Lewandowski, Tom; Kučević, Emir; Leible, Stephan; Poser, Mathis; Böhmann, Tilo

Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement

Electronic Markets

Provided in Cooperation with:
Springer Nature

Suggested Citation: Lewandowski, Tom; Kučević, Emir; Leible, Stephan; Poser, Mathis; Böhmann, Tilo (2023) : Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 33, Iss. 1, https://doi.org/10.1007/s12525-023-00662-3

Electronic Markets (2023) 33:39
https://doi.org/10.1007/s12525-023-00662-3

RESEARCH PAPER

Enhancing conversational agents for successful operation: A multi-perspective evaluation approach for continuous improvement

Tom Lewandowski[1] · Emir Kučević[1] · Stephan Leible[1] · Mathis Poser[1] · Tilo Böhmann[1]

Received: 31 March 2023 / Accepted: 11 July 2023 / Published online: 5 August 2023

Abstract

Contemporary organizations increasingly adopt conversational agents (CAs) as intelligent and natural language-based solutions for providing services and information. CAs offer new forms of personalization, speed, (cost-)effectiveness, and automation. However, despite their hype in research and practice, many organizations still fail to seize CAs' potential because they lack knowledge of how to evaluate and improve the quality of CAs to sustain them in organizational operations. We aim to fill this knowledge gap by conducting a design science research project in which we aggregate insights from the literature and practice to derive an applicable set of quality criteria for CAs. Our article contributes to CA research and guides practitioners by providing a blueprint to structure the evaluation of CAs and to discover areas for systematic improvement.

Keywords Artificial intelligence assistants · Conversational agents · Chatbots · Quality criteria set · Design science research (DSR)

JEL Classification M15 · O32/O33

Introduction

Recent technological advancements in intelligent and natural language-based information systems (IS) transform everyday life, work, and interactions (Brynjolfsson & McAfee, 2014; Davenport & Kirby, 2016; Diederich et al., 2020). As a result of ongoing developments in artificial intelligence (AI) and improvements in the underlying machine learning (ML) and natural language processing (NLP) algorithms, conversational agents (CAs) are becoming increasingly relevant in organizations as essential gateways to digital services and information (Følstad et al., 2021; Gnewuch et al., 2018). In this context, a recent analysis valued the global market for CAs at $3.49 billion in 2021 and expects it to grow to $22.9 billion by 2030, indicating their increasing importance (Research & Markets, 2022). Primarily operating in external or internal organizational environments (Patel et al., 2021), CAs interact with users (e.g., customers and employees) via natural language to provide convenient access to information from multiple connected systems and data sources. Moreover, CAs can perform standardizable processes and (cost-) effectively automate or assist tasks conventionally performed by employees (Meyer von Wolff et al., 2020). In terms of automation, Gartner predicts that CAs will automate one in ten agent interactions by 2026 (Rimol, 2022). Consequently, CAs are expected to deliver significant economic value in existing and future applications, businesses, and digital ecosystems (Seeger et al., 2021; Seiffer et al., 2021).

Due to their potential, an extensive stream of research has focused on these AI-based systems (Cui et al., 2017; Zierau et al., 2020b). Since 2016, known as the "year of the chatbot" (Dale, 2016, p. 811), interdisciplinary research has explored various aspects related to CAs (Diederich et al., 2019a; Janssen et al., 2020), leading to a significant increase in both scientific and practical knowledge (Zierau et al.,

Communicated by Rainer Schmidt

✉ Tom Lewandowski
tom.lewandowski@uni-hamburg.de

Emir Kučević
emir.kucevic@uni-hamburg.de

Stephan Leible
stephan.leible@uni-hamburg.de

Mathis Poser
mathis.poser@uni-hamburg.de

Tilo Böhmann
tilo.boehmann@uni-hamburg.de

[1] University of Hamburg, 22527 Hamburg, Germany

2020a). More specifically, previous research has examined technical aspects (e.g., NLP improvements) as well as framework and platform selection (Diederich et al., 2019a, 2019b; Følstad et al., 2021). Scholars have also investigated user attitudes toward CAs, including acceptance, motivation, and behavioral implications, such as user trust (e.g., Brandtzaeg & Følstad, 2017; Go & Sundar, 2019; Seeger et al., 2017). In addition, prior studies have also focused on interaction design (e.g., Bittner et al., 2019; Gnewuch et al., 2018) and user preferences for visual cues and conversational design of CAs (e.g., Feine et al., 2019a; Schuetzler et al., 2021). Furthermore, social, ethical, and privacy challenges associated with CAs' implementation and use have been explored (e.g., Ischen et al., 2020; Ruane et al., 2019; Wambsganss et al., 2021).

Despite the steep increase in CA research and their vast opportunities, adopting CAs in organizational environments does not always have a positive impact because the technology is still error-prone and fails in interactions (Gnewuch et al., 2017; Riquel et al., 2021). These deficiencies concern CAs of varying maturity levels and regularly result in incorrect responses and conversational breakdowns (Weiler et al., 2022). Attributable to inadequate CAs, employees have developed negative feelings toward CAs and their providers in recent years (Diederich et al., 2020; Feine et al., 2019b; Schuetzler et al., 2021). To date, several potential reasons for the shortcomings and moderate success of CAs have been identified.

First, a primary reason for the limited success of CAs is their premature deployment, often driven by high expectations and management pressure, usually combined with little knowledge of the CA development process in general and of CA quality in particular. This practice often leads to nonuse, dissent, or complete failure, as highlighted by Janssen et al. (2021) and Lewandowski et al. (2022b). Unsatisfactory CA design and limited capabilities can result in a frustrating user experience that triggers resistance and a loss of trust in the CA, further hindering its successful adoption in real-world environments as organizations (Weiler et al., 2022). The failure of CAs is frustrating not only for employees, but also for the CA vendor, who has invested significant effort, time, and money in developing the CA (Janssen et al., 2021; van der Goot et al., 2021).

Second, CAs are only marginally or not continuously evaluated to ensure their improvement, successful operation, and overall progress in organizations (Janssen et al., 2021; Meyer von Wolff et al., 2021). Therefore, previous research has proposed continuous evaluation (e.g., via monitoring (Corea et al., 2020) or chatlog data (Kvale et al., 2019)) and operation and improvement processes (Lewandowski et al., 2022b; Meyer von Wolff et al., 2022) to regularly assess their use, quality, and added value (Brandtzaeg & Følstad, 2018; Meyer von Wolff et al., 2022). However, little is known about how to systematically organize this operation and improvement process to improve CAs. Previous studies have focused on single perspectives, such as continuous technical adaptations (e.g., retraining the NLP algorithm (Meyer von Wolff et al., 2022)), adjusting the knowledge base (Janssen et al., 2021; Jonke & Volkwein, 2018), and improving individual CA functionalities and the dialog flow based on previous failures identified by chatlogs (e.g., Kvale et al., 2019). Despite these insights, there is a lack of knowledge on how CAs can be evaluated with criteria to test and improve their quality throughout their lifecycle (Lewandowski et al., 2022b). In addition, the current findings are often relatively fragmented across disciplines and application domains, and they therefore lack a cohesive axis of transferability for sustained practical usage (Elshan et al., 2022a; Følstad et al., 2021; Li & Suh, 2022). In this regard, experts in the field urge for more collaboration and aggregation in interdisciplinary research on CAs, and encourage further research on the topics around measurement, modeling, and evaluation approaches for CAs, as outlined, for example, in the CA research agenda by Følstad et al. (2021). To the best of our knowledge, there is no holistic overview of criteria for researchers and practitioners and approaches to continuously evaluate, improve, and sustain CAs that would facilitate organizations in this problem context. Therefore, this article explores the following research question:

What are relevant criteria for continuously evaluating the quality of CAs, and how can they be applied?

By addressing the research question, this article aims to systematize the continuous evaluation and improvement of CAs to counteract CA failure in organizational environments. To successfully operate a CA, measurements or criteria are needed for orientation to adapt CAs to user needs (Følstad et al., 2021; Meyer von Wolff et al., 2022). Therefore, we pursued a twofold contribution: (1) a set of relevant criteria to evaluate the quality of CAs and (2) a procedure model as part of the instantiation of the quality criteria set in an IT organization, prescribing its application and evaluation activities. The criteria set and procedure model define a cyclical criteria-based evaluation process that can be triggered by different impulses. These results address the identified research gap and present an approach for practice. Specifically, our proposed quality criteria set addresses this lack of knowledge about the successful operation of CAs. In this context, the evaluated set of quality criteria and the procedure model can serve as an initial overview for organizations to structure CA evaluations systematically and discover areas for improvement. Following design science research (DSR) activities mapped to the three-cycle view by Hevner (2007), we approach the derivation of these results

with the following structure: First, we present the related CA research and delineate the research gap in more detail. Next, we describe our research approach to developing our artifact. We then present the findings of our work, including an overview of our final quality criteria set. Subsequently, we outline the instantiation of the quality criteria set using a real-life case in an IT organization. Finally, we discuss our findings as well as their implications and conclude with our limitations and potential future research.

Related research

Text-based conversational agents as specific AI-based IS

Research on AI-based IS has attracted substantial attention (Elshan et al., 2022b; Felderer & Ramler, 2021) and transformed from a technical trend to a pervasive phenomenon in our daily lives (Maedche et al., 2019). AI-based systems proliferate in various application domains and contribute to multiple innovations (Wang et al., 2020). One application area that has seen renewed interest and increasingly utilizes AI is communication with computers via natural language, which has been a topic of research and practice for several decades (Gnewuch et al., 2017). Since the 1960s, researchers have worked on text-based and later speech-based CAs to automate procedures and assist users with various tasks (Følstad et al., 2021). An early example is ELIZA, which allows initial natural language-based interactions with a computer (Weizenbaum, 1966). However, technical limitations (e.g., computational power and storage capacity) and overly simplistic capabilities (e.g., non-learning algorithms) restricted early attempts at CAs, as they could not meet the high expectations (Diederich et al., 2019a; Gnewuch et al., 2017). According to Dale (2016) and Klopfenstein et al. (2017), ELIZA and other previously developed CAs used simple rule-based mechanisms to generate responses.

Nevertheless, in recent decades, technological progress has allowed the development of more sophisticated CAs that utilize novel AI, ML, and NLP algorithms and models (Gnewuch et al., 2017). In this context, the CA attempts to understand the user's intention behind the input prompt to provide an adequate response output. In particular, the techniques of supervised learning, unsupervised learning, and human-in-the-loop (where humans are involved in the training process) lead to increasingly better CAs (Radziwill & Benton, 2017; Wiethof & Bittner, 2021). As a result, they have gained widespread adoption and can now better address the needs of the general public and the mass market (Maedche et al., 2019).

CAs support the ongoing digitalization and automation of organizations by performing various activities, such as filtering information or efficiently assisting employees in their daily tasks (Zierau et al., 2020a). Hence, with their scalability and 24/7 availability (Gnewuch et al., 2017; Xu et al., 2017), CAs can have a transformative impact on business operations by acting as a central service platform and first point of contact for customers, providing a convenient way to handle service requests more individually before human intervention (Zierau et al., 2020a), and reducing information overload for users (Xu et al., 2017). Accordingly, employees can concentrate on more complex, creative, and non-routine tasks.

