6-18-2022

Designing Effective Conversational Repair Strategies for Chatbots

Fabian Reinkemeier
University of Goettingen, fabian.reinkemeier@wiwi.uni-goettingen.de

Ulrich Gnewuch
Karlsruhe Institute of Technology (KIT), ulrich.gnewuch@kit.edu

DESIGNING EFFECTIVE CONVERSATIONAL REPAIR STRATEGIES FOR CHATBOTS

Research Paper

Fabian Reinkemeier, University of Goettingen, Goettingen, Germany, fabian.reinkemeier@wiwi.uni-goettingen.de

Ulrich Gnewuch, Karlsruhe Institute of Technology (KIT), Institute of Information Systems and Marketing (IISM), Karlsruhe, Germany, ulrich.gnewuch@kit.edu

Abstract

Conversational breakdowns often force users to go through frustrating loops of trial and error when trying to get answers from chatbots. Although research has emphasized the potential of conversational repair strategies in helping users resolve breakdowns, design knowledge for implementing such strategies is scarce. To address this challenge, we are conducting a design science research (DSR) project to design effective repair strategies that help users recover from conversational breakdowns with chatbots. This paper presents the first design cycle, proposing, instantiating, and evaluating our first design principle on identifying the cause of conversational breakdowns. Using 21,736 real-world user messages from a large insurance company, we conducted a cluster analysis of 5,668 messages leading to breakdowns, identified four distinct breakdown types, and built a classifier that can be used to automatically identify breakdown causes in real time. Our research contributes with prescriptive knowledge for designing repair strategies in conversational breakdown situations.

Keywords: Conversational Breakdown, Repair Strategies, Chatbot, Design Science Research.

1 Introduction

Many companies use artificial intelligence (AI)–based chatbots to respond to customer service requests, provide personalized product information, and support customers' purchase decisions in e-commerce (Adam et al., 2021; Følstad and Brandtzæg, 2017; Go and Sundar, 2019). As customer demand for round-the-clock support grows (Klopfenstein et al., 2017), businesses are increasing their use of chatbots in customer interactions (Gartner, 2020); the market size of chatbots worldwide is expected to expand to $142 billion by 2024 from just $2.8 billion in 2019 (Insider Intelligence, 2021).

Despite ongoing advances in AI and natural language processing (NLP), chatbots often fail to understand user input and provide an appropriate response, resulting in negative and frustrating experiences for users (Følstad et al., 2018; Klopfenstein et al., 2017; Takayama et al., 2019). These conversational breakdowns in human–chatbot interactions—defined as "failures of the system to correctly understand the intended meaning of the user's communication" (Li et al., 2020, p. 1)—have been identified as a critical factor in how users perceive a chatbot and the company offering it (Diederich et al., 2021; Seeger and Heinzl, 2021). If a chatbot fails to provide an answer, users frequently abandon the conversation (Akhtar et al., 2019) and may even turn away from these services entirely (Ashktorab et al., 2019; van der Goot et al., 2021). Consequently, researchers consider these breakdowns as a key challenge in chatbot development (Følstad et al., 2018; Janssen et al., 2021).

Conversational breakdowns occur because chatbots lack a full understanding of natural human language. Achieving such an understanding remains a formidable task because of the complexity and variety of possible user queries; users will inevitably say things to a chatbot that it cannot understand,

even with advanced chatbot design and training (Rasa, 2021; van der Goot et al., 2021). Furthermore, from a business perspective, it is often difficult and impractical to build in a response to every potential user query because of the time and effort involved in training a chatbot. Chatbots often have a limited number of responses in the early stages of their implementation, which is practical for collecting real user questions and scaling processes (van der Goot et al., 2021).

Given that conversational breakdowns in human–chatbot interactions are unavoidable, research has emphasized the importance of *conversational repair strategies* to mitigate the negative effects of breakdowns (Ashktorab et al., 2019; Benner et al., 2021; Takayama et al., 2019). Some repair strategies aim to solve conversational breakdowns outside the ongoing chatbot interaction (e.g., by handing over the user to a live chat employee) (Go and Sundar, 2019). However, such strategies lead to additional costs for companies and run counter to the purpose of implementing a chatbot to automatically handle customer inquiries. In contrast, other repair strategies offer users the opportunity to resolve breakdowns themselves during interactions (Moore and Arar, 2018). However, current repair messages are often rather generic: "I'm sorry, I didn't understand you. Can you please try again?" Such messages have been found to even increase frustration because they do not help users understand why the chatbot failed or how to rephrase their questions to get appropriate answers (van der Goot et al., 2021). As a result, users must engage in a "haphazard trial-and-error process to recover from the breakdown" (Ashktorab et al., 2019, p. 1).

Existing research emphasizes that in order to realize the full potential of chatbots, breakdowns in chatbot interactions must be repaired (Moore and Arar, 2018; Takayama et al., 2019). For example, in addition to indicating that it did not understand the user, a chatbot could also initiate a repair of the breakdown by asking the user to replace an unrecognized word with a synonym (Li et al., 2020). However, scholarly research has tended to focus on proposing the existence of various repair strategies (Ashktorab et al., 2019; Benner et al., 2021) rather than developing knowledge on how to actually design such strategies for chatbots (Benner et al., 2021; Diederich et al., 2021; Janssen et al., 2021). To address this research gap, we pose the following research question: *How to design effective conversational repair strategies for chatbots to help users recover from conversational breakdowns in human–chatbot interactions?*

To address this question, we followed the design science research (DSR) approach (Hevner et al., 2004). Drawing on communication theories on conversational repair as our kernel theories, we iteratively design, implement, and evaluate conversational repair strategies for chatbot breakdown situations. In our ongoing DSR project, we are collaborating with a large international insurance company that offers multiple chatbots for its customers.

In this paper, we present the results of our first design cycle, which focused on addressing the following subquestion: *What are the different breakdown types and how can they be identified from user messages in human–chatbot interactions?* First, we reviewed the existing literature on conversational breakdowns and their potential effects, as well as ways to address them. We then derived two design principles to guide the design of our effective conversational repair strategies. To instantiate our first design principle, we conducted a cluster analysis of 5,668 breakdown messages from a dataset of over 21,000 interactions with the chatbots of our collaboration partner. Identifying four distinct types of conversational breakdowns, we created a breakdown type classifier and evaluated its classifications with two human coders. Our DSR project contributes both descriptive (i.e., types of breakdowns and their characteristics) and prescriptive design knowledge (i.e., how to identify and classify breakdown types from user messages) that can help researchers and practitioners design more effective conversational repair strategies for human–chatbot interactions.

