Leveraging the Potential of Conversational Agents: Quality Criteria for the Continuous Evaluation and Improvement

Tom Lewandowski
University of Hamburg
tom.lewandowski@uni-hamburg.de

Mathis Poser
University of Hamburg
mathis.poser@uni-hamburg.de

Emir Kučević
University of Hamburg
emir.kucevic@uni-hamburg.de

Marvin Heuer
University of Hamburg
marvin.heuer@uni-hamburg.de

Jannis Hellmich
University of Hamburg
jannis.hellmich@uni-hamburg.de

Michael Raykhlin
University of Hamburg
michael.raykhlin@uni-hamburg.de

Stefan Blum
iteratec GmbH
stefan.blum@iteratec.com

Tilo Böhmann
University of Hamburg
tilo.boehmann@uni-hamburg.de

Abstract

Contemporary organizations are increasingly adopting conversational agents (CAs) as intelligent and natural language-based solutions for providing services and information. CAs promote new forms of personalization, speed, cost-effectiveness, and automation. However, despite their hype in research and practice, organizations fail to sustain CAs in operations. They struggle to leverage CAs' potential because they lack knowledge on how to evaluate and improve the quality of CAs throughout their lifecycle. We build on this research gap by conducting a design science research (DSR) project, aggregating insights from the literature and practice to derive a validated set of quality criteria for CAs. Our study contributes to CA research and guides practitioners by providing a blueprint to structure the evaluation of CAs to discover areas for systematic improvement.

Keywords: artificial intelligence assistants, conversational agents, chatbots, quality criteria set, design science research (DSR)

1. Introduction

Due to ongoing developments in artificial intelligence (AI) and improvements in underlying machine learning (ML) algorithms, CAs are becoming increasingly relevant in organizations as essential gateways to digital services and information (Følstad et al., 2021; Gnewuch et al., 2018). Primarily operating in external or internal organizational environments, CAs can conveniently provide users (e.g., customers and employees) access to information from several connected systems and data sources. In addition, CAs are able to execute standardizable processes and tasks that have conventionally been performed by employees (Meyer von Wolff et al., 2020). Equipped with these capabilities, organizations can deploy CAs in various work contexts to efficiently and cost-effectively automate routine tasks or assist users in performing tasks (Meyer von Wolff et al., 2020). Due to their massive economic potential and capability to deliver personalized services, much research has been conducted on these AI-based systems (Cui et al., 2017; Zierau, Hausch, et al., 2020). More specifically, previous research has focused on aspects that are technical (e.g., technology selection and NLP improvements), behavioral (e.g., user trust), and conceptual or design-oriented (Diederich et al., 2019; Meyer von Wolff et al., 2021; Zierau, Elshan, et al., 2020).

Despite its promising potential, the adoption of 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; Janssen et al., 2021). Therefore, recent research has adopted a management perspective to identify the reasons for the moderate success of CAs. In this vein, factors for success and failure, as well as a continuous evaluation (e.g., monitoring) and improvement process, have been proposed to ensure the successful operation of CAs (Janssen et al., 2021; Meyer von Wolff et al., 2021). Thus far, however, 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., 2022). Therefore, this paper explores the following research question: *What are relevant criteria for continuously evaluating the quality of CAs, and how can they be applied?*

URI: https://hdl.handle.net/10125/103055
978-0-9981331-6-4

In this paper, a set of relevant criteria was developed to evaluate the quality of CAs, and a procedure model to apply the criteria was derived and evaluated. Since, in practice, many CAs fail due to a lack of knowledge concerning evaluation, a criteria-based approach can close this gap in CA research on lifecycle topics and support the CAs' operation phase (Lewandowski et al., 2022). From a practical lens, the quality criteria and procedure model can serve as an initial overview for organizations to systematically structure CA evaluation to discover areas for improvement. Following DSR activities, we present insights from the literature and practice to derive a validated set of quality criteria for CAs. Hence, the remainder of our paper is structured as follows. Section 2 presents the related CA research and delineates the research gap. In Section 3, we describe the DSR approach to developing our artifact. In Section 4, we present the findings of our study, including an overview of our final quality criteria set. Subsequently, Section 5 outlines the instantiation of the quality criteria set using a real-life case in an IT organization. We discuss our findings and conclude with our limitations and contributions in Section 6.