The widespread use of CAs has generated significant research interest, with a rapidly growing body of contributions. However, CA research has a strong interdisciplinary character and is fragmented into several research streams (Følstad et al., 2021): Multiple perspectives and disciplines, including "informatics, management and marketing, media and communication science, linguistics and philosophy, psychology and sociology, engineering, design, and human–computer interaction" are employed to study CAs (Følstad et al., 2021, p. 2916). This interdisciplinary research has introduced numerous designations, such as chatbots (e.g., Dale, 2016), conversational (user) interfaces (e.g., Herrera et al., 2019), or dialog systems (e.g., McTear, 2021), leading to debates in the literature about their terminology and classifications. In this context, the authors Gnewuch et al. (2017), for example, have divided these AI-based IS into two subclasses: text-based CAs (e.g., chatbots or natural dialog systems) and speech-based CAs (e.g., smart speakers or virtual assistants).

In this article, we use the term "conversational agent" to refer to all AI- and text-based representations, such as chatbots. Although some research indicates that the distinction between text- and speech-based CAs is marginal since speech-based input can be transferred to text-based input and vice versa from a technical viewpoint (Diederich et al., 2019b), research has also revealed that evaluating speech-based CAs requires distinct criteria compared to text-based CAs. For instance, evaluating the quality of a smart speaker involves design elements, such as overall (hardware) appearance (including styling elements and imagery), as discussed in Su and Hsia (2022). In addition, privacy handling is an important issue, for instance, when referring to the proactive (i.e., listening continuously to react) or reactive (i.e., reacting restricted to specific keywords) activation of speech-based CA, as discussed by Burbach et al. (2019). Furthermore, the ability of smart speakers to process audio speech and handle different dialects, tonalities, and noise in different input environments is crucial (e.g., Bisio et al., 2018), as is robust output generation (e.g., text-to-speech generation and perception of understandability and naturalness (Schmitt et al., 2021)). In summary, while text- and speech-based CAs share some

commonalities, evaluating their quality requires considering specific criteria unique to each modality.

Continuous evaluation and improvement of conversational agents

While the development of a CA has become much more accessible, the underlying IS is complex by nature (Maroengsit et al., 2019). Besides the described possibilities and applications of CAs, the management, evaluation, and improvement of these AI-based systems pose new challenges for organizations. These activities are essential because disregarding them can result in high failure and discontinuation rates (Diederich et al., 2020; Janssen et al., 2021). Many CAs have failed in real-world environments due to, among other reasons, frustrating user experiences (Følstad et al., 2018a). As a result, multiple organizations have taken their CAs offline since they lack knowledge of how to ensure continuous evaluation and improvement, leading to an uncoordinated and highly exploratory development process (Janssen et al., 2021). As CAs represent a novel form of IS with distinct characteristics that differentiate them from traditional IS and other AI-based systems, they require new approaches for design, evaluation, and improvement.

One unique characteristic of CAs is their sociability. As *social* IS, they are capable of interacting with users via natural language, representing a new sociotechnical application class (Maedche et al., 2019). These AI-based systems impact traditional service delivery and enable new individualized and convenient sociotechnical interactions (Klaus & Zaichkowsky, 2020), requiring humanlike, user-centered, and socially interactive IS design (Lewandowski et al., 2022a). Contrary to the classification of AI-based CAs as IS from a technological perspective, the existing literature shows that the organizational adoption and practical use of CAs must be viewed in fundamentally different ways (e.g., Corea et al., 2020; Lewandowski et al., 2021). As a result, CA teams are designing chatbots differently from traditional IS and from multiple new perspectives, having equipped them with social features, names, avatars, and communicative behaviors to attract users' attention and simulate natural conversation (McTear et al., 2016). Nonetheless, enhancing the user experience of CAs remains a crucial challenge owing to the absence of a comprehensive overview to determine whether they are well-designed and useful, and because of the lack of widely applied approaches to evaluate and improve them, as described in the interdisciplinary chatbot agenda by Følstad et al. (2021).

Another unique characteristic of current CAs is their level of intelligence and ability to learn and improve via naturalistic interactions. As such, they can be classified as a form of *learning and intelligent IS,* depending on the ongoing development and introduction of, so far, unsolved challenges (Lewandowski et al., 2021; Zierau et al., 2020a). CAs often have limited skills initially, and learning progress depends on the application area and the actors' engagement in training these systems. Accordingly, CAs' learning progress is highly context-driven and thus dependent on actual application and usage (Clark et al., 2019; Zierau et al., 2020c). The learning nature of CAs indicates the necessity for novel approaches to their evaluation and improvement (Lewandowski et al., 2022b; Meyer von Wolff et al., 2021).

Consequently, the highest effort needs to be invested in operations where CAs require continuous evaluation and later training and improvement in a real-world context. This endeavor is complicated by rapid changes and high dynamics, in which it is generally impossible to predict how users will interact and what information will be retrieved long-term (Janssen et al., 2021). CAs have gained a great deal of research attention, with perspectives ranging from specific conceptual- or usability-related aspects to technical design. However, detailed theoretical and practical knowledge is lacking for the operation in general and the continuous improvement process of CAs in particular (Lewandowski et al., 2022b; Meyer von Wolff et al., 2021). Hence, a comprehensive and systemized criteria-based approach to continuously evaluate CAs' quality can help to improve and sustain them.

Evaluation criteria for conversational agents

In recent years, the overall user experience and improvement of CAs have been prominent topics in research endeavors. There is a growing body of knowledge on methods and measures to evaluate the overall user experience with CAs, resulting in initial factors contributing to a positive or negative user experience (Følstad et al., 2021; Zarouali et al., 2018). In addition, authors have examined various effects of CAs at the individual level, either on perceived human likeness, trust, perceived social support, enjoyment, affordance theory (Lee & Choi, 2017; Stoeckli et al., 2019; Zierau et al., 2020b), or in the broader context of IS acceptance theories, such as in the "Technology Adoption Model" (e.g., Pillai & Sivathanu, 2020). However, there is little research on concrete quality criteria that can be applied to ensure systematic CA evaluation and improvement. Thereby, scholars call to establish convergence in interdisciplinary CA research in measurements, models, and approaches for evaluating CAs (Følstad et al., 2021).

Contributions referring to the design and evaluation of CAs are beginning to emerge. According to Følstad et al. (2021), there is a rapidly growing body of work on CA interaction design (e.g., Ashktorab et al., 2019), CA personalization (e.g., Laban & Araujo, 2020; Shumanov & Johnson, 2021), use of interaction elements (e.g., Jain et al., 2018), social cues

(e.g., Feine et al., 2019a; Seeger et al., 2021), and capability representation. However, current research is often confined to (1) single design issues or the effects of dedicated design elements (e.g., Seeger et al., 2021), (2) technical measurements or technical performance (e.g., Alonso et al., 2009; Goh et al., 2007), (3) other agent classes, such as embodied or speech-based CAs (e.g., Kuligowska, 2015; Meira and Canuto 2015), and (4) individual design aspects (e.g., Seeger et al., 2021), while (5) being segregated through the interdisciplinary CA research landscape. In addition, CA-oriented research has (6) focused on satisfaction issues, such as human behavior or ethical aspects (e.g., Neff & Nagy, 2016; Radziwill & Benton, 2017), affect and emotions, such as mood adjustment, entertainment, and authenticity (e.g., Meira and Canuto 2015; Pauletto et al., 2013; Radziwill & Benton, 2017) and (7) initial classifications and typologies for high-level analysis and guidance on interaction design (Følstad et al., 2018b), which only play an overarching role for development.

Important preliminary work includes ISO 9241-oriented CA evaluation criteria sets, such as those of Radziwill and Benton (2017), Casas et al. (2020), and Johari and Nohuddin (2021), representing first CA quality criteria sets and approaches. However, they tend to focus on improvements at a high meta-level, such as those regarding efficiency (e.g., robustness to manipulation or unexpected input), effectiveness (e.g., if the CA passes the Turing test), impact and accessibility (e.g., meets neurodiverse needs), trustworthiness and transparency (e.g., security and intrusiveness), and humanity and empathy (e.g., the realness of a CA or personalization). While these criteria may provide valuable guidance in the initial evaluation and improvement of CAs by addressing technical concerns, such as increasing the accuracy of NLP components or conducting user surveys to gauge initial perceptions, they have limited utility for CA teams in organizations seeking to ensure the long-term success of CAs within an application environment. This limitation necessitates a comprehensive system-wide perspective, for example, with respect to the overall input processing, the output presentation, representation elements, or the design of the dialog flow. To fill this research gap, it is essential to develop a more detailed, multi-perspective, and comprehensive set of quality criteria for researchers and practitioners that addresses a broader range of requirements for the long-term success of CAs.

Research approach

Our objective is to create a set of quality criteria for CAs as a central artifact that allows organizations to continuously evaluate and improve their CAs. To achieve this objective, we used the DSR paradigm and applied the three-cycle view presented by Hevner (2007). DSR is well-established in IS research and appropriate for our research because we aim to create an artifact that addresses a real-world problem and enables the continuous improvement of CAs to counteract their failure (Gregor & Hevner, 2013). Following the classification of contribution types in the DSR of Gregor and Hevner (2013), this research contributes knowledge at different levels. We contribute to level two by creating an operational artifact in the form of a set of quality criteria, including a procedure model (design knowledge). We also contribute to level one (artifact instantiation) by applying the quality criteria in a real-world context. We aim to derive and generate prescriptive knowledge from the descriptive knowledge extracted and evaluated from the knowledge base (Drechsler & Hevner, 2018). This knowledge will serve as a normative blueprint for practitioners and starting point for further research. To structure our research endeavor according to the established ground rules of DSR, we conducted seven research steps, as illustrated in Fig. 1.

Step 1 of the DSR approach refers to the identification and formulation of a pervasive real-world problem. The initial situation was investigated through two semi-structured interviews following a prepared interview guide (see Table 1), revealing that the overall quality and usage rate of their used CA (*ExpertBot*) was insufficient, and at the same time, the IT organization lacked concrete criteria and an improvement process for it. Supplementing these insights, we examined the successful and failed use cases of organizational CAs in the current body of literature to highlight the practical relevance of the problem beyond our specific case. This status quo demonstrates the need for a solution approach that defines the addressed overarching problem class. Therefore, our research is based on the current knowledge gap regarding how a criteria-based endeavor could sustain the operation and continuous improvement of CAs to ensure their long-term success. This knowledge gap was grounded and described in the *Introduction* and *Related research* sections. As a result, we adopted a problem-centered perspective at the beginning of our research, based on Peffers et al. (2007).