2 Theoretical Foundations and Related Work

2.1 Chatbots

With the recent improvements in AI, chatbots have been gaining visibility and can be found in a wide range of channels, including corporate websites and social media platforms (e.g., Facebook). In this study, we use the term *chatbots* to refer to software applications that communicate with their human users in written conversations using natural language (Dale, 2016). Chatbots belong to the more general class of conversational agents (CAs) (Feine et al., 2019; McTear et al., 2016), which are categorized as either text-based (e.g., chatbots on websites) or voice-based (e.g., Amazon's Alexa). These CAs represent a new class in information systems (IS) that is distinctly different from other IS due to their greater interactivity and intelligent capabilities (Maedche et al., 2019; Zierau et al., 2020). Instead of communicating in an unnatural form using only graphical elements in online interactions, users can also interact with AI-based chatbots through their natural language and thus communicate in a more intuitive and interactive way, just as they do with other humans (McTear, 2020). From the user perspective, chatbots offer a valuable and convenient service that is available instantly and 24/7 for a variety of purposes (van der Goot et al., 2021); from a business perspective, chatbots provide a round-the-clock service that is less costly than other contact channels (Skjuve and Brandtzæg, 2019) and can scale quickly to handle a higher volume of customer service interactions (Luo et al., 2019).

To take advantage of these opportunities, there are now numerous technologies on the market that allow companies to develop such chatbots (e.g., IBM Watson or Google Dialogflow), ranging from scripted or rule-based bots to more advanced AI-based bots that learn from previous conversations (Tuzovic and Paluch, 2018). To understand the meaning of a user query, the underlying intention (referred to as intent) is analyzed by the system recognition algorithm (e.g., using Artificial Intelligence Markup Language (AIML), symbolic AI, or machine learning approaches). Then, the machine searches for a match in the knowledge base in which chatbot developers have programmed different intents and utterances (different formulations of expected user queries that should trigger a particular intent) (McTear, 2020).

Although a large number of chatbots have been implemented in recent years, many have not met expectations and have disappeared due to technical or design flaws (Gnewuch et al., 2017; Janssen et al., 2021). Communicating with chatbots is unfamiliar to some people (Luo et al., 2019) and chatbots sometimes fail to understand the simplest user queries, such as when the user makes multiple typos or adds a small word like "not" (Mitrevski, 2018).

Despite these drawbacks, practitioners use chatbots in a variety of fields (McTear, 2020). In research, the number of publications is growing tremendously, and many papers address chatbots in terms of attitudinal and perceptual outcomes as well as design elements (Zierau et al., 2020), which is still a challenge in practice (Følstad and Brandtzæg, 2017; Gnewuch et al., 2017). For example, recent research has highlighted the problem of users experiencing conversational breakdowns and getting stuck in error loops when conversing with chatbots, leading to negative experiences (Ashktorab et al., 2019; Benner et al., 2021; Diederich et al., 2021; van der Goot et al., 2021).

2.2 Conversational breakdown and repair in human–chatbot interactions

Conversational breakdowns are a critical challenge in human–chatbot interactions (Janssen et al., 2021; Takayama et al., 2019). These breakdowns occur when a chatbot is unable to understand a user's input (i.e., recognize the intent) and provide an answer. Due to the high relevance of the topic, conversational repair strategies are gaining importance (Benner et al., 2021). Repair strategies refer to those that support "recovering from the breakdown to accomplish the task goal" (Ashktorab et al., 2019, p. 7), such as having the user clarifying the intent. In general, successful repair mechanisms enable people to maintain communication and mutual understanding despite the inevitable ambiguities and errors present in natural conversations (Albert and Ruiter, 2018; Moore and Arar, 2018). Overall,

this can minimize the negative consequences of breakdowns, including those related to user satisfaction, adoption, and experience (Lee et al., 2010; Sheehan et al., 2020; Takayama et al., 2019).

In general, there are different strategies for conversational repair that are initiated by either the chatbot or the user (McTear et al., 2016). In this context, Ashktorab et al. (2019) investigated existing conversational repair strategies for chatbots and developed new ones, finding seven strategies for situations with evidence of breakdowns: defer, options, repeat, confirmation, out-of-vocabulary explanation, keyword confirmation explanation, and keyword highlight explanation. Go and Sundar (2019) also suggested that the option of deferral—handing over the failed query to a human agent—is a good option to mitigate users' negative experiences. However, Sohn et al. (2021) noted that implementing a human-based live chat on websites reduces the likelihood that customers will disclose information on their own; if anything, the company should refer to the chat as an AI-based live chat. Another strategy, similar to out-of-vocabulary explanation, is to ask the user to make semantic adjustments, such as rephrasing the sentence or replacing a word with a synonym (Li et al., 2020). In addition, Skjuve et al. (2021) found that some users liked the openness and honesty when the chatbot responded transparently in a breakdown situation. This result is in line with Klopfenstein et al. (2017), who pointed out that it is particularly frustrating for users when a chatbot hides its errors. Li et al. (2020) pointed to the strategy of code switching, which consists of adapting the speaking style to the listener in order to reduce breakdowns. Unfortunately, previous studies on new repair strategies have not been tested in field experiments (Ashktorab et al., 2019) and lack precise guidance on how these strategies should be implemented (e.g., which strategy is effective for which user message), indicating that more specific design-oriented implications are needed (Benner et al., 2021; Diederich et al., 2021).

2.3 Communication theories on conversational repair

In human–human communication, people must constantly detect and resolve problems of speaking, hearing, and understanding (Albert and Ruiter, 2018; Ashktorab et al., 2019). This process is called conversational repair. While repair strategies can vary across languages, the principles of conversational repair in human conversations are very similar (Dingemanse et al., 2015). Communication theories generally distinguish repair strategies along a basic four-way taxonomy (Albert and Ruiter, 2018). This taxonomy has two dimensions that classify which party—self or other—initiates a repair and which party resolves it. Self is the producer/speaker of a trouble source (e.g., a misunderstanding) and other is the recipient of the trouble source (i.e., the addressee). Accordingly, there are four types of conversational repair: self-initiated self-repair, self-initiated other-repair, other-initiated self-repair, and other-initiated other-repair (Albert and Ruiter, 2018). An example of self-initiated self-repair would be if a person corrects themselves while speaking (e.g., "Yesterday, I had a nice talk with Marc … I mean Stefan."), while other-initiated self-repair is for example when the other person asks the speaker to repeat some information (e.g., "Who did you talk to?").

Generally, communication researchers have found that types of repair differ in terms of their sequential properties and preference. Opportunities for self-repairs (repairs by the current speaker) occur sequentially before opportunities for other-repair (repairs by the addressee), and self-repairs are preferred over other-repairs (McTear et al., 2016). Moreover, if one method of repair fails, people "upgrade" to incrementally more powerful methods until the repair is complete (Albert and Ruiter, 2018). A fundamental principle in human–human communication in general, and in conversational repair in particular, is the principle of least collaborative effort (Clark and Schaefer, 1987). This principle states that participants in a conversation try to minimize the total effort spent in that interactional encounter. For example, instead of choosing the simplest repair strategy possible (e.g., "huh?") and letting the other participant repeat their entire message, humans try to minimize collaborative effort by selecting the strongest repair initiator possible (e.g., "Could you please repeat your last sentence? I didn't get it.").

