for chatbots

We searched the literature and scientific databases using the keywords “chatbot AND service recovery” and extracted fifteen articles (see Table A1 in the appendix). Six articles analyse the fairness dimensions in the chatbot context. In the service literature, the theory of justice is often applied in the context of service failure and service recovery (McCollough et al., 2000; Smith et al., 1999; Wirtz and Mattila, 2004). Three dimensions of justice influence satisfaction with service recovery (Del Río-Lanza et al., 2009; Wirtz and Mccoll-Kennedy, 2010): distributive, procedural, and interactional justice. Four articles deal with emotions in the service recovery process. Cute chatbot designs, self-deprecating humour responses, humour and informal language, and humorous emojis (Liu et al., 2023) are investigated, looking at the mitigating effect on users’ negative emotions. Humour can even work better in the service recovery process than an apology or compensation (Kobel and Groeppel-Klein, 2021). Two papers examine the handover of a service failure to a human employee. While chatbots can restore satisfaction independently (Song et al., 2022), this depends on the nature of the failure – technical problems or failure to deliver the service (Xing et al., 2022). Finally, two papers evaluate different repair strategies, and one applies attribution theory to chatbots. Concerning the latter, a chatbot service failure leads customers to blame the company (Belanche et al., 2020; Merkle, 2019). However, anthropomorphic design elements can help to reduce these negative effects on the company and support problem-oriented coping strategies (Pavone et al., 2023).

One concept that is frequently used in the chatbot literature is perceived warmth and competence (Borau et al., 2021; Kull et al., 2021; Pizzi et al., 2023; Roy and Naidoo, 2021; Seiler and Schär, 2021; van Doorn et al., 2017) as constructs of social cognition (van Doorn et al., 2017). In the chatbot service recovery literature, only Han et al. (2022) and Zhou and Chang (2024) consider competence and warmth together, so further research is needed here.

2.5. Hypothesis development

Service recovery aims to restore customer satisfaction (Michel et al., 2009), which can be achieved after a successful service recovery (Miller et al., 2000). Therefore, the central dependent variable in this study is satisfaction with the chatbot. Moreover, service recovery has a positive effect on satisfaction in the chatbot context and can work just as well as immediate recovery by a human employee (Zhu et al., 2023). Satisfaction after a service failure is even higher if the chatbot uses a politeness strategy (apology, appreciation) instead of none when initiating a service recovery (Song et al., 2023). We, therefore, formulate the following hypothesis.

H1. Satisfaction with the chatbot is significantly higher in service recovery than in service failure.

Anthropomorphism is the attribution of human characteristics to non-human actors or objects. This occurs, for example, when chatbots activate knowledge about humans in their users (Epley et al., 2007). The more humanlike that chatbot avatars are designed, the more likely they are to be anthropomorphized because they possess greater similarities with the users (Epley et al., 2007). Therefore, a highly anthropomorphic chatbot design can positively influence perceived humanity (Sheehan et al., 2020). Other types of determinants of anthropomorphism are the need to build social connections with other people and the need to effectively interact with the environment (Epley et al., 2007). They lead to non-human actors being anthropomorphized due to a reduction in uncertainty since the behaviour of the non-human actor is more predictable and, thus, the interaction may be more favourable (Sheehan et al., 2020).

The more humanlike the chatbot is perceived to be, the greater the satisfaction with its use (Blut et al., 2021; Diederich et al., 2021; Söderlund and Oikarinen, 2021) and the greater the customer’s willingness to use the chatbot again (Blut et al., 2021). However, if the chatbot makes mistakes, it appears less humanlike (Diederich et al., 2021; Sheehan et al., 2020). In (online) conversation, a set of pragmatic cues is expected (Grice, 1975; Jacquet et al., 2018, 2019). Their violation can result in longer response times and lower perceived humanness (Jacquet et al., 2018, 2019). According to the theory of the uncanny valley (Mori et al., 2012, p. 98), a failure of the chatbot to respond with a meaningful response indicates the chatbot’s inability “to attain, a lifelike appearance” and results in an abrupt shift of the user’s attention “from empathy to revulsion”. According to Dietvorst et al. (2015), people lose trust in algorithms more quickly than in humans, even if humans make greater mistakes. This leads us to the following hypotheses.

H2. The chatbot is perceived as significantly more human in service recovery than in service failure.

H3. The more the chatbot is perceived as human, the more users are satisfied with the chatbot.

(Perceived) warmth and competence are the two central dimensions of social cognition and explain how people or groups are judged (Fiske et al., 2007). Taking an evolutionary perspective, Fiske et al. (2007) explain that the judgment of warmth – that is, other people's perceived intentions, trustworthiness, sincerity, friendliness, or helpfulness – precedes the judgment of competence – namely, other people's abilities and competencies. The two dimensions are also applied to non-human actors – for example, the perception of brands and organizations (Aaker et al., 2010, 2012).

When it comes to service failure and recovery, higher perceived warmth leads to higher satisfaction with recovery by a human employee (Alhouti et al., 2019; Smith et al., 2016) and a better evaluation of other loyalty intentions (Bolton and Mattila, 2015). Even after a service failure by a digital assistant, its perceived warmth increases customer satisfaction (Gelbrich et al., 2021). Moreover, the perceived competence of human employees has a positive effect on satisfaction with the service – in particular, on transactional aspects such as purchase intention – whereas for warmth this applies to relational aspects such as customer attachment to and identification with the company (Güntürkün et al., 2020). Furthermore, higher perceived competence results in customer persistence in using the service and affects positive word of mouth (Blodgett et al., 1995).

Robotics research shows that, for more human-like robots, perceived warmth is higher, and an apology can restore a service failure due to increasing perceived warmth and consequent satisfaction (Choi et al., 2021). For chatbots, Han et al. (2022) demonstrate that an empathetic response after a service failure increases both perceived warmth and (to a lesser extent) competence, which in turn has a positive influence on service quality and on ultimate satisfaction. While perceived competence positively affects trust in a chatbot, a service failure leads to a poorer evaluation of perceived competence (Toader et al., 2020). Dimensions associated with competence, such as the perceived intelligence of a chatbot (Fiske et al., 2002), also positively increase service recovery satisfaction and reuse intention (Liu et al., 2023; Yang et al., 2023). Overall, a warm response to a complaint results in higher satisfaction with the complaint handling (service recovery) than a competent response (Huang and Ha, 2020). However, the relationship orientation towards the company is important, and customers with an exchange orientation prefer a competent response (Huang and Ha, 2020). Consequently, we derive the following hypotheses.

H4. The chatbot is perceived as significantly warmer in service recovery than in service failure.

H5. The chatbot is perceived as significantly more competent in service recovery than in service failure.

The high perceived warmth and competence of a chatbot leads to a more positive chatbot-related attitude (Maar et al., 2023). Belanche et al. (2021) operationalized the humanness of robots with the three dimensions of human likeness, competence, and warmth. The degree of human likeness – for example, a (low/high) anthropomorphic design, – is often conceptualized as a prerequisite that activates perceived warmth and competence (Yang et al., 2020) not only in the robot context (Akdim et al., 2023; Choi et al., 2021; Kim et al., 2019) but also in the chatbot context (Pizzi et al., 2023). However, warmth and competence dimensions are used as important characteristics of perceptions of humanness (Alaei et al., 2022; Heflick et al., 2011). Söderlund (2021, p.17) defines perceived humanness as "the extent to which an individual is seen as having characteristics that are typical for humans". For robots, Söderlund (2021) has positively linked perceived warmth in a conversation with a robot to perceived humanness. For chatbots, perceived intelligence or expertise is often related to conversational competence displayed in the textual properties of the shared content (Laban, 2021). Schuetzler et al. (2020) found a positive effect of conversational competence in terms of tailored responses on perceived humanness. For these reasons, we suggest the following hypotheses.

H6. The higher the perceived warmth of the chatbot, the more human it is perceived to be.

H7. The higher the perceived competence of the chatbot, the more human it is perceived to be.

According to the stereotype content model (SCM), people evaluate social groups differently in terms of warmth and competence by applying stereotypes (Fiske et al., 2002). For the U.S. and also for Germany, there are significant gender differences concerning both dimensions, with women being perceived as warmer and men as more competent (Diekman and Eagly, 2000; Eagly and Steffen, 1984; Ebert et al., 2014). No self-favouritism was found for warmth and men, but both genders rated their own gender as more competent (Ebert et al., 2014). Recently, Alaei et al. (2022) confirmed gender stereotypes showing that, for females, more attractive face photos were perceived as more human whereas, for males, more intelligent faces were seen as more human. In robotics research, male robots are perceived as more intelligent and female robots as more social and collaborative (Eyssel and Häring, 2012) but also warmer (Stroessner and Benitez, 2019). However, these effects depend on the context because consumers associate product groups or services with gender (Fugate and Phillips, 2010; Roesler et al., 2022). For example, male robots are preferred for male tasks and female robots for female-associated tasks (Eyssel and Häring, 2012; Kuchenbrandt et al., 2014).

Moreover, there is a preference for gender-congruent chatbots in the chatbot context (Beldad et al., 2016; McDonnell and Baxter, 2019). Female chatbots are perceived as more authentic (Esmark Jones et al., 2022), more human (Borau et al., 2021), and warmer (Ahn et al., 2022; Borau et al., 2021). In terms of competence, there is no significant difference between male and female chatbot avatars (Borau et al., 2021; Toader et al., 2020). However, the female avatar was assigned a higher perceived competence in the study by Toader et al. (2020), contrary to the theory of SCM. Furthermore, female avatars seem to be more likely to be forgiven after a service failure (Toader et al., 2020). This is also the case for human employees using new service technologies. Customer satisfaction and revisit intention were higher in a service failure context for female service persons, but lower for males in a service success context (Wu et al., 2015). We therefore formulate the following hypotheses.

H8. The female chatbot is perceived as significantly warmer than the male chatbot.

H9. The effect of service outcome on warmth is moderated by the gender of the chatbot so that the perceived warmth of the female chatbot is higher than the male chatbot following a service failure.

H10. The effect of service outcome on humanness is moderated by the gender of the chatbot so that the perceived humanness of the female chatbot is higher than the male chatbot following a service failure.

Ebert et al. (2014) observed that the user's own gender is perceived as more competent. Similarly, Zogaj et al. (2023) found that matching the gender of the user and chatbot increases the perceived similarity with the chatbot (self-congruence), which leads to higher purchase intentions. In addition, self-congruence can lead to higher perceived authenticity of the chatbot and, ultimately, higher satisfaction (Zogaj et al., 2021). These effects are based on the similarity-attraction theory where, in interpersonal communication and human-computer interaction, people are more attracted to people or chatbots with whom they share more similarities (Gnewuch et al., 2020). However, there was no significant effect of gender matching on the competence of voice

assistants (Reinkemeier and Gnewuch, 2022a) and no interaction with the perceived humanness of chatbots (Pizzi et al., 2023). These mixed results call for further research. We put forward the following hypotheses.

H11a. Female participants perceive the female chatbot as more competent.

H11b. Male participants perceive the male chatbot as more competent.

H12. The effect of service outcome on competence is moderated by gender matching so that matching increases the effect of service recovery on the competence of the chatbot.

H13. The effect of service outcome on humanness is moderated by gender matching so that matching increases the effect of service recovery on the humanness of the chatbot

Our research model is shown in Fig. 1.

3. Methodology

We conducted a 2 (service outcome: failure vs. recovery) x 3 (chatbot gender: male, neutral, female) between-subjects online experiment to test the hypotheses. Similar to other service recovery studies in the chatbot context, respondents were asked to read a chat history (Song et al., 2023; Yang et al., 2023; Zhang et al., 2023; Zhu et al., 2023). Various chatbots were tested in practice to identify the failure and recovery scenarios. The choice fell on the WhatsApp chatbot "Klaro" from Klarmobil (2023), a German mobile communications discounter located in Hamburg and a brand of Freenet AG. Products include allnet flat rates, smartphone flat rates, and data rates. The WhatsApp chatbot was introduced for customer service in June 2019. Chatbots from the telecommunications industry have been studied more frequently in the literature (Crolic et al., 2022; Seeger and Heinzl, 2021). In addition, unlike chatbots from competitors (e.g., Vodafone and Telekom), Klaro is less rule-based, which makes it more prone to error.

We based our scenarios on a service failure from Klaro when searching for a new mobile phone contract (see Figure A2 in the appendix). More specifically, we used a response failure based on the chatbot's failure to understand (Chen et al., 2024). The chat starts with a chatbot assigned randomly from three different gender versions (avatar, name) offering to answer questions and recommend suitable tariffs for new mobile phone contracts. After the customer's approval, the chatbot asked for a specific answer (see Figure A1). Then, the chatbot detailed the gigabytes of the tariff but put forward an oversized (wrong) offer. In the service failure scenario, the chatbot was unable to present the correct tariff offer, whereas this was the case in the recovery scenario. In both scenarios, the chatbot did not understand the user input, at which point the user reformulated the inquiry. In the service failure scenario, the chatbot then asked for the invoice whereas, in the recovery scenario, the chatbot was able to fulfil the user request (for a similar approach, see Diederich et al., 2021). The chats were identical until the manipulation started to control for confounders (Mozafari et al., 2022). We used an open-source chat interface (Codepen.io) to integrate the scenarios into a messenger chat interface where the participants could read the chat history and scroll through the conversation. In the literature, a distinction is made between a process failure and an outcome failure (Sands et al., 2022). In the scenarios, we employed an outcome failure with an issue that could not be resolved. Process failure concerns the inadequate way in which the service is provided and the poor behaviour of human employees (Sands et al., 2022).

3.1. Pre-test of chatbot avatar gender and manipulations

We used cartoon-like chatbot avatars from the study by Borau et al. (2021) for the manipulation of gender (see Fig. 2). So far, these designs have been used in chatbot research, but they could well be replaced by more human-like designs using generative AI tools. We called the female chatbot Klara and the male chatbot Klaas. Similar names were used to avoid any effect of the first name on the evaluation of the chatbot (Borau et al., 2021). In addition, we used a gender-neutral chatbot as a control to provide a baseline for comparison (Mooshammer and Etzrodt, 2022). The gender-neutral chatbot took the name Klaro. For the gender-neutral avatar, we tested three avatars from the study by Crolic et al. (2022) (see Fig. 3). The scores of all three avatars were not significantly different from the scale centre on a bipolar adjective scale with the extremes of 'clearly male' and 'clearly female'. Consequently, it was classified as gender neutral by the participants (Crolic et al., 2022).

To test the manipulations, 30 respondents were randomly assigned to one of the six experimental groups and asked to evaluate the scenario and the gender of the avatars. We measured all items on a 7-point Likert scale (1 = strongly disagree to 7 = strongly agree) or a 7-point bipolar

Fig. 2. Manipulation check of chatbot gender (Borau et al., 2021, digital appendix).

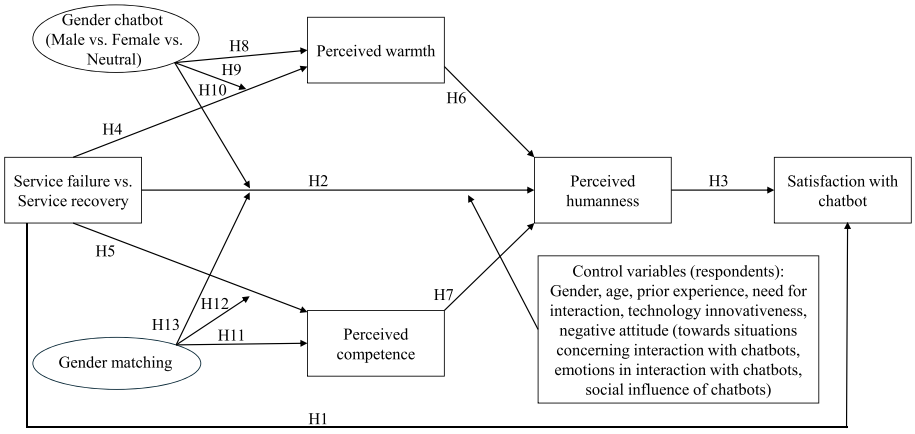


Fig. 1. Research model.

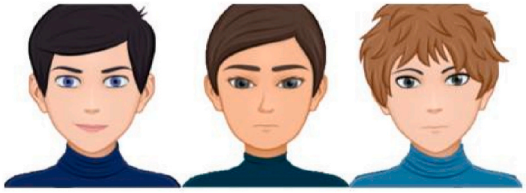


Fig. 3. Manipulation of the gender-neutral chatbot (Crolic et al. (2022), digital appendix.

scale (1 = definitely male to 7 = definitely female) except for demographics. Two participants did not pass the attention check and were excluded from the analysis. After adjustment, a total of 17 female and 11 male participants remained with an average age of 31.25 years. To check whether the manipulation of the scenarios worked, respondents answered two manipulation checks. One check concerned the service outcome (see Mozafari et al., 2022), which was correctly recognised ("The chatbot was able to solve the service inquiry"). The participants considered the service request to be resolved in the recovery scenario (n = 13; $M = 6.15$, $SD = 0.689$) but not in the service failure scenario (n = 15; $M = 1.40$, $SD = 0.828$, t (26) = -16.36; $p < 0.001$). Consequently, no adjustment of the scenarios was required for the main study.