Based on the formulated problem, in *Step 2*, we conducted a structured literature review (SLR) to derive the initial criteria for evaluating CA quality. We followed the five-step process of vom Brocke et al. (2009) in the databases of AISeL, ACM DL, IEEE Xplore, EBSCO, and ProQuest ABI/INFORM. We defined the scope of our SLR using the taxonomy proposed by Cooper (1988), as shown in Table 2. We focus on research outcomes, practices, and applications of quality criteria for CAs. Our goal is to address the lack of anchor points that allow continuous evaluation and improvement of CAs by synthesizing the relevant literature. We adopted a neutral perspective by paying

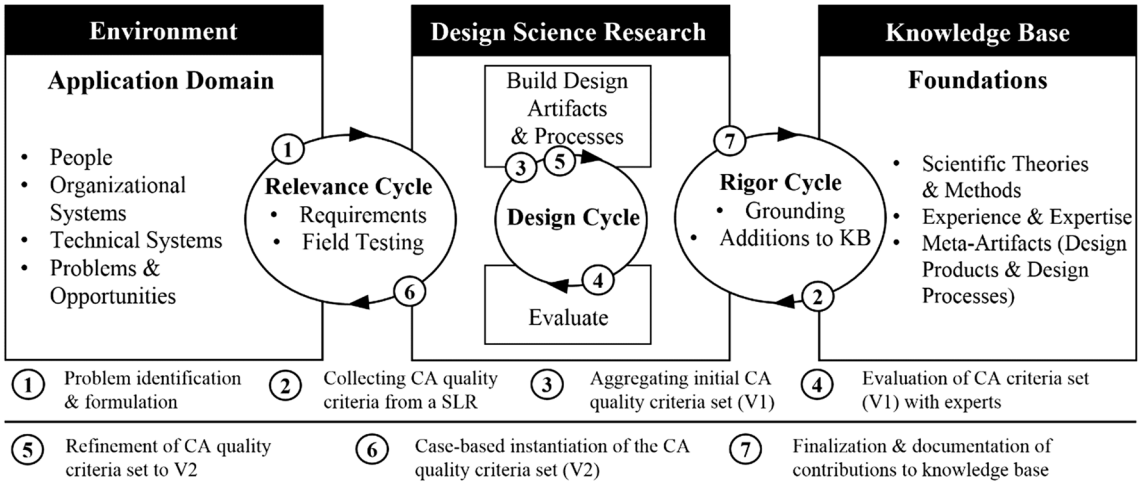

Fig. 1 DSR three-cycle view and our research steps based on Hevner (2007)

attention to different existing (interdisciplinary) criteria sets of CAs and methods of measuring their effectiveness. Our coverage strategy followed a representative nature, focusing specifically on essential and influential literature to answer our research question. A conceptual organization was chosen to cluster the existing research contributions. The results of our literature review are intended for IS researchers and interdisciplinary researchers concerned with CAs. Furthermore, practitioners can apply the derived quality criteria and procedures to improve their CAs.

We first identified the central terms in our research question and decomposed them into related concepts to construct a search term (Brink, 2013; Xiao & Watson, 2019). Next, we used the resulting terms to conduct an initial unstructured literature search of the databases: "conversational agent," "evaluation," "criteria," and "qualit*." We extracted keywords, synonyms, and homonyms from the relevant papers found (Rowley & Slack, 2004; vom Brocke et al., 2009) and used them to form the following search string: ("chatbot" OR "dialogue system" OR "conversational agent" OR "virtual assistant" OR "cognitive assistant") AND ("qualit*" OR "design" OR "criteria" OR "effectiveness" OR "evaluation" OR "usability"). We applied the search string to the aforementioned databases, resulting in 1895 articles. After screening the titles, abstracts, and keywords of each article, we selected 180 articles for in-depth analysis. To further filter the literature corpus, we established exclusion criteria to ensure that only relevant articles were included in the dataset. Two researchers independently used these criteria to screen the articles and reduce potential selection biases. Subsequently, we removed articles that addressed (1) technical or architectural aspects, (2) physical machines or robotics and their interfaces, or (3) no specific use cases for CAs. We also removed duplicates. During this process, we reduced our literature dataset to 94 articles by examining their research questions and results sections. In a final rigorous full-text analysis, we identified 67 articles as relevant, consisting primarily of journal and conference articles. Figure 2 illustrates the literature review process.

In *Step 3*, we embarked on the first design cycle to establish a quality criteria set, version 1 (V1). To do so, we followed a multi-step procedure. Initially, we independently extracted appropriate quality criteria by conducting a full-text analysis of the final 67 articles from *Step 2*. Next, we integrated the extracted criteria into a shared document containing 221 criteria, with brief descriptions and references. We then refined and streamlined the criteria based on three aspects. First, we sorted all criteria by topic and removed nonrelevant criteria for our research scope (*Step 2*). Therefore, we excluded, for example, non-CA-specific criteria, irrelevant to the evaluation of text-based CAs (e.g., those relevant only to speech-based or embodied assistants). Second, we combined criteria that were indistinguishable and removed redundant criteria. Third, we weighted the criteria based on their frequency in the reviewed literature. Due to the quantity and complexity of the collated quality criteria set, we developed a multi-level model consisting of three levels: meta-criteria, criteria, and sub-criteria. This hierarchical arrangement allows for the holistic or selective

Table 1 List of interviewees of Step 1 and the first relevance cycle

ID	Interviewee	Duration
Exp1.1	AI expert & senior project manager	56 min
Exp2.1	Software engineer & project manager	57 min

Table 2 Applied taxonomy of literature reviews by Cooper (1988)

Characteristic	Categories				
Focus	**Research outcomes**	Research methods	Theories		**Applications**
Goal	**Integration**	Criticism		**Central issues**	
Perspective	**Neutral representation**		Espousal of position		
Coverage	Exhaustive	Exhaustive (selective citation)	**Representative**		Central or pivotal
Organization	Historical	**Conceptual**		Methodological	
Audience	**Specialized scholars**	General scholars	**Practitioners or policy makers**		General public

application of the quality criteria set, enabling the evaluation of specific (topic-based) areas without needing to use the entire set.

In *Step 4*, we evaluated the initial literature-based quality criteria set (V1) through semi-structured interviews to expand the set in a second design cycle. We used Venable et al.'s (2016) Framework for Evaluation in Design Science (FEDS) throughout this process to define the overarching evaluation strategy. Our primary goal was to review and improve the quality criteria set developed for evaluating and improving CAs. Therefore, we chose a formative ex-ante approach to evaluate the quality criteria set for this design cycle. To prepare for the interviews, a semi-structured interview guide with questions about all quality criteria was created to ensure a systematic procedure and comparably gathered data. We then conducted and recorded seven interviews with experts from an IT organization with professional experience in CA projects and external researchers, following the methods of Gläser and Laudel (2009) and Meuser and Nagel (2009). We discussed the possible quality criteria of CAs with the interviewees based on their expertise before individually presenting and assessing our identified ones and having them extend our existing quality criteria set and point out missing aspects. Table 3 presents the list of interviewees of this second design cycle.

Building on these insights gained from the interviews conducted in *Step 4*, we developed V2 of our quality criteria set in *Step 5*. During the interviews, we received feedback from experts and gathered valuable input on the initial criteria set (V1). We decided whether a criterion had to be retained, revised, or added to the criteria set. In this context, we considered the experts' suggestions on the wording and structural arrangement of the criteria for reasons of comprehensibility, leading to design adjustments in V2. Furthermore, the experts' experience with CAs in real-world contexts led to the identification of additional quality criteria, which were integrated into V2 in a complementary manner where appropriate.

Table 3 List of interviewees of *Step 4* and the second design cycle

ID	Interviewee	Duration
Exp1.2	AI expert & senior project manager	42 min
Exp2.2	Software engineer & project manager	42 min
Exp3	Principal data manager	36 min
Exp4	Branch manager	35 min
Exp5.1	CA developer/engineer	29 min
Exp6	CA researcher	40 min
Exp7	CA researcher	34 min

In *Step 6*, we conducted a summative naturalistic ex-post evaluation of the quality criteria set (V2) by supervising its case-based instantiation in an IT organization using the FEDS (Venable et al., 2016). Our goal was to verify if the criteria set could be used to evaluate CA quality and whether it could help organizations improve their CAs in a structured and normative way by emphasizing its usefulness and relevance. To achieve this, we developed a procedure model for the application and instantiation of the quality criteria set and conducted two interview rounds. The first round included seven experts, three of them from *Step 4* and four new participants; in the second round, one of the new participants was not available, so 13 interviews were conducted in total, as shown in Table 4. During the first round, we asked the experts about the current state version of *ExpertBot* and its problems and potential for improvement before transitioning to the individual criteria from our set to create suitable scenarios. We used scenarios as flexible containers that include a certain number of our quality criteria that match a collective evaluation. In this context, mockups were created with Figma (2022) as prototypes to simulate each scenario with a current state version and a modified version of *ExpertBot,* incorporating altered criteria aligned with our criteria set (V2). In the second round, we used the created prototypes to simulate each scenario previously defined in A/B tests related to Young (2014). This allowed us to determine which criteria were considered highly influential and most important to the experts. In addition, we paid attention to whether the experts mentioned new criteria in the instantiation that were not yet included in our quality criteria set. The

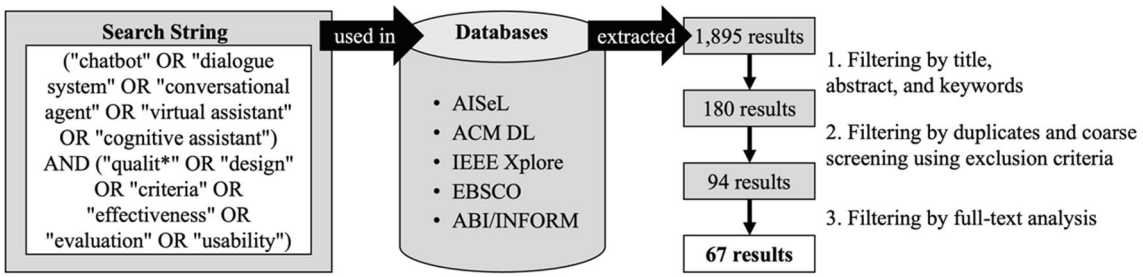

Fig. 2 Literature review process according to vom Brocke et al. (2009)

procedure is more detailed in the "Case-based instantiation of the quality criteria set" section below.

Finally, in *Step 7*, we incorporated the evaluation results from *Step 6*, the naturalistic case-based instantiation, into the quality criteria set and developed the final version presented and documented in the next section. Following Gregory and Muntermann's (2014) theorizing framework, we iteratively developed and improved an abstracted artifact version that met the larger problem class derived in *Step 1*. We communicate the quality criteria set for CAs as a rigorously elaborated prescriptive artifact, providing applicable knowledge that contributes to the knowledge base as a solution design entity for practitioners with an adaptable framework for situational instantiations to improve their CAs by applying the derived quality criteria (Drechsler & Hevner, 2018). In addition, our set provides descriptive knowledge as an observation and classification concept for researchers, with new insights and starting points for further research on evaluating, understanding, and improving CAs for their long-term success.

Quality criteria set for conversational agents

Based on the DSR research activities, we derived the final criteria set for evaluating and improving the quality of CAs. This set incorporates a hierarchical structure consisting of 6 metacriteria, 14 criteria, and 33 sub-criteria, enabling a systematic and rigorous evaluation process. The *meta-criteria* are the highest level of abstraction, representing the overarching evaluation areas of a CA. The *criteria* at the second level break them down. These can be used, for example, to create responsibilities in a CA team for (meta-)criteria areas, ensuring that accountability is clearly defined and understood. This structure also supports informed decision-making (e.g., prioritizing specific criteria of the CA). Although (meta-)criteria provide logical and structural clarity and classification, they are not sufficiently granular for evaluation purposes. Therefore, at the third level, *sub-criteria* have been defined as specific elements that can be evaluated using qualitative or quantitative methods. Overall, this approach allows for a detailed and comprehensive evaluation of CAs. The following section presents the quality criteria along the six meta-criteria and their hierarchical structures depicted in Table 5.