3 Design Science Research Project

Our research project follows the DSR approach (Hevner et al., 2004) to design effective conversational repair strategies for chatbots in order to help users recover from conversational breakdowns. The DSR approach is particularly suited to our research as it allows us to integrate scientific knowledge from our kernel theories on conversational breakdown and repair in human–human communication (Albert and Ruiter, 2018; Sacks et al., 1974; Schegloff et al., 1977), leverage existing design knowledge on conversational repair strategies (e.g., Ashktorab et al., 2019), and involve experts and real users to iteratively design and evaluate our artifact in a real-world setting. In this project, we collaborate with a large international insurance company (hereafter referred to as InsurCorp for the purpose of anonymity). InsurCorp currently runs three sales and FAQ chatbots for different insurance products on its German websites. We adopted the DSR framework proposed by Kuechler and Vaishnavi (2008) and divided our projects into three iterative build-evaluation cycles to incrementally improve the functionality of our artifact (see Figure 1). In this paper, we report on our approach and the results of the first design cycle, which focused on how to automatically identify conversational breakdown types based on the user message that led to the breakdown.

General DSR Project Phases	**Design Cycle #1:** ***Classifying Breakdown Types***	**Design Cycle #2:** ***Designing Repair Messages***	**Design Cycle #3:** ***Holistic Conv. Repair Strategies***
Awareness of Problem	Literature review and analysis of real-world chatbot interactions	Analysis of initial evaluation and literature review	Reflection and analysis of previous evaluations
Suggestion	Formulation of initial design principles	Refinement of design principles	Refinement of design principles
Development	Cluster analysis and breakdown type classifier	Identification and creation of specific repair messages	Implementation of design in InsurCorp's chatbots
Evaluation	Evaluation with human coders	Online experiment	Field experiment at InsurCorp
Conclusion			Formulation of nascent design theory

Figure 1. Design science research project (following Kuechler and Vaishnavi, 2008).

The first design cycle focused on the first part of the project, namely understanding what causes a conversational breakdown and how to automatically identify different types of breakdowns based on characteristics of users' messages. In the problem-awareness phase, we took three steps to explore the problem space: First, we reviewed existing literature to better understand the consequences of conversational breakdowns and the key issues in current conversational repair strategies used by chatbots. Second, to complement findings from the literature, we analyzed user feedback from a rating option following conversations with InsurCorp's chatbots (N = 78). Third, to explore the state of the art in real-world chatbots, we conducted an exploratory analysis of 42 other chatbots from the insurance and banking industries. Our goal was to identify how existing chatbots based on different chatbot frameworks respond when they are unable to provide answers. We collected their responses at different stages of the conversation (i.e., after the first, second, and third breakdowns) using the same set of user messages (e.g., short/medium/long messages, small talk, off-topic questions) for all chatbots. In the suggestion phase, we drew on communication theories on conversational repair (Albert and Ruiter, 2018) and the principle of least collaborative effort (Clark and Schaefer, 1987) to propose two design principles (DPs) for effective conversational repair strategies for chatbots to help users recover from conversational breakdowns. To instantiate our DPs, we leveraged existing data from InsurCorp (N = 21,736 messages) containing chatbot interactions with conversational breakdowns (N = 5,668). More specifically, we first conducted a cluster analysis of user messages that caused breakdowns. Based on the four clusters identified, we built a classifier for new messages to

assign them to clusters based on the distance between centroids and evaluated the classifier with human coders. As a next step, we plan to label all messages manually and develop a machine learning classifier for the upcoming design cycles.

As depicted in Figure 1, we have planned for two additional design cycles to incrementally add functionality, refine our design, and evaluate it in online and field settings. In the second design cycle, we intend to identify and design specific repair messages for each of the breakdown types (i.e., the clusters identified in the first cycle). These specific repair messages will be evaluated in an online experiment in which users are presented a breakdown with one of the repair messages and are asked to repair the conversation. The third and final design cycle aims to bring together both components from the previous cycles, refine the overall design, and evaluate the entire artifact in a field study at InsurCorp. Our ultimate goal is to develop a nascent design theory (Gregor and Jones, 2007) for effective conversational repair strategies for chatbots.

4 Results

4.1 Awareness of the problem

Our analysis of previous studies that have included interviews with experts (e.g., Janssen et al., 2021) and users (e.g., Følstad et al., 2018; van der Goot et al., 2021) and our investigation of real-world examples highlight that conversational breakdowns are one of the biggest challenges for chatbots in delivering a satisfying experience (Janssen et al., 2021; Takayama et al., 2019). If a chatbot fails to respond adequately to users, these breakdowns can lead to frustration and anger (Klopfenstein et al., 2017; van der Goot et al., 2021). For example, one interviewee in the Følstad et al. (2018, p. 200) study pointed out, "It is not always the chatbot understand what I say. And when I, after formulating my question in three or four different ways and the chatbot still does not understand, then I get annoyed." Moreover, breakdowns can have negative effects, such as weaker relationship development (Skjuve et al., 2021), increased perception of uncanniness (Diederich et al., 2021), and decreased trust in the chatbot (Seeger and Heinzl, 2021). Similarly, the analysis of user feedback from InsurCorp's chatbots revealed that it is particularly problematic when the chatbot does not understand its users; some wished for "maybe a few more keywords, thanks" and suggested "let[ting] the AI learn so it can answer more questions, only that increases adoption!" or "expand[ing] the program to understand questions." All in all, it can be concluded that conversational breakdowns are a serious problem that can lead to people turning away from the service (Akhtar et al., 2019; Ashktorab et al., 2019).

However, it may be difficult—if not impossible—to completely avoid conversational breakdowns, as there are inherent complexities of natural language, limitations of current technology, and other constraints in play (e.g., how much time an organization can spend training the chatbot's recognition algorithm). Likewise, chatbot technology providers, such as Rasa, have pointed out that even with perfect design, the chatbot will inevitably fail to understand something the user says, making it important for the chatbot to repair such situations in an elegant way (Rasa, 2021). The literature has also emphasized the importance of conversational repair, particularly because it allows the user to correct the conversation (Benner et al., 2021; Moore and Arar, 2018). However, research and our chatbot benchmarking indicate that chatbots often provide generic repair messages ("I'm sorry, I didn't understand you. Can you express it differently?"), regardless of how users phrase their questions (van der Goot et al., 2021). As a result, users tend to leave the chatbot directly or try to solve the problem through trial and error without really knowing what they are doing (Ashktorab et al., 2019). This can lead to frustration: In the work of van der Goot et al. (2021, p. 197), a user noted, "I already tried to ask the same question twice; now I would have to formulate it a third time; I am not going to do that." Thus, the prevailing design of repair messages in response to a breakdown is not helpful and rather increases user frustration during interaction with a chatbot. However, companies are unable to avoid such annoying breakdown loops, as they lack guidance on how to design effective conversational repair messages in individual human–chatbot interaction.

In summary, ideally all chatbot users would receive effective repair messages based on their individual input and thus experience better interactions (Moore and Arar, 2018). In reality, however, we found that users receive generic messages and experience their repeated attempts to recover from the breakdown as unsuccessful (Ashktorab et al., 2019), leading users to waste their time, get frustrated, and abandon the chatbot. We therefore suggest that chatbot designers and researchers should evaluate and test possible solutions for situations when the conversation breaks down.