2. Related Research

The vision of communicating with information systems (IS) has been around for nearly 50 years. An early example is ELIZA, which allowed initial natural language-based interactions with a computer (Weizenbaum, 1966). However, technical limitations restricted early attempts at CAs (Diederich et al., 2019). Nevertheless, in recent decades, massive technological progress has allowed the development of progressively more intelligent CAs. Consequently, CAs, known under numerous designations, such as chatbots, chatterbots, or dialog systems, have gained interest, leading to discussions in the literature about a delimitation of the terms. We use the term 'conversational agent' in this paper to refer to all AI- and text-based representations, such as chatbots (cf. Gnewuch et al., 2017), since the CAs investigated in the real-life DSR project were text-based.

Today, CAs are increasingly adopted and have attained popularity in various commercial and private application domains (Meyer von Wolff et al., 2020). Integrated into various front- and back-end systems, such as websites or messaging applications (e.g., MS Teams), CAs support organizations' ongoing digitization and automation by doing things such as filtering information or efficiently assisting employees in daily work tasks (Zierau, Elshan, et al., 2020). Hence, with their scalability and 24/7 availability (Gnewuch et al., 2017; Xu et al., 2017), CAs can make a transformative contribution by providing a convenient way for more individual interactions, such as acting as a central service platform and first point of contact for customers before they reach out to actual employees (Zierau, Elshan, et al., 2020). Thereby, users' high load of information is reduced (Xu et al., 2017). Moreover, employees can concentrate on their core and non-routine tasks.

Nevertheless, developed CAs still have a high failure rate (Janssen et al., 2021). Many fail in real-world environments due to, among other things, frustrating user experiences (Følstad, Nordheim, et al., 2018). As a result, multiple organizations take their CAs offline because they lack knowledge of quality criteria and aspects relevant to continuous evaluation and improvement, resulting in an uncoordinated and highly explorative development process (Janssen et al., 2021). Moreover, CAs represent a novel form of learning, unfinished, user-centric, and socially interactive IS that has introduced, so far, unsolved challenges (Lewandowski et al., 2021; Zierau, Elshan, et al., 2020). A distinctive feature of CAs is their capability to learn and improve via naturalistic interactions. Accordingly, CAs' learning progress is highly context-driven and thus dependent on actual application and usage (Clark et al., 2019; Zierau, Wambsganss, et al., 2020). Because of this unfinished and learning nature of CAs, novel approaches to handle their implementation and improvement in their lifecycle are required since they initially possess limited functions and require several interdisciplinary design activities (Lewandowski et al., 2022; Meyer von Wolff et al., 2021).

Consequently, the highest effort occurs in operations, where CAs require continuous evaluation and later training and improvement in a real-world context, often characterized 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). Although CAs have gained a great deal of research attention from specific conceptual, usability, or technical design perspectives, the operation in general, and continuous improvement process, specifically, lack detailed theoretical and practice-based knowledge (Lewandowski et al., 2022; Meyer von Wolff et al., 2021). Hence, a clear criteria-based approach to continuously evaluate CAs' quality in the further development is needed to sustain them. First researchers have already investigated success and failure factors for CA implementations from an organizational perspective (e.g., Janssen et al., 2021; Lewandowski et al., 2021; Meyer von Wolff et al., 2021). However, they tend to address the managerial perspective and do not focus on the continuous

improvement process. Other authors have studied the different effects of CAs on an individual level, either on perceived trust, enjoyment, or affordance theory (Stoeckli et al., 2019; Zierau, Hausch, et al., 2020) or in the wider 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 improvement. Initial contributions exist in evaluating CA design. Nevertheless, current research is (1) confined to technical measurements (e.g., Alonso et al., 2009), (2) other agent classes (e.g., Kuligowska, 2015) and (3) individual design aspects (e.g., Seeger et al., 2021), while (4) being segregated. Further, research (5) focused on human behavior or ethical aspects (e.g., Radziwill & Benton, 2017) and (6) initial classifications and typologies for only a high-level analysis and guidance on interaction design (Følstad, Skjuve, et al., 2018), which for CA teams only play a superordinate role in CA development. Thus far, a holistic overview of criteria for researchers and practitioners for constant evaluation and sustainability throughout the CA lifecycle is lacking.