The manipulation of female and male gender also worked. The mean values did not significantly differ from the scale endpoints (female chatbot avatar: $M = 6.89$, $SD = 0.323$; male chatbot avatar: $M = 1.32$, $SD = 0.820$, t (17) = -1.458; p = 0.163) (see similar Crolic et al., 2022). With regard to the gender-neutral avatars (see Fig. 3), the second avatar was the best match, with its mean value not differing significantly from the scale midpoint ($M = 4.25$, $SE = 1.669$, t (27) = 0.792; p = 0.435). The third avatar tended to appear slightly masculine ($M = 3.36$, SE = 1.521, t (27) = −2.237; $p < 0.05$). The first avatar was perceived as slightly feminine ($M = 4.61$, $SE = 1.685$, (t (27) = 1.906; p = 0.067). Thus, the second avatar worked best in the pretest and was therefore used in the main study.

Overall, the questionnaire received hardly any comments for improvement. Figure A1 in the appendix shows the final gender-based starting point of the scenario.

3.2. Data collection

A total of 333 participants took part in the survey between February 27 and March 23, 2023. We removed 23 respondents who did not pass the attention check and a further 10 respondents who failed the manipulation check, resulting in a final sample size of 300 persons, which were distributed fairly evenly across the six experimental groups (see Figure A3 in the appendix). The data cleansing concerned 7 of the 23 speeders who took less than 5 min for the survey, and 5 of the 15 respondents needing half an hour or more. We conservatively checked the speeders for straightlining. On average, the response time was 35 min to complete the survey (std. 313.13 min). We conducted a power/sample size analysis to determine the appropriateness of a sample size of 50 consumers per group for gender-based comparisons. The 300 and 6 group sample sizes were well within the reach of 44 respondents per group (in total 264) representing $\eta^2 = 0.06$ (medium effect) with p = 0.05 and $\beta = 0.9$.

To guarantee the distribution of the questionnaire, we used forums on (mobile) telephony, such as telefon-treff and mobilfunk-talk, and social media. The younger generations, especially Generation Z and millennials, as well as lower-income persons, are more willing to use chatbots than older generations (Katana, 2024; Statista, 2018), which is reflected in our sample. In addition, in Germany, the younger generations make more frequent use of mobile phone contracts (20–29: 84.2%, 30–39: 87%) instead of prepaid cards compared to the older generations (60 and older: 58%) (Statista, 2024). The average age of the respondents was 29.68 years. More females than males took part. Respondents were either students or employees and (had) attended university (of applied sciences). Correspondingly, the median net income was €1500 to €2000 (see Table A2 in the appendix).

The respondents rated their prior experience with chatbots as rather low ($M = 3.31$, $SE = 1.56$). Often, they had previously used a chatbot approximately every 2–3 months on average and for about 3 years. Approximately 12.6% of respondents (n = 33) use chatbots once a week or more frequently. Thirty-eight participants had no experience with chatbots at all (12.7%). The usage purpose of chatbots concerns most frequently service requests, search engine tasks, text creation, and testing. In contrast, chatbots are used less frequently to search for products and recommendations. Internal company chatbots are also employed sporadically. Over a third (36.3%, n = 95) have already used a chatbot from a telecommunications company and, of those, 27.5% (n = 26) searched for a mobile phone contract. On average, participants changed their mobile phone contracts every 5–6 years (see Table A3 in the appendix).

3.3. Questionnaire design and measurement items

As in the pre-test study, the term chatbot was introduced first. The participants watched a short video clip in which the term was explained and an exemplary chat process was shown (KIKI erklärt KI, 2020). General questions followed on the use of and experience with chatbots and the switching frequency of mobile phone contracts. After answering demographic questions about gender and age, respondents were randomly assigned to one of the six scenarios. Participants had to read the chat carefully and put themselves in the user's shoes. Following the manipulation checks, respondents completed the main part of the questionnaire, which included the item scales of the constructs and the control variables. Finally, the participants were asked for additional personal details.

Age and gender are typical control variables in the chatbot context (Diederich et al., 2021; Liu et al., 2023; Seeger and Heinzl, 2021). In addition, the need for interaction and negative attitudes towards chatbots influence perceived humanness (Blut et al., 2021). Furthermore, prior experience with chatbots has a (positive) effect on satisfaction (Diederich et al., 2021) and purchase intention (Luo et al., 2019) as well as on repair strategy preferences (Ashktorab et al., 2019). Van Doorn et al. (2017) assume that the technological readiness of users increases the chatbot's perceived warmth and competence. Similar to Belanche et al. (2020), we included the level of 'technology innovativeness' (Parasuraman, 2000) to control for affinity with chatbots.

The item scales used in this study were drawn from the literature (see Table A2 in the appendix). The internal consistency of the scales used is high, with Cronbach's alpha (α) values mostly above the guideline value of 0.7 (Hulland et al., 2018). Only the values of the two control variables 'negative attitude towards situations concerning interaction with chatbots' (0.65) and 'negative attitude towards emotions in interaction with chatbots' (0.55) are below this value. Since we confirmed multidimensionality for all constructs, with the variance extracted being mostly above 0.5, we calculated an average score value on the scale items. While perceived warmth and competence, need for interaction, technology innovativeness, and negative attitude towards emotions in interaction with chatbots range above the scale's midpoint, the other score values range below (see Table A4 in the appendix). In addition, we present the summary statistics and bivariate correlations for the 6 experimental groups in Table A5 in the appendix.

4. Results

4.1. Manipulation checks

As in the pre-test, the service outcome ("The chatbot was able to solve the service inquiry" (Mozafari et al., 2022) was confirmed. The mean values ($M_{Failure} = 1.59$; $M_{Recovery} = 5.75$) differ significantly

between the two scenarios (t (248,811) = −37.34; p < 0.001).

The test on the manipulation of gender ("What gender was the chatbot?") using the bipolar gender scale revealed that the mean values did significantly differ from the scale endpoints (female chatbot avatar: $M = 6.29$, $SD = 1.057$, (t (99) = -6.72; p < 0.001); male chatbot avatar: $M = 2.22$, $SD = 1.290$, t (102) = 9.62; p < 0.001) and the scale midpoint (gender-neutral chatbot avatar: $M = 3.21$, SD = 1.399 (t (96) = -5.59; p < 0.001). However, the post-hoc test with Bonferroni correction shows that the female chatbot is perceived as significantly more feminine than the male chatbot ($M_{diff} = 4.1$; p < 0.001) and the gender-neutral chatbot ($M_{diff} = 3.1$; p < 0.001). The male chatbot is perceived as significantly more masculine than the gender-neutral chatbot ($M_{diff} = -0.98$; p < 0.001), and the three mean values all differ significantly from each other.

However, the recovery scenario was perceived as significantly more credible ($M_{Recovery} = 5.3$ vs. $M_{Failure} = 4.8$; t (283) = -3.016; p < 0.01) and realistic ($M_{Recovery} = 5.37$ vs. $M_{Failure} = 5.01$; t (290) = -2.073; p < 0.05). Due to the randomised allocation to the experimental groups, there should be no structural differences in the sample between the recovery and failure scenarios (Sella et al., 2021). In contrast, the Mann-Whitney *U* test found no significant differences between the failure and recovery scenarios in terms of age (p = 0.258), net income (p = 0.285), the highest level of education (p = 0.910), prior experience with chatbots (p = 0.174), affinity for technology (p = 0.266), and need for interaction (p = 0.493).

4.2. Hypothesis testing

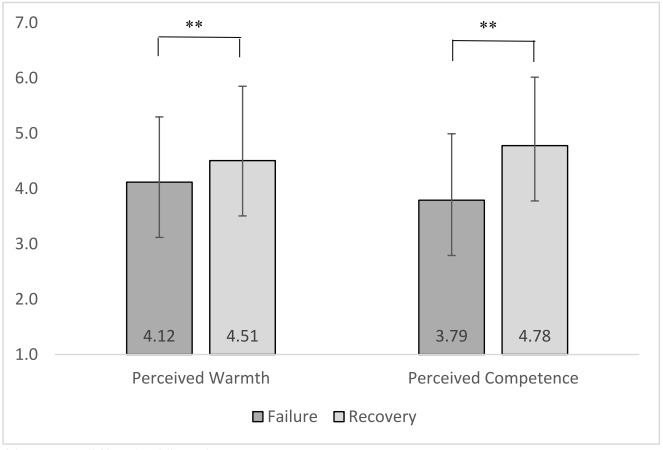
Participants were more satisfied in the recovery scenario ($M = 5.22$, $SD = 1.148$) than in the failure scenario ($M = 1.81$, $SD = 0.870$). We used Welch's F test (Derrick and White, 2016) resulting in a significant effect (F (1, 271.94) = 844.952, p < 0.001; $\eta^2 = 0.739$) and support for H1. With regard to hypothesis 2, as proposed, Welch's F test showed that the perceived humanness was significantly higher in the recovery scenario ($M = 3.41$, $SD = 1.259$) than in the failure scenario ($M = 2.88$, $SD = 1.100$) (F (1, 289.280) = 15.319, p < 0.001; $\eta^2 = 0.049$). We also confirmed the importance of perceived humanness for satisfaction with the chatbot (F (1, 272) = 3.394, p < 0.001, $\eta^2 = 0.252$) in line with H3. The service outcome has a significant positive effect on the perceived warmth of the chatbot (F (1, 298) = 7.996, p = 0.005; $\eta^2 = 0.026$) in the recovery scenario ($M = 4.51$, $SD = 1.201$) compared to the failure scenario ($M = 4.12$, $SD = 1.179$). In addition, the chatbot in the recovery scenario is perceived as significantly more competent ($M = 4.78$, $SD = 1.238$) than in the failure scenario ($M = 3.79$, $SD = 1.345$) (F (1, 298) = 43.596, p < 0.001; $\eta^2 = 0.128$). We found support for H4 and H5 (see Fig. 4). In turn, warmth (F (1, 270) = 5.403, p < 0.001; $\eta^2 = 0.367$) and competence (F (1, 275) = 6.835, p < 0.001; $\eta^2 = 0.375$) have a significant positive effect on perceived humanness supporting H6 and H7.

We used the stereotype content model (SCM) to propose that the female chatbot is perceived as significantly warmer (H8), whereas the male chatbot, contrary to the SCM, is not perceived as significantly more competent. While the gender of the chatbot affects the perceived warmth (F (2, 297) = 3.460, p = 0.033; η2 = 0.023), we applied a post-hoc test to test for mean differences regarding chatbot gender. Klara was perceived as significantly warmer as a female chatbot ($M = 4.53$, $SD = 1.208$) than the male chatbot Klaas ($M = 4.09$, $SD = 1.204$) ($M_{diff} = 0.44$; p = 0.024). There were no significant differences in warmth between the gender-neutral chatbot Klaro ($M = 4.32$, $SD = 1.167$) and Klaas ($M_{diff} = 0.23$; p = 0.368) or between Klaro and Klara ($M_{diff} = 0.21$; p = 0.425). These results confirm hypothesis H8. With regard to competence, gender has no effect (F (2,297) = 0.942, p = 0.391; $\eta^2 = 0.006$), which is reflected in the mean values ($M_{Klara} = 4.42$; $M_{Klaro} = 4.27$; $M_{Klaas} = 4.17$) (see Fig. 5).

We only partly confirmed H9 that the gender of the chatbot moderates the effect of service outcome on perceived warmth (F (2, 294) = 1.243, p = 0.290, η2 = 0.008). A pairwise comparison of the mean changes shows that the effect of service outcome on perceived warmth is significant for the male chatbot ($M_{diff} = 0.654$, p = 0.005). This effect was insignificant for the gender-neutral ($M_{diff} = 0.13$, p = 0.585) and female chatbots ($M_{diff} = 0.352$, p = 0.137). In addition, the male chatbot in the failure scenario was perceived as significantly less warm than the female chatbot ($M_{diff} = -0.582$, p = 0.013) and the gender-neutral chatbot ($M_{diff} = -0.482$, p = 0.039). The differences between the female and gender-neutral chatbots are small ($M_{diff} = 0.100$, p = 0.672). However, in the recovery scenario, the differences in the perceived warmth of the male chatbot are insignificant – in particular, compared to the gender-neutral chatbot ($M_{diff} = 0.041$, p = 0.864), but also compared to the female chatbot ($M_{diff} = -0.280$, p = 0.236). Overall, in the failure scenario, the male chatbot is perceived as less warm (and is more blamed) than the gender-neutral and female chatbots. After recovery, however, the significant differences disappear with the female chatbot still being rated as warmer and the difference with the gender-neutral chatbot increasing ($M_{diff} = 0.321$, p = 0.181) (see Fig. 6).

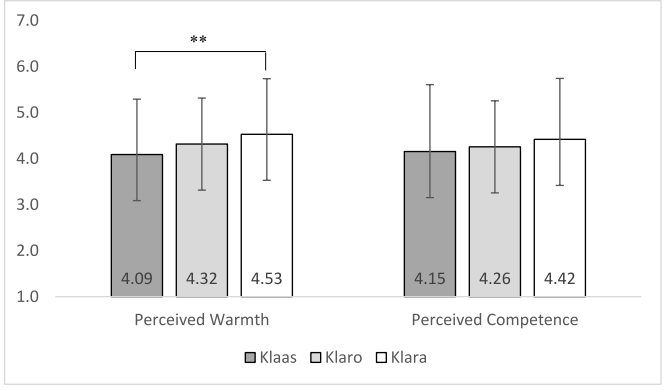
With regard to perceived competence, the mean values of the three chatbots did not differ significantly in any of the two scenarios (F (2, 294) = 0.907, p = 0.405, $\eta^2 = 0.006$). We also must reject H10, since the gender of the chatbot does not moderate perceived humanness (F (2, 294) = 0.421, p = 0.657, $\eta^2 = 0.003$). There was again no difference in the perception of the humanness of the three chatbots within the failure and recovery scenarios.

In the case of a match between the gender of the user and the gender of the chatbot, the competence of the chatbot was rated slightly higher



**: means differ significantly at p<0.01

Fig. 4. Perceived warmth and competence depending on the service outcome
**: means differ significantly at p < 0.01.



**: means differ significantly at p<0.01

Fig. 5. Perceived warmth and competence depending on chatbot gender
**: means differ significantly at p < 0.01.

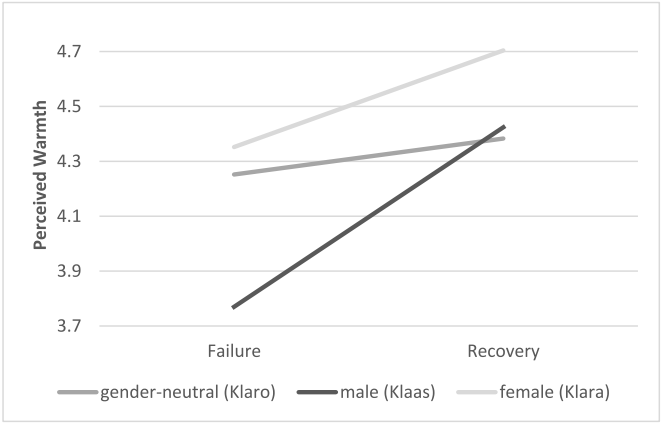


Fig. 6. Interaction effect of service outcome and chatbot gender on perceived warmth.

(M_{Match} = 4.364, SD_{Match} = 1.365; $M_{Mismatch}$ = 4.197, $SD_{Mismatch}$ = 1.436). However, this difference was not significant (F (1, 200) = 0.719; p = 0.398, η^2 = 0.004). Female participants rated the female chatbot (M = 4.466) as slightly more competent than the male chatbot (M = 4.127) (M_{diff} = 0.339; p = 0.175). For the male participants, Klara appeared also slightly more competent (M = 4.324) than Klaas (M = 4.201). Moreover, these mean values did not differ significantly (M_{diff} = 0.122; p = 0.707) and, therefore, we found no support for H11a and H11b.