4.2 Suggestion

A straightforward way to address the issues identified would be to ensure that conversational breakdowns in human–chatbot interactions do not occur in the first place. However, given the current technical limitations and organizational challenges described above, this solution is not feasible. Therefore, we argue that in the short to medium term, it is important to explore ways to mitigate the negative effects of conversational breakdowns by designing conversational repair strategies that are more effective in helping users overcome such situations. We address this challenge by drawing from research on conversational breakdown and repair in human–human interaction.

Communication theories on conversational repair generally distinguish between the person who initiates a repair and the person who resolves it (Albert and Ruiter, 2018). In our DSR project, we focus on situations in which a chatbot initiates a repair (e.g., by responding with "Sorry, I didn't understand …") and the user performs it (e.g., by reformulating the message that caused the breakdown). Hence, our focus lies on chatbot-initiated user-repair. This helps us specifically address the above-identified problem of repair messages tending to be generic and therefore not helping customers understand why the chatbot has failed and how they can rephrase their questions to get appropriate answers. We propose the following core meta-requirement: *Conversational repair strategies must include specific repair messages that help users recover from breakdowns in their interactions with chatbots.*

To address this key requirement, we draw on communication theories of conversational repair to propose two DPs based on the conceptual schema proposed by Gregor et al. (2020). First, to be able to provide more specific repair messages, it is important to identify what caused the breakdown in communication; without knowing the reason for the breakdown, it would be impossible to offer a repair message that specifically corresponds to that reason. Communication research suggests that in order to effectively repair a breakdown in communication, it is important to understand the type of trouble source (e.g., hearing, speaking, understanding) and then select an appropriate repair strategy (Albert and Ruiter, 2018). For example, when two individuals talking on the phone cannot understand each other because the connection is poor (Clark and Schaefer, 1987), they would choose a different repair strategy than when the breakdown occurs because one person cannot make sense of what the other person just said. In the chatbot context, a common cause for breakdown is submitting a long and complicated message with multiple questions that makes it difficult for the chatbot to detect the user's main intent. Therefore, we argue that as with human–human communication, understanding what caused a breakdown is the fundament of a more effective repair strategy in human–chatbot interaction. Hence, we propose our first DP:

DP1: *For chatbot designers to help users recover from a conversational breakdown, enable the chatbot to identify the breakdown's cause because it can be used as a basis for formulating more specific repair messages.*

Second, it is important to leverage the cause of the conversational breakdown identified in the first step to initiate the repair process. According to the principle of least collaborative effort (Clark and Schaefer, 1987), the person who initiates a repair should use the most specific so-called repair initiator to minimize the effort required for the repair. For example, in human–human communication, the question "Huh?" would be less specific than "Sorry, I didn't hear the last word." Similarly, a chatbot should avoid the rather unspecific "I'm sorry, I didn't understand you. Please try again" and instead offer a more specific repair initiator (e.g., "Please try again with only a few keywords"). In addition, Yuan et al. (2020) suggested that a chatbot could explain the reason for the problem and provide

guidance on how to solve it. Taken together, we argue that an effective repair strategy should consist of repair messages that are tailored to the previously identified breakdown cause and include an explanation and suggestion on how to address it. Hence, we propose our second DP:

DP2*: For chatbot designers to help users recover from a conversational breakdown, enable the chatbot to send repair messages that map to the previously identified cause of the breakdown because a more specific repair message helps users understand why the chatbot has failed and how they can resolve the problem.*

4.3 Development

In our first design cycle, we focused on the instantiation and evaluation of DP1. We adopted a data-driven exploratory approach and leveraged a dataset of chatbot interactions from InsurCorp. In the following, we briefly describe InsurCorp's chatbots, the underlying technology, and relevant variables for our analysis. We then explain the cluster analysis performed to delineate different types of conversational breakdowns based on real-world user questions that the chatbots were unable to answer. Based on the clusters, we developed and evaluated a nearest centroid classifier that enables chatbots to identify the breakdown types in real time from user messages that caused breakdowns. This artifact serves as the foundation for the remainder of our DSR project and may also be used by other chatbots to identify breakdown types and provide more specific repair messages.

4.3.1 InsurCorp's chatbots and their intent recognition technology

InsurCorp offers three different chatbots on its most visited web pages for selected insurance products (e.g., household insurance). These chatbots are designed to answer users' most common questions related to InsurCorp's products and services, with more than 300 answers available. The chatbots were developed iteratively and trained continuously over a period of at least 3 and up to 25 months (depending on the chatbot) in an intensive collaboration among several teams (technical, product specialist, sales, etc.). More specifically, the chatbots collect response quality with ratings (thumbs up or down) and gather feedback for the entire interaction. Using this information, such as unrecognized user questions, the chatbot team trains the bots every two weeks to ensure better intent recognition. However, the usefulness of further training is reaching its limits.

All chatbots were developed with the same technology provided by the software company Inbenta. The technology uses a symbolic AI approach combined with NLP and machine learning (ML) to find the correct response to a user input (i.e., intent recognition). Using a symbolic AI approach that applies linguistic (symbolic) reasoning to intent recognition has the advantage of better explaining the process of providing (or not providing) a particular answer to a user's question (Bolander, 2019). Each chatbot comes with a lexicon that consists of a representation of human language, symbols, and semantic relationships between words. Adjustments can be made to customize specific words or linguistic concepts to get more accurate chatbot responses. Compared to pure ML approaches, this also allows for a better understanding of how the intent recognition algorithm works. Real-world symbols are shown with their lexical units (an abstract representation of a word or group of words), words, typos, and lexical relations. For example, the word "study" has two lexical units: One is study as a verb, which has a lexical relation to the verb "to learn", among others. The other is study as a noun, which has a lexical relation to the noun "report", for example. The lexical unit is especially important when the company developing the chatbot wants to refine the algorithm by creating links for certain words (e.g., a product is called "Storm," and to improve the intent recognition, it makes sense to add a lexical relation between "storm" and "product" for this chatbot) (Inbenta, 2021). To further train the chatbot, the company adds a set of utterances for each intent. For example, the utterance "Are bikes insured in case of accident?" could be added to the intent "Are damages to bicycles insured?" When the user enters this or a similar message (e.g., "If I fall with the bicycle, is the bicycle insured?"), the chatbot can recognize this intent and send a predefined answer to the user (e.g., "Damages to the bicycle are not covered. However, …").

The chatbots use two different mechanisms to support the intent recognition algorithm. The first one is based on AIML (McTear et al., 2016). This is particularly useful in situations where only answers that (almost) exactly match the question should be displayed. This includes small talk questions such as "What is the name of the chatbot?" or "What is the weather?" The second and more advanced mechanism is the use of several important system variables to analyze user input; these variables also enable the chatbot manager to understand how a particular result is generated. For example, the NLP algorithm computes a so-called semantic weight for each lexical unit of the user message in the knowledge base based on its relevance (grammatical category of each lexical unit) and importance. Based on this analysis, a recognition score (ranging from 0 to 1.15) is calculated for each result from the knowledge base, reflecting the degree of similarity; a response is then displayed only if a certain threshold value is reached (0.4 for InsurCorp, which is considered strict). In this process, the values for system variables generated for each message could be used as characteristics of the breakdown to configure the chatbot's repair strategies. Two system variables should be highlighted: First, the *semantic weight* determines the total semantic weight of the user input by taking into account whether the words have semantic value (e.g., user query lacks information). For example, in the message "Where do I modify my insurance?" the terms "modify" (verb) and "insurance" (noun) have greater semantic weights because their grammatical categories make them more meaningful than "I" (pronoun) or "my" (determiner). The chatbot then calculates the semantic weight of a message by taking the sum of the semantic weights of each word in that message. Second, the chatbot tool detects the *percentage of unknown words* from the user input. Words are considered unknown in several situations, such as when a message contains serious misspellings ("drns" instead of "drinks"), an unfamiliar language, or unusual abbreviations or technical terms (Inbenta, 2021). For each user message, the chatbot calculates the percentage of unknown words by dividing the number of unknown words by the total number of words in the message (e.g., 0% = all words are known; 100% = all words are unknown).