Input

Input comprises criteria that focus on creating and submitting requests to the CA. In this context, the diverse *interaction abilities* of CAs can be evaluated (e.g., Kowald & Bruns, 2020). Many CA teams employ existing *communication channels* (e.g., messenger front ends, such as Microsoft Teams or websites), ensuring that users are comfortable and familiar with their basic functions (Feng & Buxmann, 2020). However, reflecting, exchanging, or expanding channels with progressive development is essential. Moreover, various input *control elements* can

Table 4 List of interviewees of *Step 6* and the second relevance cycle

ID	Interviewee	Duration – round 1	Duration – round 2
Exp1.3	AI expert & senior project manager	34 min	41 min
Exp2.3	Software engineer & project manager	39 min	28 min
Exp5.2	CA developer/engineer	41 min	38 min
Exp8	Product owner	33 min	28 min
Exp9	Management assistance	35 min	34 min
Exp10	Senior software architect	38 min	32 min
Exp11	Senior software engineer	34 min	-

Table 5 Final CA quality criteria set

Meta-criteria	Criteria	Sub-criteria	Example references
Input	Interaction abilities	Communication channel	(Feng & Buxmann, 2020), Interviews
		Control elements	(Kowald & Bruns, 2020; Li et al., 2020), Interviews
	Context awareness	Dialog-oriented context	(Diederich et al., 2020; Michaud, 2018; Saenz et al., 2017)
		Technical environment	Interviews
Output	Format	Visual elements	(Edirisooriya et al., 2019; Feng & Buxmann, 2020; Kowald & Bruns, 2020), Interviews
		Readability and consistency	
	Content	Transparent capabilities and limitations	(Diederich et al., 2020; Saenz et al., 2017)
		Information retrieval	(Diederich et al., 2020; Edirisooriya et al., 2019), Interviews
		Detail of knowledge	Interviews
		Solution convergence and justification	Interviews
	Calibration	Response appropriateness	(Hu et al., 2018; Jiang & Ahuja, 2020)
		Response accuracy	
	Time	Technical response time	(Edirisooriya et al., 2019; Meyer-Waarden et al., 2020), Interviews
		Balance between proactivity and interruption	(Feng & Buxmann, 2020)
Anthropomorphism	Humanlike identity	Identity and characteristics	(Schuetzler et al., 2021; Seeger et al., 2021)
		(Humanlike) visual representation	Interviews
	Verbal cues	Emotional expressions	(Saenz et al., 2017; Seeger et al., 2021)
		Chitchat/smalltalk	(Grudin & Jacques, 2019; Huiyang & Min, 2022; Schuetzler et al., 2021)
		Tailored personality and lexical alignment	
	Nonverbal cues	Emoticons	(Gnewuch et al., 2018; Schuetzler et al., 2021; Seeger et al., 2021), Interviews
		Typing delay and indicator	
Dialog control	Regular operation	Reformulate requests and alternative responses	(Diederich et al., 2020; Saenz et al., 2017), Interviews
		Conversational prompts and suggestions	(Kowald & Bruns, 2020; Li et al., 2020)
	Failure operation	(Proactive & resilient) repair strategies	(Benner et al., 2021; Diederich et al., 2020; Feng & Buxmann, 2020), Interviews
		Fallbacks and handover	(Poser et al., 2021, 2022; Wintersberger et al., 2020)
Performance	Effectiveness	Task success rate	(Peras, 2018), Interviews
		Task failure rate	
		Retention and feedback rate	Interviews
	Efficiency	Task completion time	(Holmes et al., 2019; Peras, 2018), Interviews
		Number of turns	
		Human handover rate	(Wintersberger et al., 2020), Interviews
Data privacy	Realization and communication	Privacy and anonymity	(Feng & Buxmann, 2020; Janssen et al., 2021; Lewandowski et al., 2021; Rajaobelina et al., 2021), Interviews
		Transparency	

be evaluated and integrated to facilitate dialog flow. For example, it may be helpful to allow users to interact with CA responses via buttons (Kowald & Bruns, 2020). The interviews emphasized the need to continuously evaluate and refine the selection and functionality of control elements (e.g., text, buttons, reactions, and carousel selections). In addition, the *context awareness* of CAs should be evaluated. The ability to grasp *dialog-oriented contexts* allows CAs to incorporate previous user utterances to conduct sophisticated conversations with users. These conversations should be evaluated to ensure that users do not have to enter input repetitively (Saenz et al., 2017). Connected to this, resumption and return points in the dialog tree are fundamental aspects of evaluation. A well-structured dialog flow helps users provide the correct input, achieve their goals, and avoid deadlocks (Diederich et al., 2020). Moreover, the *technical environment* needs to be established to enable unrestricted usage, especially in complex use cases. From the first to the last user touchpoint, background systems should be

conveniently accessed (e.g., single sign-on) to address background systems that resolve requests and provide information.

Output

Output refers to criteria related to the CA-generated response provided in return to the user request. Regarding output, the *format* of the CA responses should be reflected. The responses require an appropriate selection of suitable *visual elements* in terms of a user- and content-oriented presentation (e.g., with texts, images, and tiles), as well as high *readability* (Kowald & Bruns, 2020). Especially in the context of CAs, *consistency* in language and terminology is important for avoiding complexity and confusion for users (Edirisooriya et al., 2019). In terms of *content*, the CA should transparently disclose its *capabilities* and *limitations* to evoke appropriate user expectations that are consistent with the nature of the CA as a learning IS (Diederich et al., 2020). Furthermore, CA answers should be reviewed to evaluate whether users' (information) needs have been fulfilled. The relevance and meaningfulness of the presented information and the up-to-dateness of the knowledge base for *information retrieval* should be checked to determine whether background knowledge must be updated or expanded (Diederich et al., 2020). Apart from recognizing the user's intent and presenting the correct output, Feng and Buxmann (2020) emphasized the evaluation of different representations and levels of *detail of the knowledge*. Especially for more complex CAs (e.g., those that combine numerous background systems as a central platform), it is challenging to present the often complex solutions in an abstract and *convergent* way that provides users with appropriate answers to their concerns. The interview experts highlighted that solutions sorted by the relevance and *justification* of the CAs' answers could increase user trust in these outputs. For example, a CA could refer to the background system or source to make it transparent from where the knowledge was obtained (e.g., clickable link below the answer). Closely related, the CAs' *calibration* of *response appropriateness* should be evaluated to provide concise and manageable CA answers. In this context, CAs' *response accuracy* (e.g., also referred to as response quality (Jiang & Ahuja, 2020)) needs to be evaluated to present knowledge correctly (e.g., length, tonality, fluency) to the target audience. Regarding the *timing of responses*, *technical response time* is considered a relevant factor for CAs. For example, Edirisooriya et al. (2019) identified quick responses—within two to five seconds of the user's request—as essential. However, the criterion *balance between proactivity and interruption*, which refers to the fact that CAs' proactive utterances may interrupt users, indicates that this behavior and its effects on users should be evaluated.

Anthropomorphism

Anthropomorphism relates to human characteristics, such as emotions, applied to nonhuman objects (Schuetzler et al., 2021). Anthropomorphism can positively affect the use of CAs and can be divided into three aspects: humanlike identity, verbal cues, and nonverbal cues (Seeger et al., 2021). First, evaluable criteria in the context of *humanlike identity* represent aspects that strengthen CA *identity* (e.g., profile pictures or avatars) and other *characteristics*, such as demographic information, including gender, age, or name (Seeger et al., 2021). In addition, the general *visual representation* was highlighted during several interviews. A CA team should reflect on how the CA can be easily detected as the first contact point with the user, including, for example, its integration into a website, such as its position, size, responsive (humanlike) appearance, and colors. Furthermore, CAs' *verbal cues* should be reviewed. Besides the ability to engage in social dialogs, called "*chitchat*," *emotional expressions* (e.g., apologizing by the CA), verbal style, and self-reference (e.g., the CA referring to itself as "I" or "me"), or context-sensitive responses, *tailored personality and lexical alignment* (e.g., by the CA adapting its responses to the users' utterances (Saenz et al., 2017)) can also be used to make CAs seem more humanlike (Schuetzler et al., 2021; Seeger et al., 2021). In particular, chitchat and character definition were emphasized in the interviews, since many users first check the CA for its social capabilities and quickly lose interest if it fails, even at slight initial social interactions. Further possibilities of humanlike design are *nonverbal cues*, such as *emoticons*, or artificially induced *typing delays and indicators*, such as typing dots (Gnewuch et al., 2018). However, researchers have also noted that a humanlike CA can be repellent to users (e.g., Grudin & Jacques, 2019). Seeger et al. (2021) indicated that the different anthropomorphism criteria must be combined and evaluated practically.

Dialog control

For successful *dialog control*, CAs' understanding of users' requests, along with their intentions and goals, should be evaluated (Clark et al., 2019). However, CAs are learning IS and, therefore, initially error-prone. In particular, user input in lengthy and complex sentences poses a challenge for CAs (Michaud, 2018). Thus, proactive dialog handling in regular operations and reactive handling in failure operations should be evaluated to ensure that CAs avoid, reduce, or recover from failures. In *regular operations*, organizations should continuously reflect on whether the CA proactively avoids error scenarios by,

for example, asking the user *to reformulate the request* (Diederich et al., 2020) or prompting the user for more information (Chaves & Gerosa, 2021). Further, the interviews revealed the expectation that if no appropriate answer was elicited, the CA should proactively refer to misunderstandings or reintroduce his skills. Afterward, the CA could provide *alternative responses* to keep the conversation alive (Chaves & Gerosa, 2021). Another way is to provide *conversational prompts*. Through the use of prompts, the CA provides *suggestions* for prospective requests in addition to their responses (e.g., in the case of a long response time by the user). The aim is to predict the user's intentions (e.g., by offering suggestions on text buttons) and proactively avoid error cases when processing a user's text input (Li et al., 2020). In *failure operations*, it is crucial to define and evaluate (e.g., *proactive and resilient) repair strategies* to overcome conversational breakdowns, since their existence can result in a negative experience for users and impair future CA success (Benner et al., 2021). In the case of a breakdown, the CA should fail gracefully in order to maintain user trust (Feng & Buxmann, 2020). For instance, the CA can apologize and propose new solutions (Benner et al., 2021). However, if repair attempts fail repeatedly and the CA's capabilities are exceeded, the CA should encourage *fallbacks or a handover* to a service representative (Poser et al., 2021, 2022).

Performance

A holistic evaluation of CA *performance* represents a strong predictor of CA success (Peras, 2018). By combining design- and technically-oriented principles, CAs' performance relates directly to user satisfaction (Liao et al., 2016). The performance demonstrates the effective and efficient completion of tasks between the user and the CA (Peras, 2018). Regarding CAs' *effectiveness*, the *task success rate* and the *task failure rate* could be used to collect the number of successful tasks and the number of default fallback intents to trigger appropriate countermeasures (Peras, 2018). In the interviews, the *retention and feedback* rates were mentioned regarding the recordings of returning users and continuously evaluating users' average ratings to uncover weaknesses and derive improvement potential. Furthermore, it is necessary to consider CAs' *efficiency* because the adequate performance of tasks explicates only a few insights into whether the CA also performs the tasks with a resource-friendly approach. Given this perspective, evaluating the time required to complete a task (*task completion time*) and the (average) number of rounds of dialog required (average *number of turns*) is essential to capture efficiency (Holmes et al., 2019; Peras, 2018). In addition, the *human handover rate* is significant in evaluating at which points the CA cannot complete a task (Wintersberger et al., 2020).