In both outcome scenarios, females rated the female chatbot as more competent (failure: M = 4.18, M_{diff} = 0.395, p = 0.203; recovery: M = 4.77, M_{diff} = 0.265, p = 0.412). Male participants also perceived the female chatbot as slightly more competent in the two scenarios. For males, the competence values for a match or mismatch are almost identical for both a failure (M = 3.25, M_{diff} = -0.031, p = 0.941) and a recovery (M = 5.11, M_{diff} = -0.143, p = 0.724). In general, females rated a chatbot regardless of gender as more competent than males in the failure scenario ($M_{Match\text{-}diff}$ = 0.926, p = 0.010; $M_{Mismatch\text{-}diff}$ = 0.500, p = 0.195) This result was the same for males in the recovery scenario ($M_{Match\text{-}diff}$ = 0.353, p = 0.346; $M_{Mismatch\text{-}diff}$ = 0.741, p = 0.051) (see Fig. 7).

There was no moderation of gender matching on the relationship between service outcome and perceived competence (F (1, 198) = 0.085, p = 0.771). Competence in the failure scenario is not rated significantly different in the case of a gender match (M = 3.83) or mismatch (M = 3.62) (M_{diff} = 0.219; p = 0.392). Furthermore, in the service recovery scenario, there is hardly any difference in terms of perceived competence in the case of a gender match (M = 4.906) or mismatch (M = 4.792) (M_{diff} = 0.113; p = 0.661). H12 is therefore not supported.

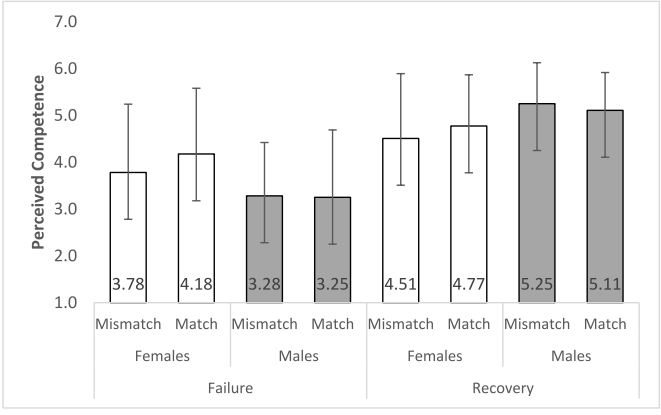


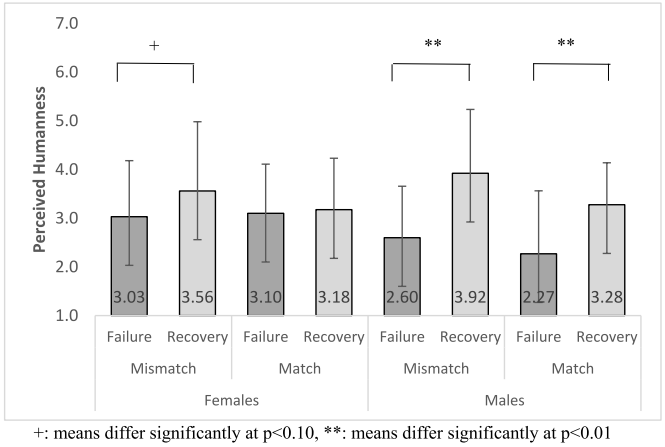
Fig. 7. Perceived competence of females and males depending on service outcome and gender (mis)matching.

Participants perceived the chatbots in the recovery scenario as significantly more human if the gender did not match their gender (M = 3.7, M_{diff} = 0.515; p = 0.031). In particular, female participants did not rate the female chatbot as more human after the recovery ($M_{Match\text{-}diff}$ = 0.075; p = 0.792) but did so for the male chatbot ($M_{Mismatch\text{-}diff}$ = 0.527; p = 0.076). Male participants rated both genders as significantly more human after the recovery ($M_{Match\text{-}diff}$ = 1.006; p = 0.006; $M_{Mismatch\text{-}diff}$ = 1.322; p = 0.001) but also showed a preference for a mismatch in the service recovery (see Fig. 8). They perceived Klara as rather more human (M = 3.92, M_{diff} = 0.646; p = 0.083) but, for females, the mean values did not differ significantly (M_{diff} = 0.384; p = 0.196). In the failure scenario, there was no effect of gender matching on perceived humanness (M_{diff} = 0.131; p = 0.587). Therefore, hypothesis 13 was only partially confirmed.

We considered the control variables following an analysis of covariance (Ancova). After including them in the model, we checked whether the effect from H2 persisted. We removed the variable 'negative attitudes towards emotions of chatbots'. There was a significantly different perception of the control variable within the groups of the service result (F (1, 298) = 4.32, p < 0.05). While we found homogeneity of the regression slopes for the remaining control variables, the control variable 'negative attitude towards the social influence of chatbots' had a significant positive effect on perceived humanness (F (1, 297) = 3.90, p = 0.049). We removed these two control variables.

4.3. Model evaluation

The effect of service outcome on perceived humanness was not significant when perceived warmth and competence were included (F (1, 296) = 1.26, p = 0.263), indicating indirect mediation (Zhao et al., 2010). We used PROCESS model 80 (Hayes, 2022) to test mediation and excluded the moderators because almost all showed no significant effect. We tested the model using a bootstrapping approach (n = 10,000) with a confidence interval of 95%. The control variables were included in the model as covariates. The results show that the satisfaction with the chatbot is very well explained with an R^2 of 0.82. The perceived warmth and competence mediate the effect of service outcome on perceived humanness. In addition, perceived warmth (b = 0.09; p = 0.019) and perceived competence (b = 0.14; p = 0.0002) have a significant positive direct effect on satisfaction with the chatbot. We report partially standardized regression coefficients due to the dichotomous outcome variable. Fig. 9 summarizes the results of the mediation analysis. The results for the three chatbot gender types confirm some results and display



+: means differ significantly at p<0.10, **: means differ significantly at p<0.01

Fig. 8. Perceived humanness of females and males depending on service outcome and gender (mis)matching
+: means differ significantly at p < 0.10, **: means differ significantly at p < 0.01.

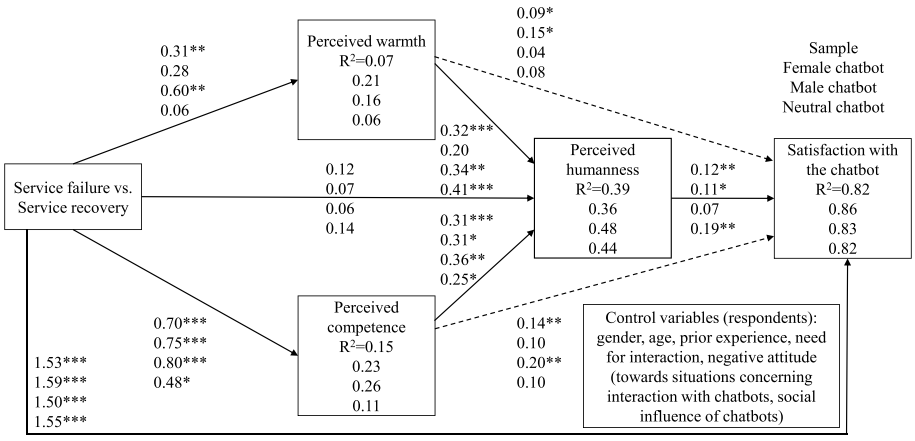


Fig. 9. Mediation analysis (PROCESS model 80)
Partially standardized regression coefficients; ***p ≤ 0.001; **p ≤ 0.01; *p ≤ 0.05, dotted line: tie not established in the model.

differences. The significant effect of the service outcome on perceived warmth for the male chatbot of the pairwise comparison is again demonstrated. For perceived competence, the effect was lower but still substantial for the gender-neutral chatbot. There is a significant direct effect of perceived warmth on satisfaction for the female chatbot. However, the indirect path over perceived humanness is driven by the perceived competence of the female chatbot. For the male chatbot, the perceived competence only increases satisfaction directly. The perceived humanness has the highest impact on satisfaction in the case of the gender-neutral chatbot.

The control variable (covariate) prior experience had a significant positive effect on perceived warmth (t = 2.75, p = 0.0063, b = 0.18) and perceived competence (t = 2.10, p = 0.0363, b = 0.13). This especially holds for the male chatbot (perceived warmth: t = 2,80, p = 0.0061, b = 0.32; perceived competence: t = 2.05, p = 0.0428, b = 0.22). A pairwise comparison shows that prior experience is significantly higher in the recovery scenario (M_{diff} = 0.48; p = 0.020). The effect disappears when experienced users are confronted with a service failure (M_{diff} = 0.17; p = 0.43). The other control variables did not influence the variables of the main model. For the female chatbot, the respondent's age negatively affected perceived warmth (t = −2.82, p = 0.0058, b = −0.30), while satisfaction was positively related to technology innovativeness (t = 2.76, p = 0.007, b = 0.14).

The total partially standardized indirect effect of service outcome on satisfaction is 0.1818, which is significant (CI = 0.1073, 0.2632). In particular, the effect of service outcome on satisfaction via perceived competence is significant (b = 0.1018, CI = 0.0448, 0.1737, path 2). In addition, the indirect effects of service outcome on satisfaction via perceived competence and perceived humanness (b = 0.0262, CI = 0.0087, 0.0486, path 5), of service outcome on satisfaction via perceived warmth (b = 0.0271, CI = 0.0034, 0.0608, path 1), and of service outcome on satisfaction via perceived warmth and perceived humanness (b = 0.117, CI = 0.0020, 0.0273, path 4) are smaller, but also significant. Only the effect of service outcome on satisfaction via perceived humanness is insignificant (b = 0.0149, CI = −0.0084, 0.0444, path 3).

The contrasts show that the paths of the service outcome on satisfaction do not differ significantly in their strength. Four out of ten comparisons are significant. Mediation on satisfaction had a significantly stronger effect via perceived competence than via perceived warmth (path 1 vs. path 2, path 2 vs. path 4, path 2 vs. path 5) and via perceived humanness (path 2 vs. path 3). Table 1 summarizes the results of the hypothesis testing.

Table 1
Summary of hypothesis testing.

Hypotheses		Method	Supported
H1	Service recovery increases satisfaction.	PROCESS, single factor variance analysis	Yes
H2	Service recovery increases perceived humanness.	PROCESS, single factor variance analysis	No
H3	Perceived humanness increases satisfaction.	PROCESS, univariate ANOVA	Yes
H4	Service recovery increases perceived warmth.	PROCESS, univariate ANOVA	Yes
H5	Service recovery increases perceived competence.	PROCESS, univariate ANOVA	Yes
H6	Perceived warmth increases perceived humanness.	PROCESS, univariate ANOVA	Yes
H7	Perceived competence increases perceived humanness.	PROCESS, univariate ANOVA	Yes
H8	Chatbot gender affects perceived warmth (female CB > male CB).	Univariate ANOVA	Yes
H9	Chatbot gender moderates the effect of the service outcome on perceived warmth.	Two-way ANOVA	No
H10	Chatbot gender moderates the effect of the service outcome on perceived humanness.	Two-way ANOVA	No
H11a	Female participants perceive the female CB as more competent than the male CB.	Two-way ANOVA	No
H11b	Male participants perceive the male CB as more competent than the female CB.	Two-way ANOVA	No
H12	Gender matching moderates the effect of the service result on expertise.	Two-way ANOVA	No
H13	Gender matching moderates the effect of the service outcome on perceived humanness.	Two-way ANOVA	No

CB = chatbot.

5. Discussion

This study compares a chatbot's service failure with a service recovery. Overall, the effect of service recovery is strongest on satisfaction with the chatbot, with the indirect effects of the chatbot's two central dimensions of social cognition being rather small. While there was a single positive effect of perceived humanness on satisfaction, the direct effects of perceived warmth and competence are more important. The recovery increases the perceived competence and the perceived warmth and, in turn, the perceived humanness of the chatbot. Warmth and competence are particularly important for chatbots that make product recommendations because users are more likely to follow the recommendations if the chatbots are perceived as warmer and more

competent (Ahn et al., 2022). While perceived warmth and competence explain perceived humanness rather well ($R^2 = 0.38$), they directly and separately affect satisfaction, with the influence of perceived competence being much stronger. This is in line with other studies in which both higher perceived warmth and competence led to higher user satisfaction (Zheng et al., 2023) or a higher rating of service quality (Han et al., 2022). In classic service research, competence and warmth also have a significant effect on customer satisfaction (Güntürkün et al., 2020).

Furthermore, the perceived humanness of chatbots leads to a higher intention to repurchase (Fota et al., 2022) and a higher intention to reuse (Sheehan et al., 2020). However, an anthropomorphic design increases expectations before the interaction, resulting in an even greater disappointment after a service failure (Crolic et al., 2022). In addition, a high degree of similarity of a robot or chatbot to a human can trigger discomfort (Thaler et al., 2021), referred to as the 'uncanny valley' (Mori et al., 2012). The feeling of discomfort is even exacerbated by a faulty chatbot service (Diederich et al., 2021). These may explain the insignificant effect of perceived humanness in the model.

With regard to research question 1 and consistent with the literature, the female chatbot is perceived to be significantly warmer than the male chatbot (Ahn et al., 2022; Borau et al., 2021). The effects of the service outcome on perceived warmth as well as the mediation of service outcome on humanness via warmth were only significant for the male chatbot. In the case of a service failure, only the male chatbot suffers from a significant loss of perceived warmth. In the recovery case, however, Klaas, Klara, and Klaro were perceived as similarly warm. Gender-neutral chatbots appear similarly warm in both the service failure and service recovery scenarios. Against a more utilitarian service context than a hedonic one in this study, it is rather surprising that warmth appears to be at least as equally prominent as competence. For service robots in tourism, the effect of appearance and service context was confirmed (Liu et al., 2022). However, the appearance in this research was robot-like and not human-like, and the latter seems to make a difference.

According to the stereotype content model, men are perceived as more competent than women (Fiske et al., 2002). In the context of human service employees, male employees were perceived as significantly more competent (Mccoll-Kennedy and Sparks, 2003) and achieved a higher assessment of service quality (Snipes et al., 2006). In this study, we could not confirm that the perception of competence of a chatbot – in contrast to warmth – is dependent on the context and can be increased if the gender of the chatbot matches the gender associated with the product (Beldad et al., 2016). The context of recommending a mobile phone contract represents more of a utilitarian purpose for which male chatbots should appear more competent (Ahn et al., 2022). However, in this study, the female chatbot was perceived as slightly, but not significantly, more competent than the male chatbot ($M_{female} = 4.42$ vs. $M_{male} = 4.15$). The result is similar to studies in the context of a hedonic product (recommending sportswear) (Toader et al., 2020) or in the health context (Borau et al., 2021). Against the widespread feminization of the service sector (Korczynski, 2005; Scholarios and Taylor, 2010) and the creation of chatbots in service with primarily female design characteristics (Feine et al., 2020), we argue that the context of the chatbots in this study could compensate for the stereotype that men are perceived as more competent. The composition of the sample with predominantly young respondents who encounter many competent women in their everyday lives (Ebert et al., 2014), points to modern gender attitudes. These can help to break down gender stereotypes and reduce the salience of gender-specific characteristics of products and chatbot avatars (Fugate and Phillips, 2010). The absence of an effect of chatbot gender on perceived competence is a surprising result since, for voice assistants, gender stereotypes indeed play a role, and male assistants are perceived as more competent (Ernst and Herm-Stapelberg, 2020). Since visual cues can define personality traits (Huang et al., 2021), glasses or reputable looks might lead to a similar evaluation.

We based our considerations concerning research question 2 on the similarity attraction theory, but we could not confirm that the matching of the chatbot and the participant's gender positively influences the former's perceived competence. There is some evidence concerning service employees (Foster and Resnick, 2013; Quach et al., 2017) or chatbot avatars (Benbasat et al., 2020; Zogaj et al., 2023) that customers prefer the same gender in service. In our study, both female and male participants rated the female chatbot slightly, but not significantly, higher in competence. In this regard, matching effects seem to be context-dependent concerning the product. In contrast to Zogaj et al. (2023) who investigated gender matching in a chatbot selling trousers, there were no matching effects for more utilitarian purposes such as buying books via a voice-controlled chatbot (Reinkemeier and Gnewuch, 2022b) or when renting a car (Pizzi et al., 2023). The more favourable attitude of women towards women compared to men towards other men (Rudman and Goodwin, 2004) is reflected in the higher competence values of the female chatbot.

In addition, the chatbot was perceived as significantly more human in the recovery scenario if the gender of the chatbot differed from the gender of the participant. Customers often experience more negative emotions during a service recovery if the gender of the customer matches that of the employee due to higher service expectations in someone similar to oneself (Boshoff, 2012). Furthermore, a higher perceived similarity of facial expressions following a service failure leads to higher dissatisfaction and a higher willingness to speak negatively about the company (Lim et al., 2017). Female customers are more willing to visit a hotel again if an empathic male service employee attempts service recovery instead of an empathetic female. They expect empathic treatment from the same gender but are positively surprised when a male service employee shows empathy in service recovery (Mccoll-Kennedy et al., 2003). This might explain why, in our study female participants did not perceive the female chatbot as more human after the recovery.