4.3.2 Cluster analysis of breakdown types

To address DP1, we used a data-driven, exploratory approach and performed a cluster analysis to identify patterns in user messages that caused conversational breakdowns (Knote et al., 2021). We analyzed 5,668 messages from over 21,000 interactions between InsurCorp's German chatbots and its customers. These were the messages that resulted in conversational breakdowns because they did not meet the minimum threshold of the intent recognition algorithm (0.4). For the cluster analysis, each chatbot message was represented by a vector of four standardized variables: (1) semantic weight, (2) percentage of unknown words, (3) total word count (WC), and (4) words per sentence (WPS). We chose this set of variables because they are relevant for a chatbot's intent recognition algorithm (Inbenta, 2021) and have been used in part in related research (e.g., Sohn et al., 2021) but most importantly because they represent generalizable structural characteristics of a message and can therefore be used beyond this DSR project.

We chose a common approach by adopting a combined cluster analysis (hierarchical and non-hierarchical algorithms) (e.g., Janssen et al., 2021; Lankton et al., 2017; Vaghefi et al., 2017) to benefit from the strength of each method (Balijepally et al., 2011). Following established guidelines for cluster analysis (Hair et al., 2019; Kaufman and Rousseeuw, 2005), we first preclustered the basic structure of the data and then performed agglomerative hierarchical clustering in SPSS version 27 using the Ward algorithm and squared Euclidean distance, which are those most commonly used in IS research (Balijepally et al., 2011). After reviewing the percent change in agglomeration coefficients and the graphical dendrogram, the optimal number of clusters was determined to be 2 or 4. Four clusters would be more reasonable to obtain meaningful clusters that can be distinguished within our data to account for the different formulations of user questions (Vaghefi et al., 2017) and thus allowing for more promising opportunities for conversation repair strategies from a practical judgment (Balijepally et al., 2011). To analyze the cluster structure empirically, we also calculated the silhouette score of cohesion and separation (Kaufman and Rousseeuw, 2005), resulting in a good quality score (> .5) for four clusters (IBM, 2021; Kaufman and Rousseeuw, 2005). Therefore, we decided to use four

clusters in our analysis. In the second step, we applied the k-means algorithm to classify each chatbot message into one of the four clusters. After inspecting messages from all clusters, we labeled the final four clusters to reflect their main characteristics: (1) elaborated, (2) specific, (3) brief, and (4) cryptic. Table 1 summarizes the distribution of variables in the clusters. For example, in Cluster 1 an average message had a semantic weight of 7.89 and contained 21.75 words—4.62% of which were unknown to the chatbots—with 14.95 words per sentence.

Description	**Dataset**	**Cluster #**			
		1	**2**	**3**	**4**
Cluster Label		elaborated	specific	brief	cryptic
Cluster Size *N* (%)	5,668	763 (13.46)	1,872 (33.03)	2,516 (44.39)	517 (9.12)
Input Variables (mean values):					
1 Semantic Weight	3.37	7.89	4.32	1.80	0.88
2 Unknown Words (%)	12.35	4.62	5.27	5.40	83.22
3 Word Count	8.59	21.75	10.96	4.20	1.97
4 Words per Sentence	7.18	14.95	9.63	4.07	1.94

Table 1. Cluster analysis results – types of breakdown.

In *Cluster 1*, messages are mostly elaborated—quite long and containing more than one sentence (reflected in the WC being much higher than the WPS). The chatbot's intention recognition algorithm fails to identify what exactly the user wants, despite the semantic weight being relatively high (several grammatically important words in a message). In addition, sometimes individual sentences or parts of them are more descriptive in nature and distract the bot from the user's actual objective. An example message from this cluster is "hello I would like to take out a household insurance with you but do not know exactly the square meter specification of the apartment. is the approximate specification also in order." For *Cluster 2*, WC almost equals WPS, indicating that these messages contain on average one sentence. With a relatively high average WC, this cluster is still defined by somewhat longer messages, such as "Isn't the bicycle as a sprts [*sic*] equipment already insured worldwide anyway?" In most cases, we found that these requests address very specific queries and are likely outside the scope of the chatbots. Messages in *Cluster 3* have little semantic weight but contain words that are familiar to the chatbots' lexicon (e.g., greetings, incomplete sentences, insults). The messages are typically quite brief and also unique, such as "Are wallpapers also insured" or "Is the way to school insured," or only have one word (e.g., "fear") and thus lack context. In *Cluster 4*, the messages are more cryptic for the chatbots because they contain a high number of unknown words for the recognition lexicon. These inputs are rather short queries with serious spelling errors or typos (e.g., "tililes" or "sk") that make no sense at all, or they are requests written in another language (e.g., "Vorrei sapere come fare l assicurazione casa a berlino") (original spelling and punctuation preserved for all examples).

To sum up, by comparing the mean value of each variable for the four clusters, we see that for Cluster 1, WC and WPS are the most important, with semantic weight also of consequence, as these variables differ the most from the other clusters. In Cluster 2, most variables are slightly above the dataset mean, while in Cluster 3 most are slightly below the mean. Only the unknown words variable is similar for Clusters 1–3, with all lying below the mean for all messages. In Cluster 4, all variables have the largest distance from their respective means—either very high (unknown words) or very low.

4.3.3 Breakdown type classifier

Based on the four clusters identified, we developed a nearest centroid classifier that categorizes a user message as one of the breakdown types. This would enable a chatbot to automatically identify the breakdown type in real time during an interaction. A nearest centroid classifier is a simple but effective linear classification model that assigns an observation to the class whose centroid is closest (Tibshirani et al., 2002). To do so, it computes the Euclidean distance of a message from each of the

class centroids. For our breakdown type classifier, we used the centroids of the previously identified clusters. In k-means clustering, the centroids are represented by the mean values of the observations in the clusters (see Table 1). We implemented the classifier using the R package lolR (Bridgeford et al., 2020) with the values of the four cluster centroids and our training dataset with 5,668 messages as inputs. Given a user message with a vector of input values (i.e., semantic weight, percentage of unknown words, WC, and WPS), the classifier computes the Euclidean distance to all cluster centroids and makes a prediction by selecting the class/cluster with the shortest distance. For example, consider the message "I have life insurance from InsurCorp, for several years now. Now I have acquired a motorcycle license, does this have any consequences?" with the following characteristics: semantic weight = 6.8, percentage of unknown words = 5, WC = 21, and WPS = 10.5. The classification results indicate that the distance is shortest to the centroid of the first cluster ($dist_{elaborated}$ = 4.66) and considerably longer to the other classes ($dist_{specific}$ = 10.38, $dist_{brief}$ = 18.67, $dist_{cryptic}$ = 81.17). As a result, this message is assigned to the first breakdown type cluster, suggesting that the breakdown occurs because the message is rather long and contains too much information pointing in different directions and therefore confuses the chatbot's intent recognition algorithm. In the next stages of our DSR project, we plan to test and compare further classification algorithms (e.g., random forest); for now, we focused on a simple classifier to determine whether our clusters and classifications work as intended.