3. Research Approach

This article aims to provide CA quality criteria that will allow organizations to continuously evaluate and improve CAs during their lifecycle, as proposed by Lewandowski et al. (2022). To achieve this goal, we adopted the DSR paradigm and applied the three-cycle view by Hevner (2007). Overall, we conducted seven research activities (see **Figure 1**).

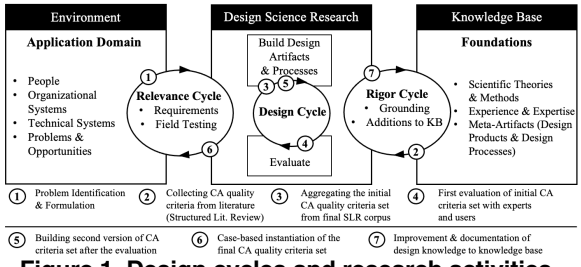

Figure 1. Design cycles and research activities, according to Hevner (2007).

The first step of the DSR approach is the identification of a pervasive real-world problem. In accordance with the Introduction and Related Research, we are building our research on the current lack of an overview in organizations concerning how a criteria-based approach could sustain the operation and continuous improvement of CAs to ensure their success throughout their lifecycle (see **Section 2**).

Based on this problem, in Step 2, we conducted a structured literature review (SLR) according to the five-step process of vom Brocke et al. (2009) in the databases of AISeL, ACM DL, IEEE Xplore, EBSCO, and ProQuest ABI/INFORM to derive initial criteria for evaluating CA quality. In this process, we based the subphases on established methods. For instance, we followed the taxonomy proposed by Cooper (1988) to define our SLR scope and Brink (2013) for the well-founded creation of a synonym list to structure the search process. We identified and verified suitable keywords via an initial database search to create 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"). Applying the search string to the aforementioned databases, we obtained 1895 articles. We selected 180 of these for an in-depth analysis by screening each article's title, abstract, and keywords. Utilizing deselection criteria ((1) technical or architectural aspects, (2) physical machines/robots, and (3) lack of CA application case) and deleting duplicates, we arrived at 94 articles. In a final full-text analysis, we classified 67 articles as relevant.

As part of Step 3, we initiated the design cycle to create an initial quality criteria set. To do so, we followed a multistep procedure. First, two researchers independently analyzed the full texts of the final 67 articles from Step 2 to identify suitable criteria. Second, the resulting criteria set (containing 221 potential criteria) was revised and condensed by (1) filtering out non-CA-specific criteria (e.g., related to the design of the messenger front end), (2) synthesizing similar and redundant criteria, (3) weighting aspects that multiple authors have addressed, and (4) deleting aspects irrelevant for evaluating text-based CAs (e.g., only relevant for speech-based assistance systems).

In Step 4, the initial literature-based criteria set was evaluated and expanded by interviewing seven CA users and experts. These interviews were conducted in December 2021 and lasted an average of 37 minutes. To ensure a systematic procedure, we developed a semi-structured guide, following the instructions of Gläser and Laudel (2009). Experts were asked (1) about their CA experience and possible quality criteria, and afterward, (2) we presented the quality criteria from the literature to let them rate existing criteria and point out missing aspects.

Building on these insights, as part of Step 5, we created the final criteria set consisting of meta-criteria, criteria, and sub-criteria (see **Section 4**). Utilizing the insights from Step 2, as well as the statements from the expert interviews, we decided whether (1) a criterion had to be retained, revised, or added to the criteria set and (2) whether the criteria set was comprehensible or needed to be restructured overall.

In Step 6, we conducted a naturalistic evaluation of the final quality criteria set by supervising its instantiation in an IT organization (see **Section 5**)**.** The goal was to verify whether the criteria set can be utilized to evaluate CA quality and whether it has the potential to help organizations continuously improve CAs in a structured way. Finally, we incorporated the findings from the naturalistic instantiation into the criteria set and improved and communicated them.

4. Quality Criteria Set for CAs

Based on the DSR research activities, we have derived a final criteria set consisting of 6 meta-criteria and 15 criteria with 33 sub-criteria to evaluate and improve CAs' quality throughout their lifecycle. The criteria set supports a cyclical evaluation process carried out at specified intervals in CAs' lifecycle which is performed based on previously collected data (e.g., monitoring, performance, or user data). In **Table 1**, we list the criteria along with example references that provide corresponding sources and interview insights.