5.1. Theoretical implications

This study shows that error-prone anthropomorphic chatbots can restore user satisfaction by resolving the issue following a service failure, thereby increasing the perceived warmth and competence. So far, only Han et al. (2022) and Zhou and Chang (2024) have used the two dimensions in the service recovery case. They found that only competence, but not warmth, influences service quality. In this respect, this study highlights that warmth can also have an influence on important downstream variables in chatbot service recovery and, thus, complements research in this area. Users perceive chatbots as social actors with human characteristics (Reinkemeier and Gnewuch, 2022b), which allows, for example, the integration of certain design elements, such as gender, which subconsciously activate gender stereotypes (McDonnell and Baxter, 2019), so that female chatbots appear warmer. A high level of perceived warmth is also important in reducing user scepticism, increasing trust in the chatbot (Pizzi et al., 2023), and improving the attitude towards the chatbot (Maar et al., 2023). User engagement with the chatbot operator's brand can be increased (Kull et al., 2021). In the robot context, warmth increases the emotional value and ultimately leads to a greater willingness to use the service again (Belanche et al., 2021). Perceived warmth plays an important role since people rate warmth above competence (Fiske et al., 2007), which is also noted in the service context (Castro et al., 2012).

Higher perceived competence, on the other hand, goes hand in hand with an increase in functional and monetary value, which again leads to a higher willingness to reuse (Belanche et al., 2021), and it increases trust (Toader et al., 2020) and favourable attitudes (Maar et al., 2023). Our study reveals that competence is more important than warmth in restoring satisfaction with a chatbot after a result error. However, both effects on perceived humanness are equally high (van Doorn et al., 2017) and improve the perception of chatbots in service recovery. Our

results on warmth and competence and their relationship with satisfaction are consistent with previous research. In service recovery, a friendly chatbot is more likely to be forgiven (Xing et al., 2022), and an empathetic communication style leads to a higher repurchase intention (Fota et al., 2022). A sincere chatbot can achieve better satisfaction scores in service recovery, and high emotional intelligence is indirectly associated with higher satisfaction (Zhang et al., 2023).

In contrast to previous research in social psychology (Holoien and Fiske, 2013; Kervyn et al., 2009) or organization and brand perception research (Aaker et al., 2012), there is a linear relationship between warmth and competence – for example, a more competent recovery chatbot was perceived as significantly warmer (Spearman-Rho 0.681, $p < 0.001$). Poor service from a chatbot leads to lower perceived humanness (Diederich et al., 2021; Sheehan et al., 2020). The results of this study show that this is due to a lower assessment of the competence and warmth of the chatbot. With regard to the importance of the perceived humanness of chatbots in the service context, our study shows mixed results. This is in line with previous studies, which found that anthropomorphizing chatbot avatars in the event of a service failure can have both positive effects through a lower loss of trust (Seeger and Heinzl, 2021) and negative effects through higher expectations (Crolic et al., 2022).

This study contributes to the role of similarity-attraction theory in the chatbot context. Previous research has confirmed that a higher perceived similarity of the user with the chatbot leads to them perceiving the chatbot as more authentic and having greater satisfaction with it (Zogaj et al., 2021). Furthermore, users trust the chatbot more when there is a match of personalities (Reinkemeier and Gnewuch, 2022b). Regarding gender matching, there are mixed results (Reinkemeier and Gnewuch, 2022b; Zogaj et al., 2023). In this study, too, there were no effects of matching.

We found that only the male chatbot loses warmth following a service failure. If there is a service recovery, there are no significant differences in terms of warmth. This result is in contrast to the stereotype content model, which proposes that men appear less warm in general (Fiske et al., 2002). A reason could be that the service failure was performance related – that is, a result error – and not relationship oriented, such as a process error. Customers are more likely to forgive a performance-related error if the error was made by a female employee (Wei and Ran, 2019). The expectations regarding the male chatbot's performance might have been higher in advance.

5.2. Management implications

Users have high expectations of chatbots (Rozumowski and Haupt, 2021), which often cannot be met in practice – especially for more complex tasks – because they are too prone to error (Tran et al., 2021). This study shows that chatbots can restore user satisfaction following a self-induced service failure. It is therefore important that chatbots are capable of determining – at the latest, after the user's reaction – that they were unable to solve the request to initiate a service recovery. Because the majority of users leave the chat conversation relatively quickly after a chatbot service failure (Dharaniya et al., 2020), the chatbot should be able to detect the likelihood that the user's request has not been correctly understood and classified.

Accordingly, the chatbot should acknowledge and deal with misunderstandings using different messages and explain how the algorithm works (Ashktorab et al., 2019). In this way, chatbots can appear more intelligent in service recovery (Ashktorab et al., 2019) and have a higher perceived functional value (Song et al., 2022). Thus, according to the findings of our study, chatbots can seem more competent and restore the user's satisfaction. The chatbot could therefore ask follow-up questions, which would not lead to a loss of willingness to use the chatbot in the future, given that follow-up questions are part of a natural human conversation (Sheehan et al., 2020). In this respect, an interactive communication style can make chatbots appear more human (Go and Sundar, 2019) and, thus, increase satisfaction.

Based on the results of this study, we recommended that chatbots must appear competent but also warm in executing a successful service recovery. There are already some design recommendations in the literature for increasing the warmth and competence of chatbots – for example, a direct gaze direction (Pizzi et al., 2023), realistic pictures, (Pizzi et al., 2023), a human name (Zheng et al., 2023), an interactive communication style (Go and Sundar, 2019), and delaying the response to simulate the typing of a human (Gnewuch et al., 2018). To make the chatbot appear warmer and more empathetic, sentiment analysis can be used to adapt the chatbot's language to a more emotional style (Huang and Rust, 2022).

However, it can be difficult to design a chatbot with higher warmth and competence at the same time, because a warmer design can lead to lower competence – for example, when using emojis (Huang et al., 2021). This suggests that the designers of chatbots would have to decide whether to use a warm or competent style. This study shows that competence is more important in service recovery. When positioning a brand, it is difficult to establish both a warm and a competent brand (Aaker et al., 2012). However, there are brands, such as Johnson & Johnson and Coca-Cola, that have both a warm and competent brand perception (Kervyn et al., 2022). Therefore, it should be possible for chatbots to appear both warm and competent.

In our study, the female chatbot in the service failure scenario appeared significantly warmer than the male chatbot and still achieved better competence scores in this scenario. This suggests that the female chatbot is more likely to be forgiven than the male chatbot (Toader et al., 2020). Error-prone chatbots should, therefore, have more female design characteristics to avoid negative evaluations following a service failure. However, young respondents perceived the female chatbot as less warm, pointing to a need for visual improvements of the avatar. The gender-neutral chatbot also performed relatively well and did not lose significant warmth following service failure. In this respect, a gender-neutral design is possible in the service failure and recovery context, especially since there are contexts that are not associated with any gender (Fugate and Phillips, 2010). The recommendation to use more female designs in service is in line with current practice, with chatbot design being predominantly female (Feine et al., 2020). If a male chatbot is used, experienced users should be involved. In addition, the chatbot can use a more feminine communication style – for example, qualifying recommendations by using words such as "maybe" or "possibly" (Mou et al., 2019). This is consistent with users' preference not to receive unspecific recommendations, but nuanced responses where possible alternatives are presented (Ashktorab et al., 2019; Følstad and Taylor, 2020).

To increase perceived competence, chatbots can introduce humour to service recovery by employing humorous emojis (Liu et al., 2023) or self-deprecating humour (Yang et al., 2023). For interpersonal interaction, humour is positively related to social competence and emotional intelligence (Yip and Martin, 2006), which also holds for human-chatbot interactions (Xie et al., 2024). Therefore, customers may view a humorous chatbot as more competent in solving their task-oriented needs and problems. For brands, clever humour leads to both higher perceived warmth and higher competence (Howe et al., 2023). Humour in service employees leads to a higher willingness to give the service provider a second chance, compared to an apology or compensation (Kobel and Groeppel-Klein, 2021). Getting a second chance is particularly relevant for chatbots because many users give up following a service failure (Dharaniya et al., 2020).

Companies employing a chatbot in their service should be aware of the susceptibility of chatbots to errors and the degree of anthropomorphic design. Anthropomorphic design elements can help to increase satisfaction with a chatbot with a low error rate and the ability to self-recovery (Choi et al., 2021). When resolving issues following a chatbot's service failure, a high degree of anthropomorphism does not seem necessary (Song et al., 2022). In our study, human-like attributes can

increase the competence (and warmth) level of the chatbot and, in turn, the degree of satisfaction. For chatbots that produce many service failures, caution should be advised regarding design because anthropomorphic design elements raise expectations of chatbots before interaction takes place (Ben Mimoun et al., 2012) and can trigger user anger in the event of poor service (Crolic et al., 2022). In contrast to the first rule-based chatbots, generative chatbots communicate in a more human-like fashion, being able to consider the user's last messages in their response. However, they can be difficult to build and train (Adamopoulou and Moussiades, 2020).

Contrary to Zogaj et al. (2023), our results suggest that, in service recovery, a chatbot with attributes of the opposite gender appears more human than a chatbot of the same gender. Therefore, in complaint management, chatbots should have the opposite gender. This finding is consistent with studies on service recovery with human employees (Boshoff, 2012; Lim et al., 2017; Mccoll-Kennedy and Sparks, 2003). Gender matching seems reasonable when one's gender is attributed to more competence and knowledge about a product (e.g., clothing, Quach et al., 2017) or when dealing with presumably unpleasant topics (e.g., health, Foster and Resnick, 2013).

5.3. Limitations and outlook

Our study has several limitations. We based the scenarios on an outcome error, but users react differently to the process errors of chatbots (Sands et al., 2022), which might lead to different results. Another limitation is the significantly lower scenario credibility in the failure scenario. After the user expressed irritation using three question marks, the failure scenario ended abruptly. This abrupt end can negatively impact the perception of the conversation because it violates the norms of a good conversation (Guydish and Fox Tree, 2021). In addition, a chatbot in practice would respond to every user input. However, a response to the question marks was omitted to ensure similar scenario lengths and to avoid considering the response as part of the service recovery. In particular, participants with higher prior experience considered the recovery scenario significantly more realistic ($M_{Failure} = 4.73$; $M_{Recovery} = 5.61$; $M_{diff} = 0.879$; $p < 0.01$), while participants with lower experience rated the scenarios equally credible and realistic ($M_{diff} = 0.108$; $p = 0.592$). Participants with less experience seem to have little faith in chatbots and perceive a chatbot failure as equally credible as a recovery. In addition, our study did not include a success scenario without requiring recovery – for example, a routine service interaction – making it impossible to determine whether the chatbot could fully restore satisfaction in the recovery scenario. The severity of a service failure (Xing et al., 2022), the type of service failure (Xing et al., 2022; Chen et al., 2024) – for example, low personalization or task complexity (Murtaza et al., 2024) – should be considered in order to enhance practical relevance.

Another limitation concerns the design of the gender-neutral chatbot. We relied on examples from the literature. However, Klaas and Klara wear glasses and smile, unlike Klaro. Wearers of glasses are stereotypically perceived as more intelligent, hardworking, and successful (Grant et al., 2016; Harris et al., 1982) but also as less attractive and social (Grant et al., 2016). Other studies conclude that glasses lead to higher warmth and competence (Fetscherin et al., 2020) or they found no effect on perceived intelligence (Lundberg and Sheehan, 1994). At the same time, smiling in service leads to higher perceived warmth and lower perceived competence (Wang et al., 2016). Moreover, smiling has the strongest behavioural influence on perceived warmth and friendliness (Bayes, 1972; Sundaram and Webster, 2000). The absence of a smile from the gender-neutral chatbot could, therefore, have influenced perceptions of warmth and competence. Other visual limitations concern the cartoon-like design of the three chatbots. The upcoming AI painting tools (Ma and Huo, 2024) allow the design of chatbots to appear more human-like, which might affect our findings. While we tested three variants of gender-neutral chatbot avatars, we did not systematically evaluate different facial features. We also did not measure the attractiveness of the three different gender versions which might have influenced their perception (Aumüller et al., 2024).

Furthermore, the limitations of the sample must be taken into account. Due to the young sample, demographic biases might affect generalizability. The weaker implicit gender stereotyping (Ebert et al., 2014) in the younger generations could be the reason for the non-significant effects of matching and chatbot gender on competence. Here, studies with different samples (older respondents or people from other cultures) or other product/service categories could contribute to improving the generalizability of the findings. In addition, the sample sizes for gender matching among male participants were rather small, with the sample sizes of gender mismatch falling below the recommended minimum of 20 participants per cell.

Further research should consider other potential variables in the model. For example, attitude satisfaction as a dependent variable (Zhou and Chang, 2024), different levels of severity of the service failure (Shams et al., 2024), and different types of service failure (outcome errors vs. process errors) (Liu et al., 2023) could be included. Moreover, research is needed on gender matching between users and chatbots – for example, the contexts in which matching is adequate. Furthermore, the potential negative effects of gender matching in service recovery should be investigated – for example, when a higher similarity between the users and the chatbot can actually be a hindrance. With regard to theory, the focus has been placed on the impact of failure and recovery on justice dimensions (Blut et al., 2021). Further research should address issues of attribution theory – for example, is the chatbot responsible for the service failure or the company? How does this change after a service recovery, especially considering the chatbot's design (e.g., anthropomorphism, gender)? Moreover, research should investigate when warmer and more competent designs are more important in the service failure and recovery process. Warmth could be more important for relationship-oriented people and competence for transaction-oriented users (Huang and Ha, 2020). Researchers should also examine when gender-neutral designs work well. In addition, the use of field experiments would facilitate a realistic appraisal of unresolved research areas.

6. Conclusion

The potential of chatbots to increase a company's efficiency is hampered by their susceptibility to errors, which often leads to user expectations of chatbots not being met. This exacerbates the existing conflict between customer service costs and service quality (Adam et al., 2021). Self-recovery by chatbots is strategically important to reduce the susceptibility to errors and ensure service quality. Anthropomorphic design elements in particular with regard to perceived competence are less important. Ultimately, a high error rate can lead to the discontinuation of the chatbot (Feine et al., 2019). So far, chatbots in customer service are mainly suitable for simple, repetitive service requests and less serviceable for more complex tasks (Gnewuch and Maedche, 2022). However, recent developments in NLP further simplify imitating human conversations and implementing chatbots (Caldarini et al., 2022). Thus, a reduction in susceptibility to errors is foreseeable with the practical implementation of increasing intelligence in chatbots. However, in the medium term, close cooperation between customer service employees and chatbots is likely to continue (Gnewuch and Maedche, 2022; Huang and Dootson, 2022), spotlighting again anthropomorphic design elements.