4.4 Evaluation

To evaluate the breakdown type classifier (i.e., our main artifact of the first design cycle) and the clusters identified, we relied on human coders and followed the approach described by Huang et al. (2019). For the evaluation, we used a separate test dataset from InsurCorp's chatbots with over 1,000 interactions that were not used for the cluster analysis. In total, this test dataset contained 282 messages that caused conversational breakdowns. From this dataset, we drew a random sample of 180 messages. We then instructed two human coders who were not involved in this study to manually classify these messages to one of the four breakdown types. Their instructions included a detailed explanation of the clusters along with a list of ten representative examples for each of them. After the coders had clarified any questions with us and confirmed their understanding of the clusters, they started coding the messages in the dataset. Coding was performed independently, but to obtain one set of classified data for comparison with the classifier, disagreements were resolved through discussion. Finally, we compared the results of the manual classification with the results of the automated classification by our breakdown type classifier. The calculated accuracy of the classifier was 79.44%, demonstrating that our classifier can be used to identify breakdown types from user messages that cause a breakdown during human–chatbot interaction. This result highlights the validity and usefulness of our artifact and suggests that it can serve as the starting point for our second design cycle.

5 Discussion

Due to the complexity of natural language, limitations of current technology, and organizational constraints (e.g., time spent for training), chatbots often fail to understand user input and provide answers, resulting in conversational breakdowns. Further complicating the situation, chatbots often provide rather generic repair messages (e.g., "I'm sorry, I didn't understand you. Can you please try again?") that do not help users but force them to engage in a frustrating trial-and-error process to recover from the breakdown (Ashktorab et al., 2019). To address this problem, we are conducting a DSR project to design effective conversational repair strategies for chatbots that help users recover from conversational breakdowns. Drawing on communication theories on conversational repair (Albert and Ruiter, 2018; Sacks et al., 1974; Schegloff et al., 1977), we proposed two DPs in the context of chatbot-initiated user-repair that guide our design of a solution that enables chatbots to identify causes of conversational breakdowns (DP1) and send more specific repair messages that map to the identified causes and allow users to recover from breakdowns themselves (DP2). In the first cycle of our DSR project, we focused on the instantiation and evaluation of DP1. More specifically,

we leveraged a dataset of chatbot interactions with 5,668 breakdowns from our collaboration partner InsurCorp to conduct a cluster analysis and build a breakdown type classifier. Our evaluation of the classifier with two human coders demonstrates its robustness and suggests that it can be applied to automatically identify breakdown types from user messages in real time. As the main artifact of the first design cycle, this classifier serves as a basis for the remainder of our DSR project. The results of this research provide valuable theoretical and practical implications, which we discuss in the following.

First, our research contributes to the body of design knowledge for chatbots. While previous research has provided valuable insights into the design of chatbots in terms of appearance and personality (e.g., Ahmad et al., 2020; Benner et al., 2021; Diederich et al., 2022; Seeger et al., 2021; Zierau et al., 2020), our research extends this literature by providing prescriptive knowledge on how to deal with conversational breakdowns. The results of our first design cycle demonstrate how data from existing chatbot interactions can be leveraged to identify distinct clusters of breakdowns, which can then be used as inputs to build a classifier. The evaluation of the breakdown type classifier reveals that it is possible to automatically recognize the reason for the breakdown from the messages alone. Such a classifier can be integrated into a running chatbot to enable the provision of specific repair messages that facilitate a user's understanding of why the chatbot failed and how to resolve the breakdown (see Ashktorab et al., 2019) for initial design ideas, such as acknowledging misunderstandings and helping users rephrase their requests). Similar approaches may be used in other areas to identify conversational breakdowns (e.g., voice-based CAs).

Besides prescriptive knowledge, our research offers several contributions in the form of descriptive knowledge (Gregor and Hevner, 2013). Rather than reducing conversational breakdowns to the capabilities of chatbot technologies alone, our analysis demonstrates that the extent of detection also depends on users and their different message-writing styles. This study provides novel insights on conversational breakdowns by revealing what types of breakdown messages exist and how they can be identified; although users have different queries, the underlying structure of messages can be divided into four types. Our classification of messages helps extend the theoretical understanding of conversational breakdowns by providing a novel and complementary perspective on the factors that can be extracted from messages. Thus, our results offer a detailed understanding of message characteristics (semantic weight, unknown words, WC, and WPS) associated with conversational breakdowns. For example, when we analyzed the fourth cluster, we found the messages to contain many unknown words that can be distinguished into two categories: (1) spelling errors and single characters with no meaning (perhaps because the user sent the message by mistake) and (2) different languages. Consistent with previous studies (e.g., Adam et al., 2021; Riquel et al., 2021), we also found breakdowns in Cluster 3 revealing small talk (e.g., "Thank you goodbye" or "beer?") and insults (e.g., "You are stupid!") to be popular among users. Although it is difficult for chatbots to possess the answers to all specific user queries, their small-talk capabilities should be expanded, as such interaction is independent of a specific product and could be applied across chatbots.

From a practitioner perspective, we provide implications for (1) companies employing chatbots, (2) software companies providing chatbot technology, and (3) users who benefit from chatbots. First, our analysis of the literature and real-world examples shows that for companies employing chatbots, conversation breakdowns are a real challenge in delivering a satisfying experience. However, presenting one-size-fits-all repair messages and continuing to train the intent recognition algorithm is not an ideal solution to the problem, especially as investing more time and money in chatbot training may not deliver a good balance between effort and benefit. As research has shown that helping the user recover from a breakdown is as good as avoiding a breakdown in the first place (Mozafari et al., 2022; Sheehan et al., 2020), companies should implement effective repair strategies. This paper offers first guidance in this regard by defining four types of breakdowns and illustrating how a classifier could identify them. If companies are unable to implement such a classifier by themselves (e.g., via an API integration), they should consider the ability to detect different breakdowns as an important criterion when deciding on a chatbot software. Second, although software companies are working to optimize their recognition algorithms, they are unlikely to be scalable enough to provide answers for every user query. Therefore, helping companies develop chatbots that have effective repair strategies

seems like a promising route to expand their range of services. Our results offer a greater understanding of conversational breakdowns and highlight the opportunity for software companies to identify them, thus enabling businesses to react appropriately to the different breakdown types. Furthermore, software companies might consider additional classification strategies for messages in Cluster 4, such as checking for profanity words and different languages. All in all, by supporting effective repair strategies, software companies can differentiate themselves from the rapidly growing and competitive field of chatbot software vendors. Third, with existing chatbots, users end up trapped in breakdown situations without knowing how it happened or how to fix it. When chatbots can detect breakdown types, they can answer both of these questions, leading users to forgive chatbots for the breakdowns and improving the overall experience. The findings of this paper will help practitioners better leverage incoming messages in chatbot interactions and help users recover from these situations, taking an important step toward overcoming one of the biggest challenges in human–chatbot interaction.