Data privacy

Data privacy includes criteria related to the *realization and communication of data protection* endeavors. One important aspect is ensuring that conversations with the CA are kept as *private* and *anonymous* as possible, particularly when the CA deals with confidential or personal data (Feng & Buxmann, 2020). During the interviews, we received feedback emphasizing the importance of minimizing the storage of conversational data and ensuring that any stored data is anonymized to the greatest extent feasible, especially when such data is necessary to improve a CAs' performance. The communication of data protection contains the criterion of *transparency* toward users, meaning the disclosure of which user data is processed. In this context, it is helpful to provide data protection policies (Rajaobelina et al., 2021).

Case-based instantiation of the quality criteria set

After the research activities of the DSR project in *Steps 1* to *5*, the final quality criteria set was instantiated in *Step 6* in an IT organization to investigate, evaluate, and improve the quality of an existing AI- and text-based CA. Due to the organization's limited in-depth knowledge of a systematic CA evaluation procedure, including methods, a systemized procedure model was initiated and documented. It comprises three main phases and was applied to utilize the final CA quality criteria set throughout each phase (see Fig. 3).

Case setting for applying the procedure model

The DSR project considered the following case setting to apply the procedure model, evaluate CAs' quality, and address an existing real-world problem: (1) The procedure model requires a suitable use case to evaluate the applicability and feasibility to indicate a CAs' quality. To this end, an existing AI- and text-based CA (*ExpertBot*) was investigated, evaluated, and improved in an IT organization. Based on our interview analysis (as outlined in *Step 1* of our DSR project), *ExpertBot* was deemed to be a suitable case for a root cause analysis, since the overall quality and usage rate were insufficient. The IT organization uses *ExpertBot* within organizational boundaries to identify, prioritize, and select needed experts. Therefore, *ExpertBot* participates in chat conversations and accesses various data sources, such as skill databases, document management systems, and internal chat forums, to provide fitting recommendations

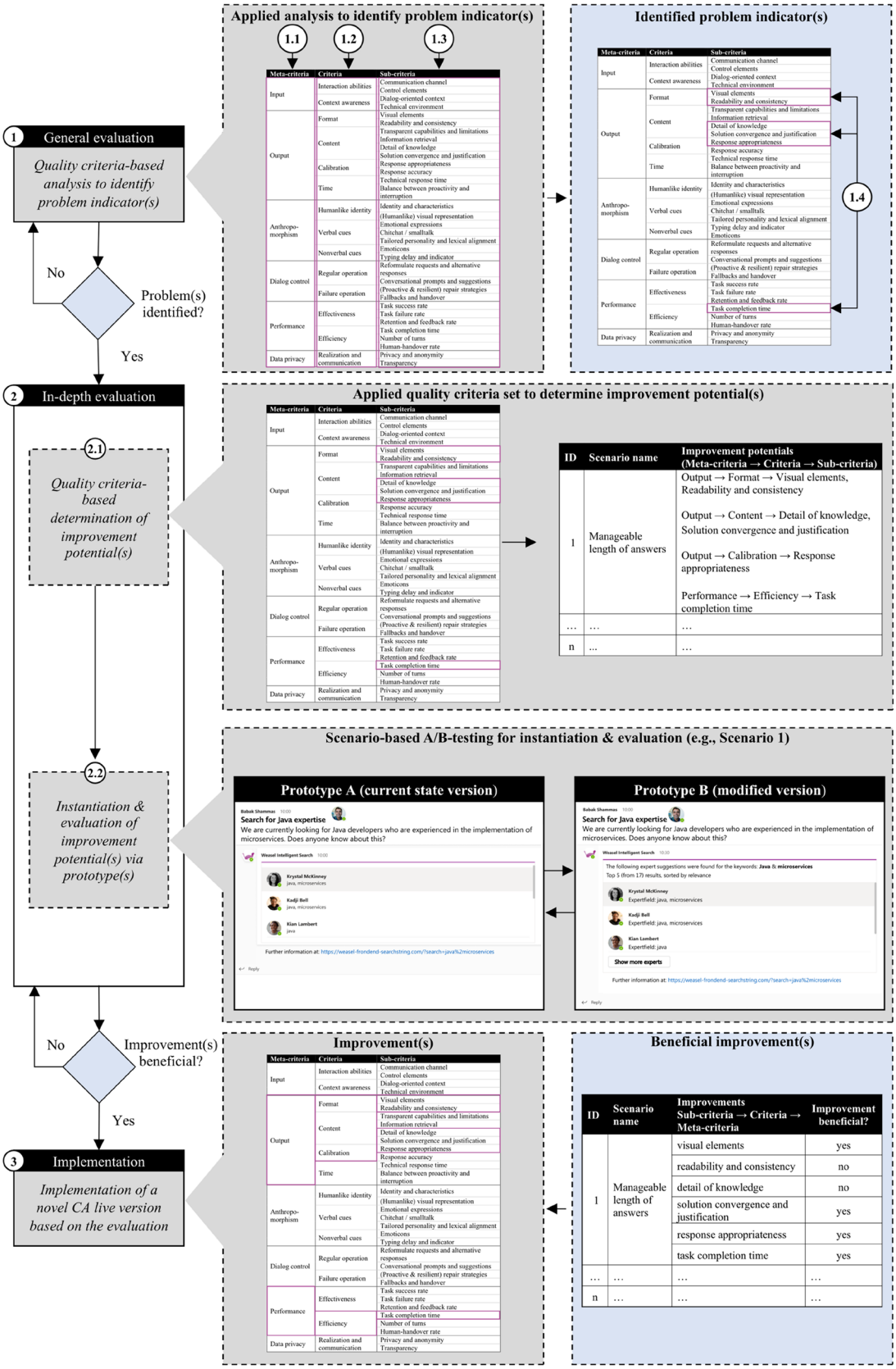

Fig. 3 Procedure model for the evaluation and improvement of CAs using our quality criteria set

for experts and their skills. *ExpertBot* is integrated into an existing text-based communication channel in Microsoft Teams and works intent-based, using Microsoft Language Understanding (LUIS) and Azure Cognitive Services in the background. (2) Furthermore, forming an expert team with varying experience levels and backgrounds regarding CAs and their application field is crucial to provide a multi-perspective view enabling a broad discussion of the quality criteria and shortcomings of CAs. In our case, to implement the procedure model for evaluating the quality of *ExpertBot* through all phases, the existing CA team has formed an interdisciplinary team of experts from the IT organization (e.g., CA developers, product owners, management assistance responsible for staffing, and employees from other departments, as outlined in *Step 6* of our DSR project). (3) Finally, an expert team requires an appropriate data basis to evaluate CAs. For this purpose, prepared data, such as the user retention rate or other criteria, can be applied. In this context, the newly formed expert team evaluated *ExpertBot* based on our quality criteria set and derived improvement potentials. Overall, this case setting served as the starting point for instantiating the CA quality criteria set through the procedure model, as shown in Fig. 3.

Utilization of the procedure model

Our procedure model is designed with three main phases and several sub-phases to provide a fine-grained approach that fosters comprehensiveness and traceability. The sub-phases enable us to (1) create progressive guidance for each phase of the procedure model, facilitating the evaluation of CAs; (2) ensure that every aspect of the procedure is thoroughly documented, which is crucial for properly evaluating CAs; and (3) create a more detailed and extensive procedure that helps the expert team to ensure a systematic CA evaluation.

Phase 1: General evaluation

In Phase 1 (general evaluation), we performed a quality criteria-based analysis to identify problems with the current CA version. More specifically, in Sub-phase 1.1, the derived meta-criteria were used to provide a starting point for the CA team's initial evaluation of the *ExpertBot* and to identify possible problem areas (see Fig. 3). In Sub-phase 1.2, the corresponding criteria of these problem areas served as a more detailed level to narrow the scope of analysis. Thereby, in Sub-phase 1.3, the sub-criteria belonging to the criteria could be used as indicators of potential problems. Based on these phases and the analysis of appropriate data related to the corresponding sub-criteria, specific problem indicators of the *ExpertBot* were identified in Sub-phase 1.4.

In our illustrated example from our instantiation (see Fig. 3), the general evaluation revealed that the overarching meta-criteria "output" and "performance" of the *ExpertBot* needed to be improved. Six problem indicators, such as "detail of knowledge," "solution convergence and justification," and "task completion time," were considered throughout the criteria-based analysis to start an in-depth evaluation. As a result, we initiated an improvement project to address the identified indicators.

Phase 2: In-depth evaluation

As part of Phase 2 (in-depth evaluation), we first conducted Sub-phase 2.1. In cooperation with the IT organization, the CAs' quality was evaluated, and the potentials for improvement were determined based on the identified problem indicators from Sub-phase 1.4. Using appropriate evaluation methods, the consideration of these improvement potentials was found to be beneficial for the expert team (comprising members from the CA team; see "case setting").

To gain this insight, we conducted seven semi-structured interviews with the expert team members. We presented the live version of the *ExpertBot* and asked the participants about the general implementation, problems, and relevance for improvement, along with the corresponding criteria from our set. The resulting evaluated improvement potentials of the *ExpertBot* were then transformed into coherent scenarios in an aggregation process. Thereby, a collective evaluation of multiple quality criteria in each scenario could be conducted. In the single scenario we outlined, as shown in Sub-phase 2.1 of Fig. 3, all six specific problem indicators were identified as improvement potentials during the interviews. Specifically, the scenario was called "manageable length of answers" and included the improvement potentials "visual elements," "readability and consistency," "detail of knowledge," "solution convergence and justification," "response appropriateness," and "task completion time."

In Sub-phase 2.2, we created mockup prototypes for the transformed scenarios to demonstrate, investigate, and evaluate the identified improvement potentials. In this context, the prototypes enabled a well-founded comparison between the current state version of the CA and the proposed modified CA version(s). The expert team provided valuable feedback to verify whether the identified improvement potentials would be beneficial if implemented or needed to be revised or discarded.

For the creation of prototypes, we employed the Figma (2022) design tool in combination with the Microsoft Teams UI Kit (2023) to ensure a familiar and consistent visual representation during the demonstration. Furthermore, the prototypes were designed based on the previously evaluated improvement potentials corresponding to the analyzed *ExpertBot.* Subsequently, we conducted A/B tests involving six participants by presenting them with two prototypes for each scenario during semi-structured interviews to achieve

a data basis for deciding whether to implement the proposed changes. One prototype contained the current CA version, while the other represented the assumed improvements (modified version, as depicted in Fig. 3). For each scenario, questions were asked in three areas during the interviews. First, we asked participants to evaluate which of the two prototypes was more effective at first glance and which aspects were crucial to this impression (e.g., perceptions of the prototype features and differences). Second, the improvement potentials were addressed individually, and the participants were asked to determine which sub-criteria were conceivable for increasing CA quality. Third, we asked which of the addressed sub-criteria was rated the most important in improving CA quality to prioritize the highest-ranked improvement potentials (e.g., number of mentions) in preparation for the last phase.