CRediT authorship contribution statement

Alexandra Rese: Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Lennart Witthohn:** Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

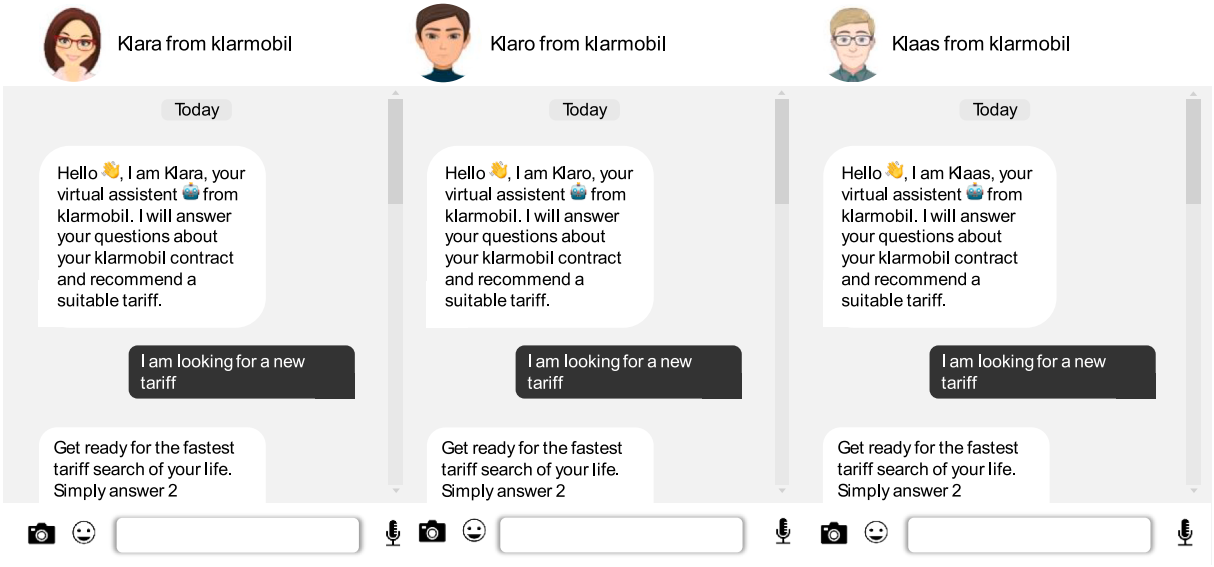
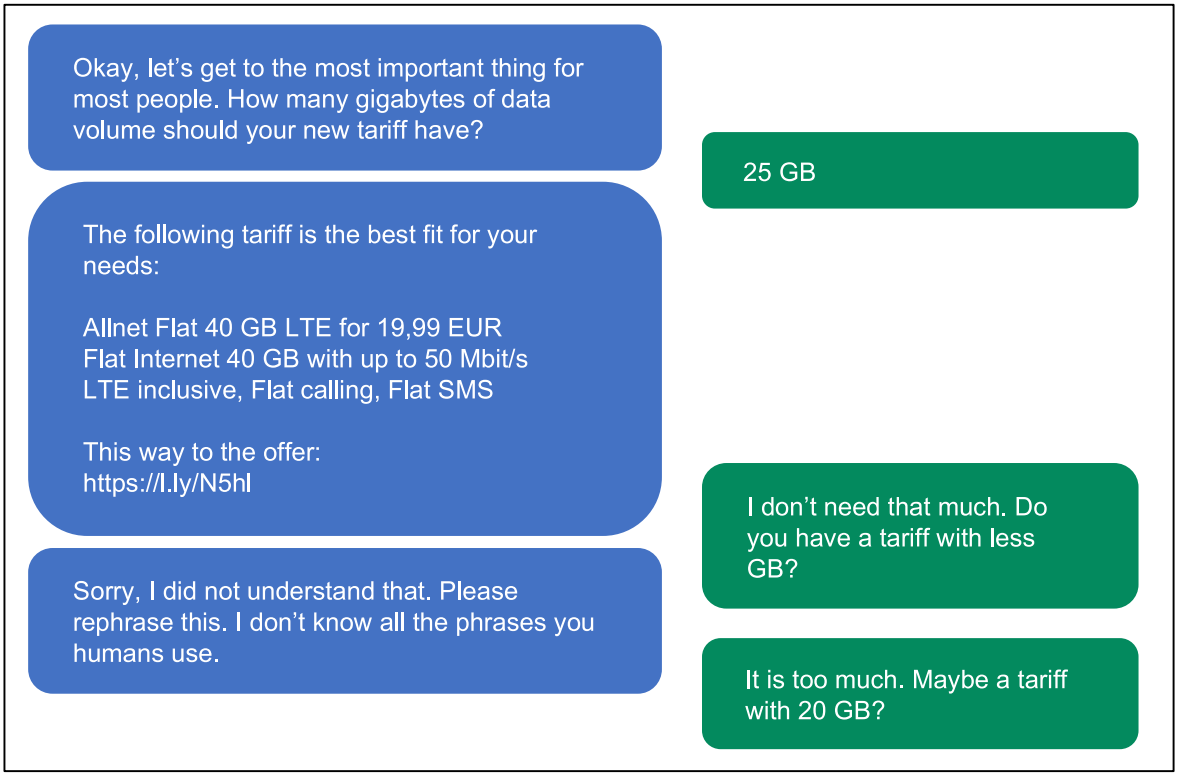


Fig. A1. Stimuli and starting point of the chatbot conversation (translated from German).



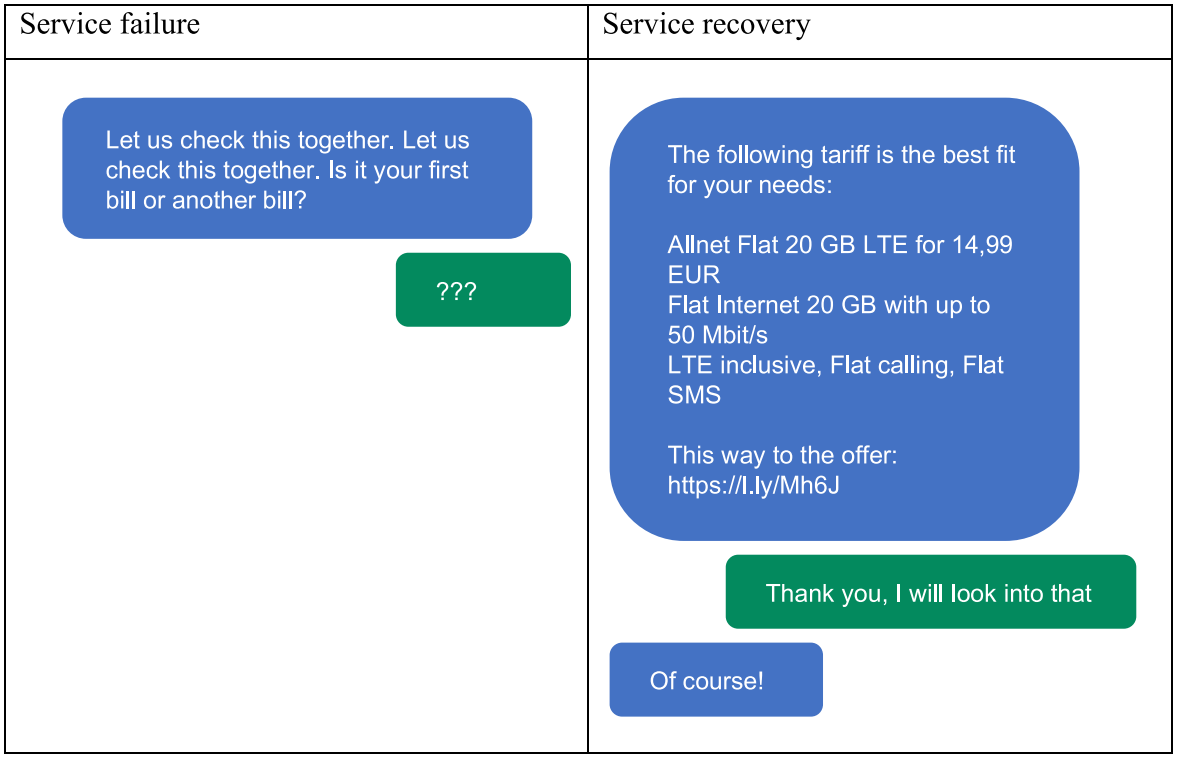


Fig. A2. Service failure and recovery scenario of the chatbot conversation.

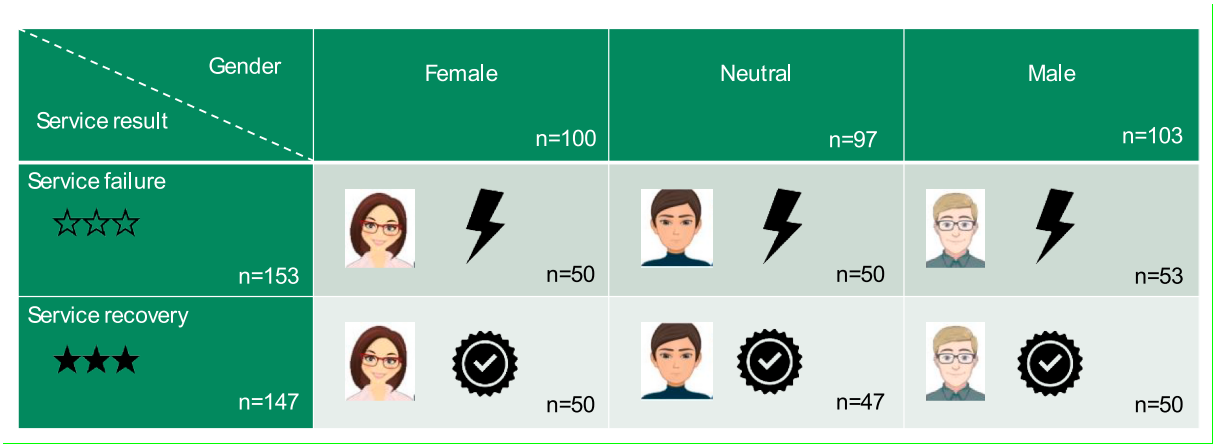


Fig. A3. Sample distribution across the six experimental groups.

Table A1
Literature review of chatbot service recovery

Reference	Context	Theory bases	Methodology	Sample	Results
Theory of justice					
Fota et al. (2022)	Chatbot in the complaint management of a retailer. Complaint: broken headphones	Social presence theory computers are social actors paradigm, distributive justice (voucher)	Scenario-based experiment	389 German participants, recruited randomly (social media channels, online forums)	A human-like avatar, an empathetic response and compensation (voucher) increase the intention to repurchase and have a positive effect on perceived humanness and evaluation of service recovery. Anthropomorphism and evaluation of redress positively influence repurchase intention (mediation).
Han et al. (2022)	Chatbot of an online food delivery service	Interactional justice (empathy), social cognition	Scenario-based experiment	Study 1: 95 US participants Study 2: 98 US participants (recruited from students)	Empathy leads to a better evaluation of both the perceived warmth and competence of the chatbot, which in turn increases the perceived service quality and satisfaction with the chatbot. In the case of a conversational breakdown resulting from a chatbot failure, empathy makes the chatbot appear significantly less competent. There is no significant effect on perceived warmth.
Markovitch et al. (2024)	Chatbot of an online service (vacation, smartphone purchase, medical advice)	Interactional justice (empathy)	Scenario-based experiment, quasi-experiment	Study 1: 199 participants Study 2: 200 participants Study 3: 315 participants (recruited from MTurk and Prolific) Study 4: 100 participants	Users were less satisfied with the chatbot in negative outcome situations compared with human employees. However, if the chatbot uses an empathic communication style, the perceived empathy can increase chatbot's evaluation so that it catches up with the human employee.
Song et al. (2023)	Chatbots in different contexts (retail, hotel industry, delivery service)	Politeness theory, procedural justice (time pressure), interactional justice (apology, appreciation)	Scenario-based experiment	Study 1: 187 Chinese participants, Study 1B: 214 Chinese participants; Study 2: 125 Chinese participants, Study 3: 221 Chinese participants (recruited from Credamo)	Appreciation and apology have a positive effect on satisfaction after recovery. The appreciation works significantly better than the apology. A combination of strategies is not significantly better than the appreciation strategy alone. Perceived face concern mediates the effect of the strategy on satisfaction after recovery. When time pressure is high, the main effect and the mediation are no longer significant.
Zhang et al. (2023)	Chatbot in the tourism sector (airline, hotel)	Interactional justice (apology), symbolic service recovery, emotional competence theory	Scenario-based experiment	Study 1: 163 Chinese participants, Study 2: 390 Chinese participants (recruited from Credamo)	Users were less satisfied with the service recovery when the apology came from a chatbot (vs. employee). This is due to a lower perceived naturalness and sincerity of the chatbot. A higher perceived emotional intelligence of the chatbot leads to a

(*continued on next page*)

Table A1 (*continued*)

Reference	Context	Theory bases	Methodology	Sample	Results
					higher perceived naturalness and sincerity.
Zhu et al. (2023)	Voice-based chatbot in the tourism sector (hotel, restaurant)	Interactional justice (apology), distributive justice (coupon), procedural justice (response time)	Scenario-based experiment	Study 1: 220 Chinese participants, Study 2: 430 Chinese participants, (recruited from Credamo)	An apology from the chatbot resulted in lower satisfaction scores and revisit intentions than with an employee. With economic recovery, there was no difference between the chatbot and the employee. The effect occurs in the case of delayed recovery, but not immediate recovery.
Attribution theory					
Pavone et al. (2023)	Chatbots in the airline industry	Cognitive appraisal theory of emotions, attribution theory, theories on anthropomorphism	Scenario-based experiment	Study 1: 122 respondents. Study 2: 120 participants (recruited by a professional panel provider), Study 3: 120 U.S. participants (recruited by a professional panel provider)	Customers consider chatbots are not responsible for a service failure due to having no control or specific intentions; they blame the company. Anthropomorphic design elements help to place less blame on the company and support problem-oriented coping strategies.
Emotions in service recovery					
Zhang et al. (2022)	Chatbot in online shopping, service failure: delivery delay	Cute apology strategies (childish, playful/ humorous)	Scenario-based experiment	Study 1: 157 Chinese participants, Study 2: 316 Chinese participants, (recruited from WenJuanXing)	The chatbot with kindchenschema and the whimsical chatbot reduce negative emotions. Both chatbot types trigger significantly fewer negative emotions at a low failure severity level. At a high failure severity level, this holds to a lesser extent only for the kindchenschema chatbot. The kindchenschema is more effective for female users and those with a low fear of technology. The playful chatbot is more effective for male users and users with a higher fear of technology.
Liu et al. (2023)	Chatbot in customer service (household devices, printing devices, energy provider)	Emojis	Scenario-based experiment	Study 1: 142 Chinese participants recruited from undergraduate students, Study 2: 131 Chinese participants recruited randomly (social media channels), Study 3: Chinese participants recruited from undergraduate students	The use of a smiley after a service failure (process or outcome failure) leads to a higher willingness to use the chatbot again. This effect is mediated by the perceived intelligence of the chatbot.
Yang et al. (2023)	Chatbot in online shopping (VR glasses, chocolate)	Self-deprecating humour responses	Scenario-based experiment	Study 1: 117 Chinese participants Study 2: 196 Chinese participants	Satisfaction with the service recovery was significantly higher with the humorous chatbot. The perceived sincerity and perceived intelligence of the chatbot mediate the effect of humour on satisfaction. Sense of power moderates the effect.
Shams et al. (2024)	Chatbot in customer service (hotel) service failure: waiting time for room service	Humour, chatbot communication styles	Scenario-based experiment	Study 1: 460 respondents Study 2: 333 respondents (recruited from Prolific)	Satisfaction with the service recovery can be increased if humour and informal language are matched, especially in the case of low-equity brands and failures of low severity.
Handover to employees					
Song et al. (2022)	Chatbots in the hotel industry and online shopping	Social response theory	Scenario-based experiment	Study 1: 107 Chinese participants Study 2: 104 Chinese participants	Satisfaction with the chatbot is higher if it is able to provide service recovery itself (vs. handing it over to employees). This is due to a higher perceived functional value (mediator) in the case of chatbot self-recovery. Privacy risk concerns are higher when handing over to the employee, which leads to a low level of satisfaction with the handover strategy. Intelligence has a positive effect on the functional value of the chatbot but also increases privacy risk concerns.
Xing et al. (2022)	Chatbots in online retail	Role congruence theory, mental accounting theory	SEM (AMOS)	N = 521 Chinese participants, recruited from WenJuanXing	In the case of an outcome service failure, the participants favour the chatbot; in the case of a process service failure, they favour the employee. Service recovery by the chatbot increases perceived fairness, perceived data protection, and perceived friendliness. Participants are more willing to forgive the service failure if the friendliness and perceived data

(*continued on next page*)

Table A1 (*continued*)

Reference	Context	Theory bases	Methodology	Sample	Results
					protection are higher. If the intelligence of the chatbot is higher, there is a greater need for employee involvement – both for an outcome and a service failure process.
Repair strategies					
Ashktorab et al. (2019)	Chatbot in the retail/ banking/ravel industry	Framework 'Grounding in Communication'	Paired comparison experiment	N = 203 participants from Amazon Turk, 1624 pairwise comparisons	The chatbot should show initiative and provide options and explanations. Chatbots should openly admit when they do not understand something but avoid redundancies. Repair strategies should be adapted to the context and individuals. Users prefer co-creation in the repair process.
Zhou and Chang (2024)	Chatbot in the airline industry	Social support theory, social cognition	Scenario-based experiment	Study 1: 382 Chinese participants Study 2: 771 Chinese participants Study 3: 769 participants (recruited from WenJuanXing)	The effect of informational self-recovery is higher on consumer quality satisfaction than emotional self-recovery, while the effect is reversed for consumer attitude satisfaction. Informational self-recovery relates to perceived competence and service process failure, while for emotional self-recovery these are perceived warmth and service outcome failure.
Our study	Chatbot in the telecommunication industry	Interactional justice (empathy), social cognition	Scenario-based experiment	300 German participants recruited randomly (social media channels, online forums)	Service recovery is most relevant for satisfaction with the chatbot. Anthropomorphic design elements come second place. Perceived competence of the chatbot is more important than perceived warmth, and gender mismatch can increase perceived humanness.