Despite these contributions, our ongoing work presents certain limitations that reveal opportunities for further research. First, in the current stage of our DSR project, we have yet to evaluate the holistic repair strategy, as our first cycle focused on identifying the causes of a breakdown. Therefore, more research is needed to design and evaluate the repair strategy end to end. Moreover, due to the lack of labeled training data, we currently rely on a rather simple nearest centroid classifier based on the results of our cluster analysis. However, as there are more advanced classification algorithms (e.g., random forest, gradient boosting) that may perform better, future research is needed to test and compare our classifier against ones using different algorithms. In addition, future research could validate the results of our cluster analysis by applying multiple cluster optimization techniques, reducing possible methodological shortcomings, and demonstrating the robustness of our four-cluster solution (Balijepally et al., 2011). Second, our cluster analysis specifically focused on structural characteristics of user messages (e.g., content words versus function words through semantic weight, word count) rather than the content of a message itself. Therefore, future research could extend our analysis by including content characteristics of user messages (e.g., type of request). Third, we examined real-world examples from two industries (insurance and banking) and analyzed user questions directed to chatbots of a specific insurance company. Our data and cluster analyses may not be generalizable to all types of chatbots and could be extended to chatbot interactions in other sectors. For example, chatbots designed for long-term interactions (e.g., Replika) may need different repair strategies than our chatbots, which were designed primarily to answer FAQs. Fourth, we focused on text-based CAs. As voice-based CAs are becoming more common (Reinkemeier and Gnewuch, 2022) but have different reasons for breakdowns—for example, different accents or background noise (Weber and Ludwig, 2020)—our DPs could be reviewed and adapted to such interactions. The last limitation involves our focus on messages in German; future research should apply our proposed DPs to chatbots operating in other languages.

6 Conclusion

Following the DSR approach (Hevner et al., 2004), we address the real-world challenge of repairing conversational breakdowns in human–chatbot interaction. In this paper, we report the results of our first design cycle, in which we conducted a cluster analysis of breakdown messages and then built a classifier that can automatically identify the cause of a breakdown from the user's message. The evaluation of the breakdown type classifier demonstrates its utility in addressing the first DP. The findings of the first design cycle serve as the starting point for the subsequent cycles, with the ultimate goal of designing effective conversational repair strategies for chatbots. In the next cycle we plan to conduct an experimental evaluation of different repair messages that map to our identified breakdown types. Finally, we will combine our artifacts (classifier and repair messages) and evaluate them together in a randomized field experiment with our collaboration partner. Overall, through our DSR project, we aim to advance the body of design knowledge for chatbot repair strategies and enable practitioners to design effective conversational repair strategies that provide a better user experience.

References

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

Ahmad, R., D. Siemon, D. Fernau and S. Robra-Bissantz (2020). "Introducing "Raffi": A Personality Adaptive Conversational Agent" *Twenty-Third Pacific Asia Conference on Information Systems*.

Akhtar, M., J. Neidhardt and H. Werthner (2019). "The Potential of Chatbots: Analysis of Chatbot Conversations". In: *21st IEEE Conference on Business Informatics*. *Proceedings: 15-17 July 2019, Moscow, Russia*. Ed. by J. Becker, D. A. Novikov. Los Alamitos, California: Conference Publishing Services, IEEE Computer Society, pp. 397–404.

Albert, S. and J. P. de Ruiter (2018). "Repair: The Interface Between Interaction and Cognition" *Topics in cognitive science* 10 (2), 279–313.

Ashktorab, Z., M. Jain, Q. V. Liao and J. D. Weisz (2019). "Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Ed. by S. Brewster, G. Fitzpatrick, A. Cox, V. Kostakos. New York, NY, USA: ACM.

Balijepally, V., G. Mangalaraj and K. Iyengar (2011). "Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research" *Journal of the Association for Information Systems* 12 (5), 375–413.

Benner, D., E. Elshan, S. Schöbel and A. Janson (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 Information Systems (ICIS)*.

Bolander, T. (2019). "What do we loose when machines take the decisions?" *Journal of Management and Governance* 23 (4), 849–867.

Bridgeford, E., M. Tang, J. Yim and J. Vogelstein (2020). *Package 'lolR'* URL: https://CRAN.R-project.org/package=lolR (visited on 11/12/2021).

Clark, H. H. and E. F. Schaefer (1987). "Collaborating on contributions to conversations" *Language and Cognitive Processes* 2 (1), 19–41.

Dale, R. (2016). "The return of the chatbots" *Natural Language Engineering* 22 (5), 811–817.

Diederich, S., A. B. Brendel, S. Morana and L. M. Kolbe (2022). "On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research" *Journal of the Association for Information Systems* 23 (1), 96–138.

Diederich, S., T.-B. Lembcke, A. B. Brendel and L. M. Kolbe (2021). "Understanding the Impact that Response Failure has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights from an Online Experiment" *AIS Transactions on Human-Computer Interaction* 13 (1), 82–103.

Dingemanse, M., S. G. Roberts, J. Baranova, J. Blythe, P. Drew, S. Floyd, R. S. Gisladottir, K. H. Kendrick, S. C. Levinson, E. Manrique, G. Rossi and N. J. Enfield (2015). "Universal Principles in the Repair of Communication Problems" *PloS one* 10 (9).

Feine, J., U. Gnewuch, S. Morana and A. Maedche (2019). "A Taxonomy of Social Cues for Conversational Agents" *International Journal of Human-Computer Studies* 132, 138–161.

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

Følstad, A., C. B. Nordheim and C. A. Bjørkli (2018). "What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study". In S. S. Bodrunova (ed.) *Internet Science*, pp. 194–208. Cham: Springer International Publishing.

Gartner (2020). *Top CX Trends for CIOs to Watch*. URL: https://www.gartner.com/smarterwithgartner/top-cx-trends-for-cios-to-watch/ (visited on 11/13/2021).

Gnewuch, U., S. Morana and A. Maedche (2017). "Towards Designing Cooperative and Social Conversational Agents for Customer Service". Short Paper. In: *Proceeding of the 38th International Conference on Information Systems (ICIS)*.

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

Gregor, S. and A. R. Hevner (2013). "Positioning and Presenting Design Science Research for Maximum Impact" *MIS Quarterly* 37 (2), 337–355.

Gregor, S. and D. Jones (2007). "The Anatomy of a Design Theory" *Journal of the Association for Information Systems* 8 (5), 312–335.

Gregor, S., L. Kruse and S. Seidel (2020). "Research Perspectives: The Anatomy of a Design Principle" *Journal of the Association for Information Systems* 21, 1622–1652.