Phase 3: Implementation

Finally, in Phase 3 (implementation), the improvement potentials, identified in Sub-phase 2.1 and evaluated as beneficial in Sub-phase 2.2 for increasing the quality of the CA, were implemented in a revised CA live version. These improvements were communicated to the users to ensure their visibility in the organization. After Phase 3, the procedure should be repeated to improve the CA on a long-term basis, for instance, if problems are identified based on existing data, or as part of a general cyclical evaluation to examine the quality of the new CA live version as a whole or in defined segments, which, however, was not part of the instantiation.

Discussion

Organizations strive to implement CAs due to their potential to increase business value with their ability to assist or automate processes, tasks, and activities (Lewandowski et al., 2021). However, despite their strengths in improving organizational efficiency (Zierau et al., 2020c), many CAs across industries are still error-prone and fail during interactions (Gnewuch et al., 2017), leading to a high discontinuation rate (Janssen et al., 2021). To strengthen the management of CAs in an organizational context and improve their success, we have developed a quality criteria set and procedure model for conducting holistic evaluations and improvements of CAs. In a multi-step DSR project, criteria were identified, aggregated, and evaluated ex-ante for applicability and operationality in real-world environments. In addition, a procedure model for the application of the quality criteria set was determined as part of a naturalistic ex-post evaluation. The conducted evaluation activities demonstrate that the incorporated criteria provide an integrated view of a CA evaluation. Regarding the procedure model, the results indicate that a systematic analysis of the problems, requirements, and status quo of a specific CA is supported to identify and improve its most relevant aspects. In combination, these findings have implications for research and practice.

Theoretical implications

First, our quality criteria set and procedure model contribute to CA research by providing a synthesized and systematized approach to improving the success of contemporary CAs. To achieve this, we contribute the quality criteria set derived from strongly dispersed CA research streams (Følstad et al., 2021). A large share of this research has focused on specific design and technical issues to elevate the user experience (e.g., Seeger et al., 2021), as these issues were considered the main challenges in the implementation of CAs (Følstad et al., 2018a; Janssen et al., 2021; van der Goot et al., 2021). However, CAs are inherently complex IS (Maroengsit et al., 2019) with distinctive characteristics that require a comprehensive view and analysis, as failures can arise from multiple (interrelated) factors (Janssen et al., 2021; Meyer von Wolff et al., 2021). Therefore, we extend the focus of current CA research to a consolidated set of essential quality criteria that should be considered to support the prevention of CA failure. We also provide a starting point for a more structured CA evaluation with our procedure model, as recommended by Følstad et al. (2021). The quality criteria set addresses the type of AI and text-based CAs in general domains, as classified by Gnewuch et al. (2017). Nevertheless, the results may also apply to other types of CAs. Additionally, our work complements other preliminary efforts, such as the evaluation criteria sets of Radziwill and Benton (2017) and Casas et al. (2020), to provide a better understanding of CAs in the improvement process with a system-wide view.

Recent technical advancements in the field of NLP and ML applications should also be highlighted, especially the emergence of large language models (LLMs). These models are pre-trained on billions of text samples from specific data sources on the Internet and can generate diverse types of content (Brown et al., 2020; Jiang et al., 2022). In particular, these models become widely available to (non-technical) users via the release of intuitive and conversational interfaces, such as OpenAI's ChatGPT or Google's Bard (Jiang et al., 2022; Teubner et al., 2023). These releases implicate a remarkable movement in CA research exploring new application scenarios and their potential, which is also referred to as a new "AI wave" by Schöbel et al. (2023). Consequently, the question arises of which quality criteria are affected in this new wave and what requirements result for CAs and their (further) development. We, therefore, expect a

paradigm shift in the perception and utilization of the different criteria from our set in regard to the novel LLM applications, which could lead to high dynamics and flexibility in their adaptation and use.

Second, this article contributes to management research on CAs, which encompasses various aspects of the CA lifecycle, such as critical phases, factors, or tasks within CA development (Lewandowski et al., 2022b; Meyer von Wolff et al., 2022). However, these initial studies neither provide deeper insights into CAs' evaluation and improvement nor explain how a cyclical evaluation process can be executed. In this regard, our research provides an approach for evaluating and improving CAs, which can serve as a meta-model for other researchers using different qualitative and quantitative methods within the lifecycle of a CA. While researchers focus on the design of CAs by targeting specific aspects, such as increasing user trust or anthropomorphism (e.g., Seeger et al., 2017), they often disregard the importance of evaluating CAs on an ongoing basis, as elaborated in the "Related research" section. Thereby, we provide knowledge regarding a structured and continuous CA evaluation to ensure the improvement of CAs during their operation in organizations (Janssen et al., 2021; Meyer von Wolff et al., 2021). In addition, the quality criteria set and procedure can assist in other lifecycle phases, for instance, by providing an overview of initial design issues in the initiation phase. Furthermore, the quality criteria set can support a multi-perspective and comprehensive development process and the detection of problems before going live in the integration phase to avoid direct failure. Our article aggregates design knowledge, supplemented by practical insights, and introduces a structured approach that provides initial insights into activities, people, and data, which can foster operations and enhance the performance of CAs.

Third, from the DSR lens, we contribute prescriptive design knowledge with the quality criteria set and procedure model for their application. Both form our developed artifact. This artifact provides a foundation that can be applied in the identified higher-level problem class in other solution spaces (Hevner et al., 2004). The application of the artifact can be utilized to address and explore the problem class in more depth and further improve the ability to apply it in a generalized manner or to design more sophisticated artifacts as tools for similar problems.

Practical implications

In addition to its theoretical contributions, our article has practical implications for organizations. By providing a systematic approach to evaluation and improvement, our artifact can guide CA teams in various ways to support the successful development, operation, and evolution of CAs. First, the combination of the quality criteria set and procedure model allows practitioners to obtain a comprehensive overview of relevant criteria and to narrow down the evaluation of their CA to identify specific problems and improve the overall quality of CAs. Second, the procedure model can serve as a blueprint for CA teams to systematize the evaluation process. The delineation of content and the sequence of relevant steps provides a feasible approach for practitioners to structure their evaluation and improvement activities of existing CAs. In addition, the criteria set can serve as a basis for CA teams to decide whether a CA project should be established and whether requirements are present (e.g., prepared data, an interdisciplinary team) to enable a comprehensive and multi-perspective evaluation of the quality of CAs. Thereby, the execution of evaluation and improvement tasks could be accelerated. Apart from the description of relevant criteria and the evaluation steps, the artifact's application may positively affect organizations. For instance, following the systematized procedure to improve CAs, the perceived user satisfaction could increase, thus resulting in an improved acceptance and usage rate and consequently counteracting the discontinuation rate of CAs. In addition, the evolution of CAs and related positive effects could not be limited to the CA domain, as their success could foster the overall AI transformation of an organization, so that the increased quality and use of CAs can influence other learning and AI-based IS.

Limitations and future research

Our research is not without limitations that have implications for further research. The developed artifact comprises a comprehensive set of quality criteria. However, its application does not inevitably guarantee success in the deployment and continuous improvement of CAs. To achieve this broad goal, additional aspects, such as technical requirements (e.g., AI, ML, and NLP algorithms and tools), a fit of the technology to the use case (e.g., using a CA for complex tasks or in emotionally-sensitive environments), design (e.g., human–computer interface), and organizational communication to users (e.g., tutorials, highlighting benefits and restrictions) have to be considered. All factors in interaction lay the foundation for a successful CA operation. In pursuit of this goal, the quality criteria set and procedure model can be considered one piece of the greater puzzle.

The instantiation of the quality criteria set revealed several challenges and aspects that need further research. First, in Phase 1 of our instantiation, we determined the need for a quality criteria-based, in-depth evaluation of CAs' output and performance. These overarching meta-criteria proved to be valid starting points for exploring the improvement potentials of the *ExpertBot*. Nevertheless, further investigation is required to identify additional triggers that warrant in-depth evaluation. Broadening the perspective, triggers from

outside the organizational boundaries, such as feedback from customers, are possible. However, we did not identify any of these triggers in our project due to the inward-facing use case of the *ExpertBot*.

Second, further research on how organizations can generate a CA evaluation strategy, including aspects such as evaluation intervals, criteria selection, and suitable evaluation methods, is needed. In our real-world instantiation, we have resorted to semi-structured interviews and A/B testing as qualitative evaluation methods, which are not necessarily suitable for all criteria. Overall, a general framework could assist organizations and researchers in the selection of suitable evaluation and data analysis methods for relevant areas (such as our meta-criteria). In particular, longitudinal studies that explore application of the quality criteria set and procedure model in real-world environments can provide deeper insights into evaluation strategies and the impact of their use.

Third, we observed that different quality criteria of our set have varying levels of impact on CAs' quality. The criteria exhibit an indeterminate degree of interdependence as they influence each other. In addition, we found that the skills of the participants (e.g., CA team) can influence this factor. We derive three directions for further research: (1) conception, design, and evaluation of a modular, context-adaptive procedure model that can be tailored to arbitrary CA application environments and their essential quality criteria, including the individual conditions; (2) investigations to determine the needs of AI and data literacy experts for designing, continuously evaluating, and improving CAs, as well as the needs of (non-expert) users for utilizing and validating the information output to counteract their failure; (3) identification of criteria in our set that may need to be evaluated more or less frequently. A combined classification or ranking of the influence and importance of the criteria (e.g., by an empirical research approach) offers additional potential.

Moreover, we expect further technical progress and research in the context of customizable CAs, as also described in a study by Schöbel et al. (2023). Aspects such as social presence and anthropomorphism, as well as personalization and empathy of human-AI interactions, are to be considered. Especially the new wave of AI technologies and LLM could lead to improvements in terms of better customization and contextualization. By exploring these areas, further research could counteract the skepticism of users toward conventional CAs, perceiving them as unnatural, impersonal, or deceptive (Schöbel et al., 2023), and reduce the overall failure of CAs (Gnewuch et al., 2017). In this regard, our set of criteria contains anthropomorphism criteria, such as identity, visual representation, and tailored personalization. However, these criteria need to be further explored in the wake of the recent customization and contextualization capabilities of LLMs that could make CAs more adaptive to users' emotional states, for example, by tailoring responses to individual needs and preferences, fostering a wider acceptance in the future.

In addition to information retrieval scenarios, CAs augmented with LLM capabilities could act in a broader spectrum of possible use cases. In conjunction with our demonstrated procedure model, new approaches for evaluating and improving CAs extended by LLMs may become mandatory. For example, generative activities (e.g., content created based on statistical methods and available data) should be handled differently from information retrieval activities (e.g., content extracted unchanged from a connected data source). The question whether the generated content corresponds to the truth or contains false and misleading facts arises. The range of application scenarios in practice and the exploration of these technologies are in their infancy and will be an engaging field of research (Schöbel et al., 2023).

While our article provides insights and an approach to the evaluation and improvement of CAs, it has methodological limitations. Although our artifact proved to be applicable by instantiating it in an IT organization, its transferability to other application environments with CAs of different use cases, other CA teams, and conditions remain to be proven to further address the overarching problem class. In this vein, our set of quality criteria could be a building block for adaptation. Overall, several foundations are laid for research on the design, validation, and adaptation of the quality criteria set and procedure model.