Table A2
Socio-demographics

Demographics	Specifications	n	%
Gender	Female	178	59.3
	Male	121	40.3
	Diverse	1	0.3
Age	16–24 years	91	30.3
	25–34 years	153	51.0
	35–44 years	21	7.0
	45–54 years	17	5.0
	55–64 years	16	5.0
	Over 65 years	2	0.7
Monthly household net income	<500 euros	21	
	<500 euros	21	7.0
	500-999 euros	65	21.7
	1000–1499 euros	48	16.0
	1500–1900 euros	26	8.7
	2000–2499 euros	14	4.7
	2500–2999 euros	18	6.0
	3000–4999 euros	55	18.3
	5000 euros and more	29	9.7
	No answer	24	8.0
Employment status	Pupil	4	1.3
	Student	186	62.0
	Employed	98	32.7
	Self-employed	6	2.0
	Housewife/houseman	3	1.0
	No answer	3	1.0
Education	Intermediate school certificate	3	1,0
	University (of Applied Sciences) entrance level or equivalent level	69	23.0
	Completed vocational training	17	5.7
	Completed vocational training at a master craftspeople or technical school	24	8.0
	Bachelor's degree	119	39.7
	Master's degree, diploma, state examination	58	19.3
	PhD	2	0.7
	No answer	3	1.0
Rate of change of mobile phone contracts	Every year	1	0.3
	Every two years	51	17.0

(*continued on next page*)

Table A2 (*continued*)

Demographics	Specifications	n	%
	Every 3–4 years	68	22.7
	Every 5–6 years	40	13.3
	Less often	81	27.0
	Never	59	19.7

n = 300.

Table A3
Chatbot usage behaviour

Characteristics	Specifications	n	(%)
Usage frequency (n = 262)	Several times a week	18	6.9
	About 1 time per week	15	5.7
	Several times a month	37	14.1
	About 1 time per month	30	11.5
	About every 2–3 months	67	25.6
	About every six months	45	17.2
	Less often	50	19.1
Usage duration (n = 262)	For more than 5 years	27	10.3
	For about 5 years	31	11.8
	For about 3 years	106	40.5
	For about 1 year	55	21.0
	For a few months	16	6.1
	Only recently	27	10.3
Usage purpose (multiple answers) (n = 262)	For service requests	218	72.7
	As a search engine (e.g., ChatGPT)	111	37.0
	For creating texts (e.g., ChatGPT)	99	33.0
	To try out	78	26.0
	Internally as a company chatbot (e.g., in the HR department, IT department)	29	9.7
	To search for or purchase products	26	8.7
	For recommendations	19	6.3
	Others	7	2.3
Usage of a chatbot from a telecommunications company (n = 262)	Yes	95	36.3
	No	167	63.7
Usage of a chatbot to search for a mobile phone contract (n = 95)	Yes	26	27.4
	No	69	72.6
Brand (multiple answers) (n = 26)	German Telekom	13	50.0
	Vodafone	5	19.2
	O2	2	7.7
	A1	2	7.7
	Klarmobil, Magenta, Congstar, 1&1	1	3.8

Table A4
Measurement items

Factor (Source)	Item	Mean	Std.	Factor loading	α	Variance extracted
Warmth (Choi et al., 2021)	Please rate Klara/Klaro/Klaas regarding the following characteristics:	4.31	1.20		0.86	0.65
	[Chatbot name] is …					
	… caring.	3.72	1.60	0.820		
	… friendly.	5.33	1.29	0.777		
	… kind.	4.25	1.52	0.842		
	… warm.	3.74	1.56	0.825		
	… sociable.	4.49	1.49	0.759		
Competence (Choi et al., 2021)	Please rate Klara/Klaro/Klaas regarding the following characteristics:	4.27	1.38		0.87	0.72
	[Chatbot name] is …					
	… intelligent.	3.55	1.66	0.795		
	… energetic.	4.42	1.62	0.884		
	… organized.	4.29	1.63	0.894		
	… motivated.	4.84	1.64	0.806		
Perceived humanness (Bartneck et al., 2009)	Please rate Klara/Klaro/Klaas regarding the following characteristics (seven-point bipolar adjective scale):	3.14	1.21		0.85	0.65
	1 = fake vs. 7 = natural	3.42	1.64	0.886		
	1 = machinelike vs. 7 = humanlike	2.88	1.59	0.872		
	1 = artificial vs. 7 = lifelike	2.94	1.46	0.896		
	1 = unconscious vs. 7 = conscious	2.35	1.46	0.674		
	1 = communicates rigidly vs. 7 = communicates elegantly	4.11	1.52	0.587		

(*continued on next page*)

Table A4 (*continued*)

Factor (Source)	Item	Mean	Std.	Factor loading	α	Variance extracted
Satisfaction (Chung et al., 2020)	To what extent do you agree with the following statements?	3.48	1.99		0.97	0.92
	The user can be satisfied with [chatbot name].	3.34	1.98	0.962		
	[Chatbot name] did a good job.	3.47	1.96	0.970		
	[Chatbot name] did what was expected of him/her.	3.79	2.17	0.935		
	The conversation with [chatbot name] was satisfactory.	3.33	2.18	0.968		
Prior experience with chatbots (Lacey et al., 2010)		3.31	1.56		0.94	0.85
	I have a lot of experience with chatbots.	3.33	1.66	0.957		
	I am very familiar with chatbots.	3.55	1.68	0.955		
	I know my way around chatbots.	3.36	1.68	0.928		
	I use chatbots on a regular basis.	3.02	1.76	0.847		
Need for interaction (Dabholkar, 1996)		4.67	1.39		0.73 (0.65)	0.66 (0.52)
	Human contact in providing services makes the process enjoyable for the customer.	5.05	1.55	0.834 (0.826)		
	I like interacting with the person who provides the service.	4.95	1.61	0.876 (0.848)		
	Personal attention by the service employee is not very important to me (reverse-scored).	*3.90*	*1.893*	*0.409*		
	It bothers me to use a machine when I could talk to a person instead.	4.02	1.98	0.725 (0.717)		
Technology innovativeness (Parasuraman, 2000)		4.30	1.41		0.91	0.74
	I always like to try out the latest technologies.	4.86	1.46	0.849		
	In my circle of friends, I am among the first when it comes to using new technologies.	3.68	1.70	0.874		
	I enjoy the challenge of figuring out high-tech gadgets.	4.10	1.73	0.886		
	I have fewer problems than other people in making technology work for me.	4.65	1.54	0.865		
	Other people ask me for advice when it comes to using new technologies.	4.21	1.77	0.814		
Negative attitude towards situations concerning interactions with chatbots (Nomura et al., 2006)		3.40	1.22		0.65 (0.62)	0.49 (0.41)
	I would feel uneasy if I was given a job where I had to use chatbots.	3.72	1.89	0.764 (0.743)		
	The word "chatbot" means nothing to me.	*3.77*	*1.933*	*0.381*		
	I would feel nervous operating a chatbot in front of other people.	2.62	1.62	0.638 (0.614)		
	I would hate the idea that chatbots or artificial intelligences were making judgments about things.	4.40	1.78	0.654 (0.660)		
	I would feel paranoid talking with a chatbot.	2.87	1.66	0.742 (0.734)		
Negative attitude towards social influence of chatbots (Nomura et al., 2006)		3.87	1.28		0.76	0.52
	I would feel uneasy if chatbots really had emotions.	4.81	1.86	0.619		
	Something bad might happen if chatbots developed into living beings.	4.54	1.91	0.808		
	I feel that if I depend on chatbots too much, something bad might happen.	3.70	1.73	0.793		
	I am concerned that chatbots would be a bad influence on children.	3.42	1.75	0.743		
	I feel that, in the future, society will be dominated by chatbots.	2.89	1.66	0.606		
Negative attitude towards emotions in interaction with chatbots (Nomura et al., 2006)		5.00	1.13		0.55	0.55
	I would feel relaxed talking with a chatbot (reverse-scored).	3.86	1.67	0.478		
	If chatbots had emotions, I would be able to make friends with them (reverse-scored).	5.81	1.56	0.828		
	I feel comforted being with chatbots that have emotions (reverse-scored).	5.46	1.42	0.859		

In italics: items dropped; in brackets: values before removing items.

Table A5
Descriptive statistics and bivariate correlations across the six experimental groups.

	Mean (std.)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Perceived warmth (1)								
a	4.35 (1.199)							
b	4.25 (1.057)							
c	3.77 (1.208)							
d	4.70 (1.203)							
e	4.38 (1.282)							
f	4.42 (1.116)							
Perceived competence (2)								
a	3.89 (1.379)	0.721***						

(*continued on next page*)

Table A5 (*continued*)

	Mean (std.)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
b	3.92 (1.189)	0.643***						
c	3.58 (1.447)	0.737***						
d	4.95 (1.037)	0.656***						
e	4.62 (1.455)	0.761***						
f	4.76 (1.204)	0.632***						
Perceived humanness (3)								
a	2.94 (1.041)	0.469**	0.532**					
b	2.96 (1.006)	0.486***	0.408**					
c	2.74 (1.239)	0.582***	0.704***					
d	3.44 (1.197)	0.473**	0.434**					
e	3.35 (1.387)	0.674***	0.663***					
f	3.44 (1.216)	0.502***	0.341*					
Satisfaction (4)								
a	1.83 (0.761)	0.380**	0.521***	0.381**				
b	1.90 (0.930)	0.270+	0.277+	0.516***				
c	1.72 (0.913)	0.400**	0.524***	0.466***				
d	5.25 (1.107)	0.631***	0.525***	0.442**				
e	5.32 (1.296)	0.620***	0.623***	0.540***				
f	5.10 (1.051)	0.421**	0.505***	0.294*				
Prior experience with chatbots (5)								
a	3.08 (1.498)	0.169	0.158	0.249+	0.003			
b	3.16 (1.430)	−0.057	−0.064	0.110	−0.023			
c	3.30 (1.485)	0.216	0.070	−0.011	−0.089			
d	3.87 (1.507)	0.201	0.174	0.216	0.226			
e	3.28 (1.838)	0.205	0.106	0.004	0.152			
f	3.21 (1.554)	0.313*	0.366**	0.114	0.148			
Need for interaction (6)								
a	4.79 (1.330)	−0.143	−0.072	−0.190	0.119	−0.274+		
b	4.44 (1.524)	−0.033	−0.003	−0.046	−0.330*	−0.143		
c	4.62 (1.550)	−0.118	0.016	0.039	0.127	−0.322*		
d	4.78 (1.321)	−0.287*	−0.214	−0.281*	−0.261+	−0.218		
e	4.90 (1.194)	0.095	0.032	0.098	0.053	0.202		
f	4.53 (1.355)	0.223	0.111	0.071	0.177	−0.388**		
Technology innovativeness (7)								
a	4.29 (1.480)	−0.107	0.007	−0.107	0.015	0.232	−0.149	
b	4.31 (1.144)	−0.020	−0.007	0.017	0.050	0.422**	−0.115	
c	4.04 (1.538)	−0.123	0.179	−0.234+	−0.163	−0.339*	0.006	
d	4.40 (1.396)	0.162	0.142	0.144	0.402**	0.564**	−0.219	
e	4.54 (1.385)	−0.058	−0.164	−0.009	−0.024	0.529***	0.070	
f	4.24 (1.473)	0.278+	0.336*	0.067	0.146	0.486***	−0.232	
Negative attitude towards situations concerning interactions with chatbots (8)								
a	3.37 (1.171)	−0.138	−0.091	−0.069	0.164	−0.181	0.469**	−0.217
b	3.13 (1.084)	0.052	0.128	0.032	0.083	−0.089	0.426**	−0.152
c	3.47 (1.255)	−0.209	−0.143	−0.071	−0.046	−0.119	0.300*	0.181
d	3.28 (1.280)	−0.355*	−0.359*	−0.464**	−0.368*	−0.281*	0.566***	−0.321*
e	3.60 (1.206)	0.110	0.157	0.203	0.150	−0.212	0.362*	−0.196
f	3.57 (1.296)	0.078	−0.003	−0.117	−0.035	−0.245+	0.403**	−0.207

a: service failure – female chatbot, b: service failure – gender-neutral chatbot, c: service failure – male chatbot, d: service recovery – female chatbot, e: service recovery – gender-neutral chatbot, f: service recovery – male chatbot Significance levels: + $p < 0.10$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Data availability

Data will be made available on request.

References

Aaker, J., Vohs, K.D., Mogilner, C., 2010. Nonprofits are seen as warm and for-profits as competent: firm stereotypes matter. J. Consum. Res. 37 (2), 224–237.

Aaker, J.L., Garbinsky, E.N., Vohs, K.D., 2012. Cultivating admiration in brands: warmth, competence, and landing in the "golden quadrant". J. Consum. Psychol. 22 (2), 191–194.

Adam, M., Wessel, M., Benlian, A., 2021. AI-based chatbots in customer service and their effects on user compliance. Electron. Market 31 (2), 427–445.

Adamopoulou, E., Moussiades, L., 2020. An overview of chatbot technology. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (Eds.), Artificial Intelligence Applications and Innovations, vol. 584. Springer International Publishing, Cham, pp. 373–383.

Ahn, J., Kim, J., Sung, Y., 2022. The effect of gender stereotypes on artificial intelligence recommendations. J. Bus. Res. 141, 50–59.

Akdim, K., Belanche, D., Flavián, M., 2023. Attitudes toward service robots: analyses of explicit and implicit attitudes based on anthropomorphism and construal level theory. IJCHM 35 (8), 2816–2837.

Alaei, R., Deska, J.C., Hugenberg, K., Rule, N.O., 2022. People attribute humanness to men and women differently based on their facial appearance. J. Pers. Soc. Psychol. 123 (2), 400–422.

Alhouti, S., Wright, S.A., Baker, T.L., 2019. Responding to service failures with prevention framed donations. JSM 33 (5), 547–556.

Ashktorab, Z., Jain, M., Liao, Q.V., Weisz, J.D., 2019. Resilient chatbots. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems, Glasgow Scotland UK. 04 05 2019 09 05 2019. ACM, New York, NY, USA, pp. 1–12.

Aumüller, A., Winklbauer, A., Schreibmaier, B., Batinic, B., Mara, M., 2024. Rethinking feminized service bots: user responses to abstract and gender-ambiguous chatbot avatars in a large-scale interaction study. Personal Ubiquitous Comput. 28, 1021–1032.

Bartneck, C., Kulić, D., Croft, E., Zoghbi, S., 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int. J. Soc. Robotics 1 (1), 71–81.

Bayes, M.A., 1972. Behavioral cues of interpersonal warmth. J. Consult. Clin. Psychol. 39 (2), 333–339.

Belanche, D., Casaló, L.V., Flavián, C., Schepers, J., 2020. Robots or frontline employees? Exploring customers' attributions of responsibility and stability after service failure or success. JOSM 31 (2), 267–289.

Belanche, D., Casaló, L.V., Schepers, J., Flavián, C., 2021. Examining the effects of robots' physical appearance, warmth, and competence in frontline services: the Humanness-Value-Loyalty model. Psychol. Market. 38 (12), 2357–2376.

Beldad, A., Hegner, S., Hoppen, J., 2016. The effect of virtual sales agent (VSA) gender – product gender congruence on product advice credibility, trust in VSA and online vendor, and purchase intention. Comput. Hum. Behav. 60, 62–72.

Ben Mimoun, M.S., Poncin, I., Garnier, M., 2012. Case study—embodied virtual agents: an analysis on reasons for failure. J. Retailing Consum. Serv. 19 (6), 605–612.