Hair, J. F., W. C. Black, B. J. Babin and R. E. Anderson (2019). *Multivariate data analysis.* Eighth edition. Hampshire, UK: Cengage Learning EMEA.

Hevner, A. R., S. T. March, J. Park and S. Ram (2004). "Design Science in Information Systems Research" *MIS Quarterly* 28 (1), 75–105.

Huang, K.-Y., I. Chengalur-Smith and A. Pinsonneault (2019). "Sharing Is Caring: Social Support Provision and Companionship Activities in Healthcare Virtual Support Communities" *MIS Quarterly* 43 (2), 395–423.

IBM (2021). *SPSS Statistics 28.0.0 Model Summary View*. URL: https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=viewer-model-summary-view (visited on 03/20/2022).

Inbenta (2021). *Inbenta help center*. URL: https://help.inbenta.com (visited on 11/12/2021).

Insider Intelligence (2021). *Chatbot market in 2021: Stats, trends, and companies in the growing AI chatbot industry*. URL: https://www.businessinsider.com/chatbot-market-stats-trends (visited on 11/12/2021).

Janssen, A., L. Grützner and M. H. Breitner (2021). "Why do Chatbots fail? A Critical Success Factors Analysis". In: *Proceedings of the 42nd International Conference on Information Systems (ICIS)*.

Kaufman, L. and P. J. Rousseeuw (2005). *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken, New Jersey: Wiley-Interscience.

Klopfenstein, L. C., S. Delpriori, S. Malatini and A. Bogliolo (2017). "The Rise of Bots". In: *DIS'17. Proceedings of the 2017 ACM Conference on Designing Interactive Systems*. Ed. by O. Mival. New York, NY: ACM, pp. 555–565.

Knote, R., A. Janson, M. Söllner and J. M. Leimeister (2021). "Value Co-Creation in Smart Services: A Functional Affordances Perspective on Smart Personal Assistants" *Journal of the Association for Information Systems* 22 (2), 418–458.

Kuechler, B. and V. Vaishnavi (2008). "On theory development in design science research: anatomy of a research project" *European Journal of Information Systems* 17 (5), 489–504.

Lankton, N. K., D. H. McKnight and J. F. Tripp (2017). "Facebook privacy management strategies: A cluster analysis of user privacy behaviors" *Computers in Human Behavior* 76, 149–163.

Lee, M. K., S. Kiesler, J. Forlizzi, S. Srinivasa and P. Rybski (2010). "Gracefully mitigating breakdowns in robotic services" *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–210.

Li, T. J.-J., J. Chen, H. Xia, T. M. Mitchell and B. A. Myers (2020). "Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs". In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Ed. by S. Iqbal, K. MacLean, F. Chevalier, S. Mueller. New York, NY, USA: ACM, pp. 1094–1107.

Luo, X., S. Tong, Z. Fang and Z. Qu (2019). "Frontiers. Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases" *Marketing Science* 38 (6), 913–1084.

Maedche, A., C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana and M. Söllner (2019). "AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives" *Business & Information Systems Engineering* 61 (4), 535–544.

McTear, M. (2020). "Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots" *Synthesis Lectures on Human Language Technologies* 13 (3), 1–251.

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

Mitrevski, M. (2018). *Developing Conversational Interfaces for iOS. Add responsive voice control to your apps.* Berkeley, CA: Apress.

Moore, R. J. and R. Arar (2018). "Conversational UX Design: An Introduction". In R. J. Moore, M. H. Szymanski, R. Arar and G.-J. Ren (eds.) *Studies in Conversational UX Design*, pp. 1–16. Cham: Springer International Publishing.

Mozafari, N., M. Hammerschmidt and W. H. Weiger (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*.

Rasa (2021). *Fallback and Human Handoff*. URL: https://rasa.com/docs/rasa/fallback-handoff/ (visited on 11/12/2021).

Reinkemeier, F. and U. Gnewuch (2022). "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*, pp. 4326–4335.

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

Sacks, H., E. A. Schegloff and G. Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation" *Language* 50 (4), 696–735.

Schegloff, E. A., G. Jefferson and H. Sacks (1977). "The Preference for Self-Correction in the Organization of Repair in Conversation" *Language* 53 (2), 361–382.

Seeger, A.-M. and A. Heinzl (2021). "Chatbots often Fail! Can Anthropomorphic Design Mitigate Trust Loss in Conversational Agents for Customer Service?". In: *Proceedings of European Conference on Information Systems (ECIS)*.

Seeger, A.-M., J. Pfeiffer and A. Heinzl (2021). "Texting with Human-like Conversational Agents: Designing for Anthropomorphism" *Journal of the Association for Information Systems* 4 (22).

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

Skjuve, M. and P. B. Brandtzæg (2019). "Measuring User Experience in Chatbots: An Approach to Interpersonal Communication Competence". In S. S. Bodrunova, O. Koltsova, A. Følstad, H. Halpin, P. Kolozaridi, L. Yuldashev, A. Smoliarova and H. Niedermayer (eds.) *Internet Science*, pp. 113–120. Cham: Springer International Publishing.

Skjuve, M., A. Følstad, K. I. Fostervold and P. B. Brandtzæg (2021). "My Chatbot Companion - a Study of Human-Chatbot Relationships" *International Journal of Human-Computer Studies* 149.

Sohn, S., D. Siemon and S. Morana (2021). "When Live Chats Make Us Disclose More". In: *Proceedings of the 42nd International Conference on Information Systems (ICIS)*.

Takayama, J., E. Nomoto and Y. Arase (2019). "Dialogue breakdown detection robust to variations in annotators and dialogue systems" *Computer Speech & Language* 54, 31–43.

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression" *Proceedings of the National Academy of Sciences of the United States of America* 99 (10), 6567–6572.

Tuzovic, S. and S. Paluch (2018). "Conversational Commerce – A New Era for Service Business Development?". In M. Bruhn and K. Hadwich (eds.) *Service Business Development*, pp. 82–101. Wiesbaden, Germany: Springer Gabler.

Vaghefi, I., L. Lapointe and C. Boudreau-Pinsonneault (2017). "A typology of user liability to IT addiction" *Information Systems Journal* 27 (2), 125–169.

van der Goot, M. J., L. Hafkamp and Z. Dankfort (2021). "Customer Service Chatbots: A Qualitative Interview Study into the Communication Journey of Customers". In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin and P. B. Brandtzæg (eds.) *Chatbot Research and Design*, pp. 190–204. Cham: Springer International Publishing.

Weber, P. and T. Ludwig (2020). "(Non-)Interacting with conversational agents". In: *MuC '20: Proceedings of the Conference on Mensch und Computer*, pp. 321–331.

Yuan, S., B. Brüggemeier, S. Hillmann and T. Michael (2020). "User Preference and Categories for Error Responses in Conversational User Interfaces". In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. New York, NY, United States: ACM.

Zierau, N., E. Elshan, C. Visini and A. Janson (2020). "A Review of the Empirical Literature on Conversational Agents and Future Research Directions". In: *Proceedings of the 41st International Conference of Information Systems (ICIS)*.