Conclusion

CAs have become increasingly relevant in facilitating convenient access to information and services, representing essential gateways for organizations to interact with customers or employees (Følstad et al., 2021). However, due to their frequent premature deployment and varying maturity levels, CAs can be error-prone and fail to meet the requirements of their intended use cases, ultimately leading to their abandonment. To address this challenge, we conducted a DSR project that demonstrates how organizations can leverage a systematized procedure model based on criteria-based analysis to foster continuous evaluation and improvement of CAs. Our article provides guidance for organizations to better understand and evaluate the quality of their CAs, thereby laying the foundation for their long-term success. As a result, this article expands the knowledge base on CAs and emphasizes that evaluating and improving them is an ongoing challenge due to their complex and unique nature. Additional research is

needed to further explore how organizations can conduct criteria-based evaluations of their CAs and develop effective evaluation strategies.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Alonso, A. F., Fuertes Castro, J. L., Martínez Normand, L., & Soza, H. (2009). Towards a set of measures for evaluating software agent autonomy. *Mexican International Conference on Artificial Intelligence (MICAI)*, Guanajuato, México. https://doi.org/10.1109/MICAI.2009.15

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. *Conference on human factors in computing systems (CHI)*, Glasgow, Scotland, UK. https://doi.org/10.1145/3290605.3300484

Benner, D., Elshan, E., Schöbel, S., & Janson, A. (2021). What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents. *International Conference on Information Systems (ICIS).* Austin, TX, United States

Bisio, I., Garibotto, C., Grattarola, A., Lavagetto, F., & Sciarrone, A. (2018). Smart and robust speaker recognition for context-aware in-vehicle applications. *IEEE Transactions on Vehicular Technology, 67*(9), 8808–8821. https://doi.org/10.1109/TVT.2018.2849577

Bittner, E. A. C., Oeste-Reiß, S., & Leimeister, J. M. (2019). Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, United States.

Brandtzaeg, P. D., & Følstad, A. (2017). Why people use chatbots. *International Conference on Internet Science (INSCI).* Cham

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations. *Interactions, 25*(5), 38–43.

Brink, A. (2013). *Anfertigung wissenschaftlicher Arbeiten: Ein prozessorientierter Leitfaden zur Erstellung von Bachelor-, Master-und Diplomarbeiten* (5th ed.). Springer Gabler.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *European Conference on Information Systems (ECIS)*, Verona, Italy.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Burbach, L., Halbach, P., Plettenberg, N., Nakayama, J., Ziefle, M., & Valdez, A. C. (2019). "Hey, Siri","Ok, Google","Alexa". Acceptance-relevant factors of virtual voice-assistants. *International Conference on Professional Communication (IPCC)*, Aachen, Germany. https://doi.org/10.1109/ProComm.2019.00025

Casas, J., Tricot, M.-O., Abou Khaled, O., Mugellini, E., & Cudré-Mauroux, P. (2020). Trends & methods in chatbot evaluation. *International Conference on Multimodal Interaction (ICMI)*, Virtual event, Netherlands. https://doi.org/10.1145/3395035.3425319

Chaves, A. P., & Gerosa, M. A. (2021). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human-Computer Interaction, 37*(8), 729–758. https://doi.org/10.1080/10447318.2020.1841438

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., & Munteanu, C. (2019). What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems (CHI)*, New York, NY, United States. https://doi.org/10.1145/3290605.3300705

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104–126.

Corea, C., Delfmann, P., & Nagel, S. (2020). Towards intelligent chatbots for customer care-practice-based requirements for a research agenda. *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, United States.

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017). SuperAgent: A customer service chatbot for e-commerce websites. *Meeting of the Association for Computational Linguistics-*System *Demonstrations*, Vancouver, Canada. https://doi.org/10.18653/v1/P17-4017

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817. https://doi.org/10.1017/S1351324916000243

Davenport, T. H., & Kirby, J. (2016). Just how smart are smart machines? *MIT Sloan Management Review, 57*(3), 21–25.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering, 62*(3), 193–209. https://doi.org/10.1007/s12599-020-00639-y

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

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019b). Towards a taxonomy of platforms for conversational agent design. *International Conference on Wirtschaftsinformatik (WI)*, Siegen, Germany.

Drechsler, A., & Hevner, A. R. (2018). Utilizing, producing, and contributing design knowledge in DSR projects. *International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, Chennai, India. https://doi.org/10.1007/978-3-319-91800-6_6

Edirisooriya, M., Mahakalanda, I., & Yapa, T. (2019). Generalized framework for automated conversational agent design via QFD. *Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka. https://doi.org/10.1109/MERCon.2019.8818945

Elshan, E., Engel, C., Ebel, P., & Siemon, D. (2022a). Assessing the reusability of design principles in the realm of conversational agents. *International Conference on Design Science Research in Information Systems and Technology (DESRIST)*, St. Petersburg, FL, USA. https://doi.org/10.1007/978-3-031-06516-3_10

Elshan, E., Siemon, D., De Vreede, T., De Vreede, G.-J., Oeste-Reiß, S., & Ebel, P. (2022b). Requirements for AI-based teammates: A qualitative inquiry in the context of creative workshops. *Hawaii Conference on System Sciences (HICSS)*, Hawaii, HI, USA.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019a). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies, 132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Feine, J., Morana, S., & Maedche, A. (2019b). Designing a chatbot social cue configuration system. *International Conference on Information Systems (ICIS)*, Munich, Germany.

Felderer, M., & Ramler, R. (2021). Quality assurance for AI-based systems: Overview and challenges (Introduction to interactive session). *Software Quality: Future Perspectives on* Software *Engineering Quality (SWQD)*, Springer, Cham. https://doi.org/10.1007/978-3-030-65854-0_3

Feng, S., & Buxmann, P. (2020). My virtual colleague: A state-of-the-art analysis of conversational agents for the workplace. *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, United States.

Figma. (2022). Retrieved May 31, 2022, from https://www.figma.com

Følstad, A., Araujo, T., Law, E.L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., & Ischen, C. (2021). Future directions for chatbot research: An interdisciplinary research agenda. *Computing, 103*(12), 2915–2942. https://doi.org/10.1007/s00607-021-01016-7

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018a). What makes users trust a chatbot for customer service? An exploratory interview study. *International Conference on Internet Science (INSCI)*, Springer, Cham. https://doi.org/10.1007/978-3-030-01437-7_16

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2018b). Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. *International Conference on Internet Science (INSCI)*, Springer, Cham. https://doi.org/10.1007/978-3-030-17705-8_13

Gläser, J., & Laudel, G. (2009). *Experteninterviews und qualitative Inhaltsanalyse: Als Instrumente rekonstruierender Untersuchungen* (3rd ed.). VS Verlag für Sozialwissenschaften.

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

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. *European Conference on Information Systems (ECIS)*, Portsmouth, United Kingdom.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Goh, O. S., Ardil, C., Wong, W., & Fung, C. C. (2007). A black-box approach for response quality evaluation of conversational agent systems. *International Journal of Computational Intelligence, 3*(3), 195–203.

van der Goot, M. J., Hafkamp, L., & Dankfort, Z. (2021). Customer service chatbots: A qualitative interview study into customers' communication journey. *International Workshop on Chatbot Research, CONVERSATIONS 2020*, Amsterdam, the Netherlands. https://doi.org/10.1007/978-3-030-68288-0_13

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *Management Information Systems Quarterly (MISQ), 37*(2), 337–355. https://doi.org/10.25300/misq/2013/37.2.01

Gregory, R. W., & Muntermann, J. (2014). Research note—Heuristic theorizing: Proactively generating design theories. *Information Systems Research, 25*(3), 639–653. https://doi.org/10.1287/isre.2014.0533

Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. *Conference on Human Factors in Computing Systems (CHI)*, Glasgow, Scotland. https://doi.org/10.1145/3290605.3300439

Herrera, A., Yaguachi, L., & Piedra, N. (2019). Building conversational interface for customer support applied to open campus an open online course provider. *International Conference on Advanced Learning Technologies (ICALT)*, Maceio, Brazil. https://doi.org/10.1109/ICALT.2019.00011

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly (MISQ), 28*(1), 75–105.

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *European Conference on Cognitive Ergonomics (ECCE)*, Belfast, UK. https://doi.org/10.1145/3335082.3335094

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., & Akkiraju, R. (2018). Touch your heart: A tone-aware chatbot for customer care on social media. *Conference on Human Factors in Computing Systems (CHI)*, Montréal, Canada. http://dx.doi.org/https://doi.org/10.1145/3173574.3173989

Huiyang, S., & Min, W. (2022). Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. *International Journal of Human-Computer Interaction, 38*(1), 28–41.

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., & Smit, E. (2020). Privacy concerns in chatbot interactions. Chatbot Research and Design. CONVERSATIONS. (2019). Amsterdam. *The Netherlands*. https://doi.org/10.1007/978-3-030-39540-7_3

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. *Designing Interactive Systems Conference (DIS)*, New York, NY, USA. https://doi.org/10.1145/3196709.3196735

Janssen, A., Passlick, J., Cardona, D. R., & Breitner, M. H. (2020). Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering, 62*(3), 211–225. https://doi.org/10.1007/s12599-020-00644-1

Janssen, A., Grützner, L., & Breitner, M. H. (2021). Why do chatbots fail? A critical success factors analysis. *International Conference on Information Systems (ICIS)*, Austin, TX, United States.

Jiang, J., & Ahuja, N. (2020). Response quality in human-chatbot collaborative systems. *Conference on Research and Development in Information Retrieval (SIGIR)*, Virtual Event, China. https://doi.org/10.1145/3397271.3401234

Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J. (2022). PromptMaker: Prompt-based prototyping with large language models. *Conference on Human Factors in Computing Systems (CHI)*, New Orleans, LA, United States. https://doi.org/10.1145/3491101.3503564

Johari, N. M., & Nohuddin, P. (2021). Quality attributes for a good chatbot: A literature review. *International Journal of Electrical Engineering and Technology (IJEET), 12*(7), 109–119. https://doi.org/10.34218/IJEET.12.7.2021.012

Jonke, A. W., & Volkwein, J. B. (2018). From tweet to chatbot–Content management as a core competency for the digital evolution. *Digital Marketplaces Unleashed*, 275–285. https://doi.org/10.1007/978-3-662-49275-8_28

Klaus, P., & Zaichkowsky, J. (2020). AI voice bots: A services marketing research agenda. *Journal of Services Marketing, 34*(3), 389–398. https://doi.org/10.1108/jsm-01-2019-0043

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots. *Conference on Designing Interactive Systems (DIS)*, Edinburgh, United Kingdom. https://doi.org/10.1145/3064663.3064672

Kowald, C., & Bruns, B. (2020). Chatbot Kim: A digital tutor on AI. How advanced dialog design creates better conversational learning experiences. *International Journal of Advanced Corporate Learning (iJAC), 13*(3), 26–34. https://doi.org/10.3991/ijac.v13i3.17017

Kuligowska, K. (2015). Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, *2*.

Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2019). Improving conversations: Lessons learnt from manual analysis of chatbot dialogues. *International Workshop on Chatbot Research and Design, Springer, Cham.* https://doi.org/10.1007/978-3-030-39540-7_13

Laban, G., & Araujo, T. (2020). The effect of personalization techniques in users' perceptions of conversational recommender systems. *ACM International Conference on Intelligent Virtual Agents (IVA)*, New York, NY, USA. https://doi.org/10.1145/3383652.3423890

Lee, S., & Choi, J. (2017). Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies, 103*, 95–105. https://doi.org/10.1016/j.ijhcs.2017.02.005

Lewandowski, T., Delling, J., Grotherr, C., & Böhmann, T. (2021). State-of-the-art analysis of adopting AI-based conversational agents in organizations: A systematic literature review. *Pacific Asia Conference on Information Systems (PACIS)*, Dubai, UAE.

Lewandowski, T., Grotherr, C., & Böhmann, T. (2022a). Managing artificial intelligence systems for value co-creation: The case of conversational agents and natural language assistants. In B. Edvardsson & B. Tronvoll (Eds.), *The Palgrave Handbook of Service Management* (pp. 945–966). Springer International Publishing. https://doi.org/10.1007/978-3-030-91828-6_45

Lewandowski, T., Heuer, M., Vogel, P., & Böhmann, T. (2022b). Design knowledge for the lifecycle management of conversational agents. *International Conference on Wirtschaftsinformatik (WI)*, Nürnberg, Germany.

Li, C.-H., Yeh, S.-F., Chang, T.-J., Tsai, M.-H., Chen, K., & Chang, Y.-J. (2020). A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. *Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA. https://doi.org/10.1145/3313831.3376209

Li, M., & Suh, A. (2022). Anthropomorphism in AI-enabled technology: A literature review. *Electronic Markets, 32*, 2245–2275. https://doi.org/10.1007/s12525-022-00591-7

Liao, Q. V., Davis, M., Geyer, W., Muller, M., & Shami, N. S. (2016). What can you do? Studying social-agent orientation and agent proactive interactions with an agent for employees. *ACM Conference on Designing Interactive Systems (DIS)*, Brisbane, Australia. https://doi.org/10.1145/2901790.2901842

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants. *Business & Information Systems Engineering, 61*(4), 535–544. https://doi.org/10.1007/s12599-019-00600-8

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A survey on evaluation methods for chatbots. *International Conference on Information and Education Technology, Aizu-Wakamatsu, Japan.* https://doi.org/10.1145/3323771.3323824

McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface* (Vol. 6). Springer, Cham, https://doi.org/10.1007/978-3-319-32967-3

McTear, M. (2021). Introducing dialogue systems. In *Conversational AI. Synthesis Lectures on Human Language* Technologies (pp. 11–42). Springer, Cham. https://doi.org/10.1007/978-3-031-02176-3_1

Meira, M. O., & Canuto, A. M. P. (2015). Evaluation of emotional agents' architectures: An approach based on quality metrics and the influence of emotions on users. *World Congress on Engineering (WCE)*, London, U.K. https://www.iaeng.org/publication/WCE2015/WCE2015_pp143-150.pdf

Meuser, M., & Nagel, U. (2009). The expert interview and changes in knowledge production. *Interviewing experts*, 17–42.

Meyer von Wolff, R., Hobert, S., Masuch, K., & Schumann, M. (2020). Chatbots at digital workplaces–A grounded-theory approach for surveying application areas and objectives. *Pacific Asia Journal of the Association for Information Systems, 12*(2), 3. https://doi.org/10.17705/1pais.12203

Meyer von Wolff, R., Hobert, S., & Schumann, M. (2021). Sorry, I can't understand you! –Influencing factors and challenges of chatbots at digital workplaces. *International Conference on Wirtschaftsinformatik (WI)*, Essen, Germany.

Meyer von Wolff, R., Hobert, S., & Schumann, M. (2022). Chatbot Introduction and operation in enterprises–A design science research-based structured procedure model for chatbot projects. *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, United States.

Meyer-Waarden, L., Pavone, G., Poocharoentou, T., Prayatsup, P., Ratinaud, M., Tison, A., & Torné, S. (2020). How service quality influences customer acceptance and usage of chatbots. *Journal of Service Management Research, 4*(1), 35–51. https://doi.org/10.15358/2511-8676-2020-1-35

Michaud, L. N. (2018). Observations of a new chatbot: Drawing conclusions from early interactions with users. *IT Professional, 20*(5), 40–47. https://doi.org/10.1109/MITP.2018.053891336

Microsoft Teams UI Kit. (2023). Retrieved 28.03.2023 from https://www.figma.com/community/file/916836509871353159/Microsoft-Teams-UI-Kit

Neff, G., & Nagy, P. (2016). Automation, algorithms, and politics| Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication, 10*(17), 4915–4931.

Patel, S., Chiu, Y.-T., Khan, M. S., Bernard, J.-G., & Ekandjo, T. A. (2021). Conversational agents in organisations: Strategic applications and implementation considerations. *Journal of Global Information Management (JGIM), 29*(6), 1–25. https://doi.org/10.4018/JGIM.20211101.oa53

Pauletto, S., Balentine, B., Pidcock, C., Jones, K., Bottaci, L., Aretoulaki, M., Wells, J., Mundy, D. P., & Balentine, J. (2013). Exploring expressivity and emotion with artificial voice and speech technologies. *Logopedics Phoniatrics Vocology, 38*(3), 115–125. https://doi.org/10.3109/14015439.2013.810303

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, 89–97.

Pillai, R., & Sivathanu, B. (2020). Adoption of AI-based chatbots for hospitality and tourism. International *Journal of Contemporary Hospitality Management*, *32*(10). https://doi.org/10.1108/IJCHM-04-2020-0259

Poser, M., Singh, S., & Bittner, E. (2021). Hybrid service recovery: Design for seamless inquiry handovers between conversational agents and human service agents. *Hawaii International Conference on System Sciences (HICSS)*, Hawaii, United States.

Poser, M., Wiethof, C., & Bittner, E. A. (2022). Integration of AI into customer service: A taxonomy to inform design decisions. *European Conference on Information Systems (ECIS)*, Timisoara, Romania.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint*. https://doi.org/10.48550/arXiv.1704.04579

Rajaobelina, L., Prom Tep, S., Arcand, M., & Ricard, L. (2021). Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology & Marketing, 38*(12), 2339–2356. https://doi.org/10.1002/mar.21548

Research and Markets. (2022). *Chatbot market*. https://www.researchandmarkets.com/reports/5648691/chatbot-market-by-product-type-by-application?utm_source=BW&utm_medium=PressRelease&utm_code=f7b65b&utm_campaign=1751095+-+The+Worldwide+Chatbot+Industry+is+Expected+to+Reach+%2422.9+Billion+by+2030&utm_exec=jamu273prd#

Rimol, M. (2022). *Gartner predicts conversational AI will reduce contact center agent labor costs by $80 billion in 2026*. Gartner. Retrieved March 21, 2023 from https://www.gartner.com/en/newsroom/press-releases/2022-08-31-gartner-predicts-conversational-ai-will-reduce-contac

Riquel, J., Brendel, A. B., Hildebrandt, F., Greve, M., & Kolbe, L. M. (2021). "Even the wisest machine makes errors" – An experimental investigation of human-like designed and flawed conversational agents. *International Conference on Information Systems (ICIS)*, Austin, TX, United States.

Rowley, J., & Slack, F. (2004). Conducting a literature review. *Management Research News, 27*(6), 31–39. https://doi.org/10.1108/01409170410784185

Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and ethical considerations. *AICS*.

Saenz, J., Burgess, W., Gustitis, E., Mena, A., & Sasangohar, F. (2017). The usability analysis of chatbot technologies for internal personnel communications. *IIE Annual Conference*, Norcross.

Schmitt, A., Zierau, N., Janson, A., & Leimeister, J. M. (2021). Voice as a contemporary frontier of interaction design. *European Conference on Information Systems (ECIS)*, Virtual Conference.

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2023). Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers*, 1–26. https://doi.org/10.1007/s10796-023-10375-9

Schuetzler, R. M., Grimes, G. M., Giboney, J. S., & Rosser, H. K. (2021). Deciding whether and how to deploy chatbots. *MIS Quarterly Executive*, *20*(1), 4. https://doi.org/10.17705/2msqe.00039

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with human-like conversational agents: Designing for anthropomorphism. *Journal of the Association for Information systems, 22*(4), 8. https://doi.org/10.17705/1jais.00685

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2017). When do we need a human? Anthropomorphic design and trustworthiness of conversational agents. *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on* HCI *Research in MIS* Seoul, Korea.

Seiffer, A., Gnewuch, U., & Maedche, A. (2021). Understanding employee responses to software robots: a systematic literature review. *International Conference on Information Systems*, Austin, US.

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in human behavior, 117*, 106627. https://doi.org/10.1016/j.chb.2020.106627

Stoeckli, E., Dremel, C., Uebernickel, F., & Brenner, W. (2019). How affordances of chatbots cross the chasm between social and traditional enterprise systems. *Electronic Markets, 30*(2), 369–403. https://doi.org/10.1007/s12525-019-00359-6

Su, Y.-S., & Hsia, J.-H. (2022). An evaluation model of smart speaker design. In *The Routledge Companion to Technology Management* (pp. 141–156). Routledge.

Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of ChatGPT et al. the prospects of large language models. *Business & Information Systems Engineering, 65*(2), 95–101. https://doi.org/10.1007/s12599-023-00795-x

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Wambsganss, T., Höch, A., Zierau, N., & Söllner, M. (2021). Ethical design of conversational agents: Towards principles for a value-sensitive design. *International Conference on Wirtschaftsinformatik (WI)*, Essen, Germany.

Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2020). Effects of voice-based AI in customer service: Evidence from a natural experiment. *International Conference on Information Systems (ICIS)*, India.

Weiler, S., Matt, C., & Hess, T. (2022). Immunizing with information–Inoculation messages against conversational agents' response failures. *Electronic Markets, 32*(1), 239–258. https://doi.org/10.1007/s12525-021-00509-9

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

Wiethof, C., & Bittner, E. (2021). Hybrid intelligence-combining the human in the loop with the computer in the loop: A systematic literature review. *International Conference on Information Systems (ICIS)*, Austin, TX, United States.

Wintersberger, P., Klotz, T., & Riener, A. (2020). Tell me more: Transparency and time-fillers to optimize chatbots' waiting time experience. *Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, Tallinn, Estonia.

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research, 39*(1), 93–112. https://doi.org/10.1177/0739456X17723971

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. *Conference on Human Factors in Computing Systems (CHI)*, New York, NY, United States. https://doi.org/10.1145/3025453.3025496

Young, S. W. (2014). Improving library user experience with A/B testing: Principles and process. *Weave: Journal of Library User Experience*, *1*(1). https://doi.org/10.3998/weave.12535642.0001.101

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking, 21*(8), 491–497. https://doi.org/10.1089/CYBER.2017.0518

Zierau, N., Elshan, E., Visini, C., & Janson, A. (2020a). A review of the empirical literature on conversational agents and future research directions. *International Conference on Information Systems (ICIS)*, India.

Zierau, N., Hausch, M., Bruhin, O., & Söllner, M. (2020b). Towards developing trust-supporting design features for AI-based chatbots in customer service. *International Conference on Information Systems (ICIS)*, India.

Zierau, N., Wambsganss, T., Janson, A., Schöbel, S., & Leimeister, J. M. (2020c). The anatomy of user experience with conversational agents: A taxonomy and propositions of service clues. *International Conference on Information Systems (ICIS)*, India.