Benbasat, I., Dimoka, A., Pavlou, P.A., Qiu, L., 2020. The role of demographic similarity in people's decision to interact with online anthropomorphic recommendation agents: evidence from a functional magnetic resonance imaging (fMRI) study. Int. J. Hum. Comput. Stud. 133, 56–70.
Benner, D., Elshan, E., Schöbel, S.M., Janson, A., 2021. What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents. In: Proceedings of the 42nd International Conference on Interaction Sciences Information Systems (ICIS 2021), Austin, Texas.
Bitner, M.J., Brown, S.W., Meuter, M.L., 2000. Technology infusion in service encounters. JAMS 28 (1), 138–149.
Blodgett, J.G., Wakefield, K.L., Barnes, J.H., 1995. The effects of customer service on consumer complaining behavior. JSM 9 (4), 31–42.
Blut, M., Wang, C., Wünderlich, N.V., Brock, C., 2021. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. JAMS 49 (4), 632–658.
Bolton, L.E., Mattila, A.S., 2015. How does corporate social responsibility affect consumer response to service failure in buyer–seller relationships? J. Retailing 91 (1), 140–153.
Borau, S., Otterbring, T., Laporte, S., Fosso Wamba, S., 2021. The most human bot: female gendering increases humanness perceptions of bots and acceptance of AI. Psychol. Market. 38 (7), 1052–1068.
Boshoff, C., 2012. A neurophysiological assessment of consumers' emotional responses to service recovery behaviors. J. Serv. Res. 15 (4), 401–413.
Bougie, R., Pieters, R., Zeelenberg, M., 2003. Angry customers don't come back, they get back: the experience and behavioral implications of anger and dissatisfaction in services. JAMS 31 (4), 377–393.
Cahn, J., 2017. CHATBOT: Architecture, Design, & Development. Senior Thesis (EAS499). University of Pennsylvania, School of Engineering and Applied Science, Department of Computer and Information Science.
Caldarini, G., Jaf, S., McGarry, K., 2022. A literature survey of recent advances in chatbots. Information 13 (1), 41. Information 13 (1).
Cammarota, A., D'Arco, M., Marino, V., Resciniti, R., 2023. Brand activism: a literature review and future research agenda. Int. J. Consum. Stud. 47 (5), 1669–1691.
Castro, I., Thompson, S., Ward, J., 2012. The importance of warmth and competence in the acquisition and retention of new customers. NA – Adv. Consum. Res. 40, 947–948.
Chen, Q., Gong, Y., Lu, Y., Luo, X., 2024. The golden zone of AI's emotional expression in frontline chatbot service failures. Internet Res. https://doi.org/10.1108/INTR-07-2023-0551.
Choi, S., Mattila, A.S., Bolton, L.E., 2021. To err is human (-oid): how do consumers react to robot service failure and recovery? J. Serv. Res. 24 (3), 354–371.
Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Lee, L., Zemmel, R., 2023. The Economic Potential of Generative AI. McKinsey & Company. http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/14313/1/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf.
Chung, M., Ko, E., Joung, H., Kim, S.J., 2020. Chatbot e-service and customer satisfaction regarding luxury brands. J. Bus. Res. 117, 587–595.
Clark, H.H., Brennan, S.E., 2004. Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (Eds.), Perspectives on Socially Shared Cognition, 4, printing ed. American Psychological Association, Washington, D.C., pp. 127–149
Coniam, D., 2014. The linguistic accuracy of chatbots: usability from an ESL perspective. Text Talk 34 (5), 545–567.
Crolic, C., Thomaz, F., Hadi, R., Stephen, A.T., 2022. Blame the bot: anthropomorphism and anger in customer–chatbot interactions. J. Market. 86 (1), 132–148.
Dabholkar, P.A., 1996. Consumer evaluations of new technology-based self-service options: an investigation of alternative models of service quality. Int. J. Res. Market. 13 (1), 29–51.
Del Río-Lanza, A.B., Vázquez-Casielles, R., Díaz-Martín, A.M., 2009. Satisfaction with service recovery: perceived justice and emotional responses. J. Bus. Res. 62 (8), 775–781.
Derrick, B., White, P., 2016. Why Welch's test is Type I error robust. TQMP 12 (1), 30–38.
Dharaniya, R., Vijayalakshmi, K., Tejasree, R., Naveena, P., 2020. Survey on interactive chatbot. IJRASET 8 (6), 1698–1704.
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. THCI, pp. 82–103.
Diekman, A.B., Eagly, A.H., 2000. Stereotypes as dynamic constructs: women and men of the past, present, and future. Pers. Soc. Psychol. Bull. 26 (10), 1171–1188.
Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. 144 (1), 114–126.
Eagly, A.H., Steffen, V.J., 1984. Gender stereotypes stem from the distribution of women and men into social roles. J. Pers. Soc. Psychol. 46 (4), 735–754.
Ebert, I.D., Steffens, M.C., Kroth, A., 2014. Warm, but maybe not so competent?—contemporary implicit stereotypes of women and men in Germany. Sex. Roles 70 (9–10), 359–375.
Epley, N., Waytz, A., Cacioppo, J.T., 2007. On seeing human: a three-factor theory of anthropomorphism. Psychol. Rev. 114 (4), 864–886.
Ernst, C.P., Herm-Stapelberg, N., 2020. Gender stereotyping's influence on the perceived competence of siri and Co. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 4448–4453.
Esmark Jones, C.L., Hancock, T., Kazandjian, B., Voorhees, C.M., 2022. Engaging the Avatar: the effects of authenticity signals during chat-based service recoveries. J. Bus. Res. 144, 703–716.
Eyssel, F., Häring, M., 2012. (S) he's got the look: gender stereotyping of robots. J. Appl. Soc. Psychol. 42 (9), 2213–2230.
Feine, J., Gnewuch, U., Morana, S., Maedche, A., 2020. Gender bias in chatbot design. In: Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.-C., Granmo, O.-C., Luger, E., Brandtzaeg, P.B. (Eds.), Chatbot Research and Design, vol. 11970. Springer International Publishing, Cham, pp. 79–93.
Feine, J., Morana, S., Gnewuch, U., 2019. Measuring service encounter satisfaction with customer service chatbots using sentiment analysis. In: Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI2019). Siegen, Germany. February 24–27.
Fetscherin, M., Tantleff-Dunn, S., Klumb, A., 2020. Effects of facial features and styling elements on perceptions of competence, warmth, and hireability of male professionals. J. Soc. Psychol. 160 (3), 332–345.
Fiore, D., Baldauf, M., Thiel, C., 2019. Forgot Your Password Again?" Acceptance and user experience of a chatbot for in-company IT support. In: Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia, pp. 1–11.
Fiske, S.T., Cuddy, A.J.C., Glick, P., 2007. Universal dimensions of social cognition: warmth and competence. Trends Cognit. Sci. 11 (2), 77–83.
Fiske, S.T., Cuddy, A.J.C., Glick, P., Xu, J., 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. J. Pers. Soc. Psychol. 82 (6), 878–902.
Følstad, A., Taylor, C., 2020. Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives. In: Følstad, A., Araujo, T., Papadopoulos, S., Law, E.L.-C., Granmo, O.-C., Luger, E., Brandtzaeg, P.B. (Eds.), Chatbot Research and Design, vol. 11970. Springer International Publishing, Cham, pp. 201–214.
Foster, C., Resnick, S., 2013. Service worker appearance and the retail service encounter: the influence of gender and age. Serv. Ind. J. 33 (2), 236–247.
Fota, A., Wagner, K., Roeding, T., Schramm-Klein, H., 2022. Help! I have a problem –Differences between a humanlike and robot-like chatbot avatar in complaint management. In: Proceedings of the 55th Hawaii International Conference on System Sciences 2021, pp. 4273–4282.
Fotheringham, D., Wiles, M.A., 2023. The effect of implementing chatbot customer service on stock returns: an event study analysis. JAMS 51 (4), 802–822.
Fugate, D.L., Phillips, J., 2010. Product gender perceptions and antecedents of product gender congruence. J. Consum. Market. 27 (3), 251–261.
Gelbrich, K., Hagel, J., Orsingher, C., 2021. Emotional support from a digital assistant in technology-mediated services: effects on customer satisfaction and behavioral persistence. Int. J. Res. Market. 38 (1), 176–193.
Gnewuch, U., Maedche, A., 2022. Hybride mensch-KI service-agenten. In: Bruhn, M., Hadwich, K. (Eds.), Smart Services. Springer, Fachmedien Wiesbaden, Wiesbaden, pp. 63–77.
Gnewuch, U., Morana, S., Adam, M., Maedche, A., 2018. Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. In: Proceedings of the 26th European Conference on Information Systems (ECIS2018). Portsmouth, UK.
Gnewuch, U., Yu, M., Maedche, A., 2020. The effect of perceived similarity in dominance on customer self-disclosure to chatbots in conversational commerce. In: Proceedings of the 28th European Conference on Information Systems (ECIS 2020) – A Virtual AIS Conference.
Go, E., Sundar, S.S., 2019. Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. Comput. Hum. Behav. 97, 304–316.
Grant, S.L., Mizzi, T., Anglim, J., 2016. 'Fat, four-eyed and female' 30 years later: a replication of Harris, Harris, and Bochner's (1982) early study of obesity stereotypes. Aust. J. Psychol. 68 (4), 290–300.
Grégoire, Y., Mattila, A.S., 2021. Service failure and recovery at the crossroads: recommendations to revitalize the field and its influence. J. Serv. Res. 24 (3), 323–328.
Grice, H., 1975. Logic and conversation. Syntax and Semantics 3: Speech Acts. Academic Press, New York, pp. 41–58.
Güntürkün, P., Haumann, T., Mikolon, S., 2020. Disentangling the differential roles of warmth and competence judgments in customer-service provider relationships. J. Serv. Res. 23 (4), 476–503.
Guydish, A.J., Fox Tree, J.E., 2021. Good conversations: grounding, convergence, and richness. New Ideas Psychol. 63, 100877.
Han, E., Yin, D., Zhang, H., 2022. Chatbot empathy in customer service: when it works and when it backfires. SIGHCI 2022 Proceedings 1–7. https://aisel.aisnet.org/sighci2022/1.
Harris, M.B., Harris, R.J., Bochner, S., 1982. Fat, four-eyed, and female: stereotypes of obesity, glasses, and gender 1. J. Appl. Soc. Pyschol. 12 (6), 503–516.
Hayes, A.F., 2022. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach, third ed. The Guilford Press, New York, London, p. 732.
Heflick, N.A., Goldenberg, J.L., Cooper, D.P., Puvia, E., 2011. From women to objects: appearance focus, target gender, and perceptions of warmth, morality and competence. J. Exp. Soc. Psychol. 47 (3), 572–581.
Hess Jr., R.L., Ganesan, S., Klein, N.M., 2003. Service failure and recovery: the impact of relationship factors on customer satisfaction. JAMS 31 (2), 127–145.
Holloway, B.B., Beatty, S.E., 2003. Service failure in online retailing. J. Serv. Res. 6 (1), 92–105.
Holoien, D.S., Fiske, S.T., 2013. Downplaying positive impressions: compensation between warmth and competence in impression management. J. Exp. Soc. Psychol. 49 (1), 33–41.
Howe, H.S., Zhou, L., Dias, R.S., Fitzsimons, G.J., 2023. Aha over Haha: brands benefit more from being clever than from being funny. J. Consum. Psychol. 33 (1), 107–114.

Hoyer, W.D., Kroschke, M., Schmitt, B., Kraume, K., Shankar, V., 2020. Transforming the customer experience through new technologies. J. Interact. Market. 51, 57–71.
Hsu, C.-L., Lin, J.C.-C., 2023. Understanding the user satisfaction and loyalty of customer service chatbots. J. Retailing Consum. Serv. 71, 103211.
Hu, K., 2023. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
Huang, D., Markovitch, D.G., Stough, R.A., 2024. Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust. J. Retailing Consum. Serv. 76, 103600.
Huang, M.-H., Rust, R.T., 2022. A framework for collaborative artificial intelligence in marketing. J. Retailing 98 (2), 209–223.
Huang, R., Ha, S., 2020. The effects of warmth-oriented and competence-oriented service recovery messages on observers on online platforms. J. Bus. Res. 121, 616–627.
Huang, Y., Gursoy, D., Zhang, M., Nunkoo, R., Shi, S., 2021. Interactivity in online chat: conversational cues and visual cues in the service recovery process. Int. J. Inf. Manag. 60, 102360.
Huang, Y.-S., Dootson, P., 2022. Chatbots and service failure: when does it lead to customer aggression. J. Retailing Consum. Serv. 68, 103044.
Hulland, J., Baumgartner, H., Smith, K.M., 2018. Marketing survey research best practices: evidence and recommendations from a review of JAMS articles. JAMS 46 (1), 92–108.
Janssen, A., Grützner, L., Breitner, M., 2021. Why do chatbots fail? A critical success factors analysis. In: Proceedings of the 42nd International Conference on Information Systems (ICIS 2021), Austin, Texas.
Jacquet, B., Hullin, A., Baratgin, J., Jamet, F., 2019. The impact of the Gricean maxims of quality, quantity and manner in chatbots. In: 2019 International Conference on Information and Digital Technologies (Idt). IEEE, pp. 180–189.
Jacquet, B., Baratgin, J., Jamet, F., 2018. The Gricean maxims of quantity and of relation in the Turing test. In: 2018 11th International Conference on Human System Interaction (Hsi). IEEE, pp. 332–338.
Jenneboer, L., Herrando, C., Constantinides, E., 2022. The impact of chatbots on customer loyalty: a systematic literature review. J. Theor. Appl. Electron. Commer. Res. 17, 212–229.
Kaczorowska-Spychalska, D., 2019. How chatbots influence marketing. Management 23 (1), 251–270.
Katana, 2024. 1 in 2 Customers Prefer a Real Human over an AI Chatbot when Chatting Online. https://katanamrp.com/blog/customers-prefer-a-real-human-over-an-ai-chatbot/.
Kelley, S.W., Davis, M.A., 1994. Antecedents to customer expectations for service recovery. JAMS 22 (1), 52–61.
Kelley, S.W., Hoffman, K., Davis, M.A., 1993. A typology of retail failures and recoveries. J. Retailing 69 (4), 429–452.
Kervyn, N., Fiske, S.T., Malone, C., 2022. Social perception of brands: warmth and competence define images of both brands and social groups. Consum. Psychol. Rev. 5 (1), 51–68.
Kervyn, N., Yzerbyt, V.Y., Judd, C.M., Nunes, A., 2009. A question of compensation: the social life of the fundamental dimensions of social perception. J. Pers. Soc. Psychol. 96 (4), 828–842.
Kiki erklärt, K.I., 2020. Was ist ein Chatbot? YouTube. https://www.youtube.com/watch?v=BAqfk1Kai4E&list=PPSV. (Accessed 7 May 2023).
Kim, S.Y., Schmitt, B.H., Thalmann, N.M., 2019. Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Mark. Lett. 30 (1), 1–12.
Klarmobil, 2023. Kontaktmöglichkeiten. klarmobil. https://www.klarmobil.de/service/kontakt/. (Accessed 5 May 2023).
Kobel, S., Groeppel-Klein, A., 2021. No laughing matter, or a secret weapon? Exploring the effect of humor in service failure situations. J. Bus. Res. 132, 260–269.
Korczynski, M., 2005. Skills in service work: an overview. Hum. Resour. Manag. J. 15 (2), 3–14.
Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., André, E., 2014. Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. Int. J. Soc. Robot. 6, 417–427.
Kull, A.J., Romero, M., Monahan, L., 2021. How may I help you? Driving brand engagement through the warmth of an initial chatbot message. J. Bus. Res. 135, 840–850.
Laban, G., 2021. Perceptions of anthropomorphism in a chatbot dialogue: the role of animacy and intelligence. In: Proceedings of the 9th International Conference on Human-Agent Interaction, New York, NY, USA. ACM, New York, NY, USA, pp. 305–310.
Lacey, R., Close, A.G., Finney, R.Z., 2010. The pivotal roles of product knowledge and corporate social responsibility in event sponsorship effectiveness. J. Bus. Res. 3 (11), 1222–1228.
Li, Z., 2023. The dark side of chatgpt: legal and ethical challenges from stochastic parrots and hallucination. arXiv preprint arXiv:2304.14347.
Liang, S., Li, R., Lan, B., Chu, Y., Zhang, M., Li, L., 2024. Untouchable them: the effect of chatbot gender on angry customers. JRIM 18 (6), 1099–1135.
Lim, E.A.C., Lee, Y.H., Foo, M.-D., 2017. Frontline employees' nonverbal cues in service encounters: a double-edged sword. JAMS 45 (5), 657–676.
Liu, D., Lv, Y., Huang, W., 2023. How do consumers react to chatbots' humorous emojis in service failures. Technol. Soc. 73, 102244.
Liu, X.S., Yi, X.S., Wan, L.C., 2022. Friendly or competent? The effects of perception of robot appearance and service context on usage intention. Ann. Tourism Res. 92, 103324.
Lu, Z., Min, Q., Jiang, L., Chen, Q., 2024. The effect of the anthropomorphic design of chatbots on customer switching intention when the chatbot service fails: an expectation perspective. Int. J. Inf. Manag. 76, 102767.
Lundberg, J.K., Sheehan, E.P., 1994. The effects of glasses and weight on perceptions of attractiveness and intelligence. J. Soc. Behav. Pers. 9 (4), 753–760.
Luo, X., Tong, S., Fang, Z., Qu, Z., 2019. Frontiers: machines vs. humans: the impact of artificial intelligence chatbot disclosure on customer purchases. Mark. Sci. 38 (6), 937–947.
Ma, X., Huo, Y., 2024. Drawing a satisfying picture: an exploratory study of human-AI interaction in AI Painting through breakdown–repair communication strategies. Inf. Process. Manag. 61 (4), 103755.
Maar, D., Besson, E., Kefi, H., 2023. Fostering positive customer attitudes and usage intentions for scheduling services via chatbots. JOSM 34 (2), 208–230.
Magnini, V.P., Ford, J.B., Markowski, E.P., Honeycutt, E.D., 2007. The service recovery paradox: justifiable theory or smoldering myth? JSM 21 (3), 213–225.
Markovitch, D.G., Stough, R.A., Huang, D., 2024. Consumer reactions to chatbot versus human service: an investigation in the role of outcome valence and perceived empathy. J. Retailing Consum. Serv. 79, 103847.
Matos, C.A. de, Henrique, J.L., Alberto Vargas Rossi, C., 2007. Service recovery paradox: a meta-analysis. J. Serv. Res. 10 (1), 60–77.
Mattila, A.S., Cho, W., Ro, H., 2011. The role of self-service technologies in restoring justice. J. Bus. Res. 64 (4), 348–355.
Mattila, A.S., Enz, C.A., 2002. The role of emotions in service encounters. J. Serv. Res. 4 (4), 268–277.
Mccoll-Kennedy, J.R., Daus, C.S., Sparks, B.A., 2003. The role of gender in reactions to service failure and recovery. J. Serv. Res. 6 (1), 66–82.
Mccoll-Kennedy, J.R., Sparks, B.A., 2003. Application of fairness theory to service failures and service recovery. J. Serv. Res. 5 (3), 251–266.
McCollough, M.A., Berry, L.L., Yadav, M.S., 2000. An empirical investigation of customer satisfaction after service failure and recovery. J. Serv. Res. 3 (2), 121–137.
McDonnell, M., Baxter, D., 2019. Chatbots and gender stereotyping. Interact. Comput. 31 (2), 116–121.
Merkle, M., 2019. Customer responses to service robots – comparing human-robot interaction with human-human interaction. In: Proceedings of the Annual 52nd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences, pp. 1396–1405.
Michel, S., Bowen, D., Johnston, R., 2009. Why service recovery fails. JOSM 20 (3), 253–273.
Miller, J.L., Craighead, C.W., Karwan, K.R., 2000. Service recovery: a framework and empirical investigation. J. Oper. Manag. 18 (4), 387–400.
Mooshammer, S., Etzrodt, K., 2022. Social research with gender-neutral voices in chatbots – the generation and evaluation of artificial gender-neutral voices with Praat and Google WaveNet. In: International Workshop on Chatbot Research and Design. Springer, Cham, pp. 176–191.
Mordor Intelligence, 2024. Global Chatbot Market Size & Share Analysis – Growth Trends & Forecasts, pp. 2024–2029. https://www.mordorintelligence.com/industry-reports/global-chatbot-market.
Morgeson, F.V., Hult, G.T.M., Mithas, S., Keiningham, T., Fornell, C., 2020. Turning complaining customers into loyal customers: moderators of the complaint handling–customer loyalty relationship. J. Market. 84 (5), 79–99.
Mori, M., MacDorman, K., Kageki, N., 2012. The uncanny valley [from the field]. IEEE Robot. Autom. Mag. 19 (2), 98–100.
Mou, Y., Xu, K., Xia, K., 2019. Unpacking the black box: Examining the (de)Gender categorization effect in human-machine communication. Comput. Hum. Behav. 90, 380–387.
Mozafari, N., Schwede, M., Hammerschmidt, M., Weiger, W., 2022. Claim success, but blame the bot? User reactions to service failure and recovery in interactions with humanoid service robots. In: Proceedings of the 55th Hawaii International Conference on System Sciences, pp. 4296–4305.
Murtaza, Z., Sharma, I., Carbonell, P., 2024. Examining chatbot usage intention in a service encounter: role of task complexity, communication style, and brand personality. Technol. Forecast. Soc. Change 209, 123806.
Nomura, T., Kanda, T., Suzuki, T., 2006. Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. AI Soc. 20 (2), 138–150.
Oliver, R.L., 1977. Effect of expectation and disconfirmation on postexposure product evaluations: an alternative interpretation. J. Appl. Psychol. 62 (4), 480–486.
Open AI, 2023. Optimizing Language Models for Dialogue. https://web.archive.org/web/20221130180912/. (Accessed 28 June 2023). https://openai.com/blog/chatgpt/ (Zugriff28.06.2023).
Parasuraman, A., 2000. Technology Readiness Index (TRI): a multiple-item scale to measure readiness to embrace new technologies. J. Serv. Res. 2 (4), 307–320.
Pavone, G., Meyer-Waarden, L., Munzel, A., 2023. Rage against the machine: experimental insights into customers' negative emotional responses, attributions of responsibility, and coping strategies in artificial intelligence–based service failures. J. Interact. Market. 58 (1), 52–71.
Pizzi, G., Vannucci, V., Mazzoli, V., Donvito, R., 2023. I, Chatbot! The impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions: an abstract. In: Jochims, B., Allen, J. (Eds.), Optimistic Marketing in Challenging Times: Serving Ever-Shifting Customer Needs. Springer Nature Switzerland, Cham, pp. 283–284.
Quach, S., Jebarajakirthy, C., Thaichon, P., 2017. Aesthetic labor and visible diversity: the role in retailing service encounters. J. Retailing Consum. Serv. 38, 34–43.
Rajaobelina, L., Prom Tep, S., Arcand, M., Ricard, L., 2021. Creepiness: its antecedents and impact on loyalty when interacting with a chatbot. Psychol. Market. 38 (12), 2339–2356.

Reinkemeier, F., Gnewuch, U., 2022a. Match or mismatch? How matching personality and gender between voice assistants and users affects trust in voice commerce. In: Proceedings of the 55th Hawaii International Conference on System Sciences 2021, pp. 4326–4335.
Reinkemeier, F., Gnewuch, U., 2022b. Designing effective conversational repair strategies for chatbots. In: Proceedings of the 30th European Conference on Information Systems (ECIS 2022), Timişoara, Romania.
Roesler, E., Naendrup-Poell, L., Manzey, D., Onnasch, L., 2022. Why context matters: the influence of application domain on preferred degree of anthropomorphism and gender attribution in human–robot interaction. Int. J. Soc. Robot. 14 (5), 1155–1166.
Roy, R., Naidoo, V., 2021. Enhancing chatbot effectiveness: the role of anthropomorphic conversational styles and time orientation. J. Bus. Res. 126, 23–34.
Rozumowski, A., Haupt, M., 2021. Sorry I am still learning – active expectation management of chatbots. In: Proceedings of the 19th International Conference E-Society (ES), Virtual, 3-5 March 2021. IADIS, pp. 275–278.
Rudman, L.A., Goodwin, S.A., 2004. Gender differences in automatic in-group bias: why do women like women more than men like men? J. Pers. Soc. Psychol. 87 (4), 494–509.
Sands, S., Campbell, C., Plangger, K., Pitt, L., 2022. Buffer bots: the role of virtual service agents in mitigating negative effects when service fails. Psychol. Market. 39 (11), 2039–2054.
Scholarios, D., Taylor, P., 2010. Gender, choice and constraint in call centre employment. New Technology. Work Employ 25 (2), 101–116.
Schuetzler, R.M., Grimes, G.M., Scott Giboney, J., 2020. The impact of chatbot conversational skill on engagement and perceived humanness. J. Manag. Inf. Syst. 37 (3), 875–900.
Seeger, A.-M., Heinzl, A., 2021. Chatbots often fail! Can anthropomorphic design mitigate trust loss in conversational agents for customer service?. In: Proceedings of the 29th European Conference on Information Systems (ECIS 2021), Marrakesh, Morocco - A Virtual AIS Conference.
Seeger, A.-M., Pfeiffer, J., Heinzl, A., 2018. Designing anthropomorphic conversational agents: development and empirical evaluation of a design framework. In: Proceedings of the 39th International Conference on Information Systems (ICIS 2018), San Francisco, USA.
Seiler, R., Schär, A., 2021. Chatbots, conversational interfaces, and the stereotype content model. In: Proceedings of the 54th Hawaii International Conference on System Sciences 2021, pp. 1861–1867.
Sella, F., Raz, G., Cohen Kadosh, R., 2021. When randomisation is not good enough: matching groups in intervention studies. Psychon. Bull. Rev. 28 (6), 2085–2093.
Shaier, S., Hunter, L.E., Wense, K., 2023. Who Are All the Stochastic Parrots Imitating? They Should Tell Us! arXiv preprint. https://arxiv.org/abs/2310.10583.
Shams, G., Kim, K.K., Kim, K., 2024. Enhancing service recovery satisfaction with chatbots: the role of humor and informal language. Int. J. Hospit. Manag. 120.
Sheehan, B., Jin, H.S., Gottlieb, U., 2020. Customer service chatbots: anthropomorphism and adoption. J. Bus. Res. 115, 14–24.
Shin, H., Bunosso, I., Levine, L.R., 2023. The influence of chatbot humour on consumer evaluations of services. Int. J. Consum. Stud. 47 (2), 545–562.
Smith, A.K., Bolton, R.N., Wagner, J., 1999. A model of customer satisfaction with service encounters involving failure and recovery. J. Mark. Res. 36 (3), 356–372.
Smith, N.A., Martinez, L.R., Sabat, I.E., 2016. Weight and gender in service jobs. Cornell Hosp. Q. 57 (3), 314–328.
Snipes, R.L., Thomson, N.F., Oswald, S.L., 2006. Gender bias in customer evaluations of service quality: an empirical investigation. JSM 20 (4), 274–284.
Söderlund, M., 2021. The robot-to-robot service encounter: an examination of the impact of inter-robot warmth. JSM 35 (9), 15–27.
Söderlund, M., Oikarinen, E.-L., 2021. Service encounters with virtual agents: an examination of perceived humanness as a source of customer satisfaction. EJM 55 (13), 94–121.
Song, M., Du, J., Xing, X., Mou, J., 2022. Should the chatbot "save itself" or "be helped by others"? The influence of service recovery types on consumer perceptions of recovery satisfaction. Electron. Commer. Res. Appl. 55, 101199.
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. J. Retailing Consum. Serv. 73, 103323.
Statista, 2024. The proportion of prepaid and contract customers among mobile phone users in Germany by age group in 2020 (Anteile von Prepaid- und Vertragskunden an den Mobiltelefonnutzern in Deutschland nach Altersgruppe im Jahr 2020). https://de.statista.com/statistik/daten/studie/154166/umfrage/vertrags-und-prepaid-mobilfunkkunden-nach-alter-in-deutschland/.
Statista, 2018. Can you think of communicating with a "chatbot" in general? (Können Sie sich ganz allgemein vorstellen, mit einem „Chatbot" zu kommunizieren?). https://de.statista.com/statistik/daten/studie/872937/umfrage/bereitschaft-zur-kommunikation-mit-chatbots-nach-altersgruppen-in-deutschland/.
Stroessner, S.J., Benitez, J., 2019. The social perception of humanoid and non-humanoid robots: effects of gendered and machinelike features. Int. J. Soc. Robot. 11 (2), 305–315.
Sundaram, D.S., Webster, C., 2000. The role of nonverbal communication in service encounters. JSM 14 (5), 378–391.
Suta, P., Lan, X., Wu, B., Mongkolnam, P., Chan, J.H., 2020. An overview of machine learning in chatbots. IJMERR 9 (4), 502–510.
Thaler, M., Schlögl, S., Groth, A., 2021. Agent vs. avatar: comparing embodied conversational agents concerning characteristics of the uncanny valley. In: 2020 IEEE International Conference on Human-Machine Systems (ICHMS). IEEE, pp. 1–6.
Toader, D.-C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., Rădulescu, A.T., 2020. The effect of social presence and chatbot errors on trust. Sustainability 12 (1), 256.
Tran, A.D., Pallant, J.I., Johnson, L.W., 2021. Exploring the impact of chatbots on consumer sentiment and expectations in retail. J. Retailing Consum. Serv. 63, 102718.
Valentini, S., Orsingher, C., Polyakova, A., 2020. Customers' emotions in service failure and recovery: a meta-analysis. Mark. Letture 31 (2/3), 199–216.
van Doorn, J., Mende, M., Noble, S.M., Hulland, J., Ostrom, A.L., Grewal, D., Petersen, J. A., 2017. Domo arigato Mr. Roboto: emergence of automated social presence in organizational frontlines and customers' service experiences. J. Serv. Res. 20 (1), 43–58.
van Vaerenbergh, Y., Orsingher, C., Vermeir, I., Larivière, B., 2014. A meta-analysis of relationships linking service failure attributions to customer outcomes. J. Serv. Res. 17 (4), 381–398.
Wang, Z., Mao, H., Jessica Li, Y., Liu, F., 2016. Smile big or not? Effects of smile intensity on perceptions of warmth and competence. J. Consum. Res. ucw062. 43 (5), 787–805.
Wei, H., Ran, Y., 2019. Male versus female: how the gender of apologizers influences consumer forgiveness. J. Bus. Ethics 154 (2), 371–387.
Wirtz, J., Mattila, A.S., 2004. Consumer responses to compensation, speed of recovery and apology after a service failure. Int. J. Serv. Ind. Manag. 15 (2), 150–166.
Wirtz, J., Mccoll-Kennedy, J.R., 2010. Opportunistic customer claiming during service recovery. JAMS 38 (5), 654–675.
Wu, L., Fan, A., Mattila, A.S., 2015. Wearable technology in service delivery processes: the gender-moderated technology objectification effect. Int. J. Hospit. Manag. 51, 1–7.
Xie, Y., Liang, C., Zhou, P., Jiang, L., 2024. Exploring the influence mechanism of chatbot-expressed humor on service satisfaction in online customer service. J. Retailing Consum. Serv. 76, 103599.
Xing, X., Song, M., Duan, Y., Mou, J., 2022. Effects of different service failure types and recovery strategies on the consumer response mechanism of chatbots. Technol. Soc. 70, 102049.
Yang, L.W., Aggarwal, P., McGill, A.L., 2020. The 3 C's of anthropomorphism: connection, comprehension, and competition. Consum. Psychol. Rev. 3 (1), 3–19.
Yang, Z., Zhou, J., Yang, H., 2023. The impact of AI's response method on service recovery satisfaction in the context of service failure. Sustainability 15 (4), 3294.
Yip, J.A., Martin, R.A., 2006. Sense of humor, emotional intelligence, and social competence. J. Res. Pers. 40 (6), 1202–1208.
Zhang, J., Zhu, Y., Wu, J., Yu-Buck, G.F., 2023. A natural apology is sincere: understanding chatbots' performance in symbolic recovery. Int. J. Hospit. Manag. 108, 103387.
Zhang, T., Feng, C., Chen, H., Xian, J., 2022. Calming the customers by AI: investigating the role of chatbot acting-cute strategies in soothing negative customer emotions. Electron. Mark. 32 (4), 2277–2292.
Zhao, X., Lynch, J.G., Chen, Q., 2010. Reconsidering baron and kenny: myths and truths about mediation analysis. J. Consum. Res. 37 (2), 197–206.
Zheng, T., Duan, X., Zhang, K., Yang, X., Jiang, Y., 2023. How chatbots' anthropomorphism affects user satisfaction: the mediating role of perceived warmth and competence. In: Tu, Y., Chi, M. (Eds.), E-business. Digital Empowerment for an Intelligent Future, vol. 481. Springer Nature, Switzerland, Cham, pp. 96–107.
Zhou, C., Chang, Q., 2024. Informational or emotional? Exploring the relative effects of chatbots' self-recovery strategies on consumer satisfaction. J. Retailing Consum. Serv. 78, 103779.
Zhu, Y., Zhang, J., Wu, J., 2023. Who did what and when? The effect of chatbots' service recovery on customer satisfaction and revisit intention. JHTT 14 (3), 416–429.
Zogaj, A., Mähner, P.M., Tscheulin, D.K., 2021. Similarity between human beings and chatbots – the effect of self-congruence on consumer satisfaction while considering the mediating role of authenticity. In: Bruhn, M., Hadwich, K. (Eds.), Künstliche Intelligenz im Dienstleistungsmanagement. Springer, Fachmedien Wiesbaden, Wiesbaden, pp. 427–443.
Zogaj, A., Mähner, P.M., Yang, L., Tscheulin, D.K., 2023. It's a Match! The effects of chatbot anthropomorphization and chatbot gender on consumer behavior. J. Bus. Res. 155, 113412.

Alexandra Rese (alexandra.rese@uni-bayreuth.de) is Associate Professor for Innovation Marketing at the University of Bayreuth, Germany. She completed her dissertation in sociology and entrepreneurship at the University of Karlsruhe and her habilitation in business administration at Brandenburg University of Technology Cottbus-Senftenberg. Her works have appeared in journals such as *R&D Management, Creativity and Innovation Management, International Journal of Innovation Management, Review of Managerial Science, Journal of Knowledge Management, Technological Forecasting and Social Change, Research Policy, Computers in Human Behavior, Journal of Retailing and Consumer Services* and *Journal of Marketing Management*. Her current research focuses on the acceptance of innovative applications in retailing, e.g. augmented reality or chatbots, as well as sustainability in retailing.

Lennart Witthohn has received a Master of Science degree in Business Administration from the University of Bayreuth, Germany. His main research interest are the satisfaction of consumers with a chatbot after a service failure has occurred and effects of service recovery and anthropomorphic design elements in this regard.