4.1. 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 MS Teams, or websites), ensuring 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 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). Especially in the interviews, the need to continuously refine the selection and functionality of control elements was emphasized (e.g., text, buttons, reactions, and carousel selections). In addition, the **context awareness** of CAs should be evaluated. The ability to grasp *dialog-oriented context* allows CAs to incorporate previous user utterances to conduct a conversation 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 for evaluation. A well-structured conversation flow helps users provide the correct input, achieve their goals, and avoid deadlocks (Diederich et al., 2020). In addition, the *technical context* 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 to provide correct data for the user's input (e.g., one-time user identification to address background systems to resolve requests).

4.2. Output

Regarding **Output**, the **format** of CA responses should be reflected. The responses require the appropriate selection of a *suitable output format* in terms of a user- and content-oriented presentation (e.g., with texts, images, and tiles). An *appealing output formatting or visual representation of CA responses* is recommended (Kowald & Bruns, 2020). Especially in the CA context, users prefer short and manageable CA answers (Edirisooriya et al., 2019). In terms of **content**, the CA should transparently present its *capabilities* and *limitations* to evoke an appropriate user expectation that is consistent with the nature of the CA as an unfinished and 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 presented information and the up-to-dateness of the knowledge base for *information retrieval* should be checked to decide 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 solutions that are often complex in an abstract and *convergent* way that provide users with appropriate answers to their concerns. The interview experts highlighted that solutions sorted by relevance and *justification* of the CAs' answers could increase user trust in these answers. For example, a CA could refer to the background system/source to make 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* (also referred to as response quality, e.g., Jiang & Ahuja, 2020) needs to be evaluated to present knowledge correctly (e.g., length, tonality, fluency) to the target group. Regarding the **timing of responses**, on the one hand, the *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. On the

other hand, the criterion *balance between proactivity and interruption* refers to the fact that CAs' proactive utterances may interrupt users. This behavior and its effects on users should be evaluated.

4.3. Anthropomorphism

Anthropomorphism refers 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 also 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 position, size, attractive [humanlike] appearance, and colors. Furthermore, CAs' **verbal cues** should be reviewed. In addition to the ability to engage in social dialogues, called *"chitchat," emotional expressions* (e.g., apologizing by the CA), *verbal style*, and *self-reference* (e.g., the chatbot 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). Continuously improving social skills has already had a short-term impact on the success of a CA. However, researchers (e.g., Grudin & Jacques, 2019) have also noted that a humanlike CA can be repellant to users. Seeger et al. (2021) indicated that the different anthropomorphism criteria must be combined and evaluated practically.

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 context	Interviews
Output	Format	Suitable format	(Edirisooriya et al., 2019; Feng & Buxmann, 2020; Kowald & Bruns, 2020), Interviews
		Appealing formatting and visualization	
	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)
Anthropo-morphism	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	Typing delay and indicator	(Gnewuch et al., 2018; Schuetzler et al., 2021; Seeger et al., 2021), Interviews
		Emoticons	
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; Poser et al., 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
		Average number of turns	
		Human-handover rate	(Wintersberger et al., 2020), Interviews
Data privacy	Implementation and communication	Privacy and anonymity	(Feng & Buxmann, 2020; Janssen et al., 2021; Lewandowski et al., 2021; Rajaobelina et al., 2021), Interviews
		Transparency	

Table 1. Final CA Quality Criteria Set

4.4. 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 are, therefore, initially error-prone. In particular, user input in long 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). If no appropriate answer was elicited, the CA could proactively refer to misunderstandings or reintroduce his skills (interviews). 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 its response (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 suggestions on text buttons) and to proactively avoid error cases when processing a user's free 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 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; Poser et al., 2022).

4.5. Performance

A holistic evaluation of CA performance represents a strong predictor for CA success (Peras, 2018). By combining design-and technically-oriented principles, the CAs' **performance** is directly related to user satisfaction (Liao et al., 2016). The performance demonstrates the effective and efficient completion of executed 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 average number of (successful) tasks and the average number of default fallback intents to trigger appropriate countermeasures (Peras, 2018). In the interviews, the *retention and feedback rates* were mentioned regarding recording returning users and continuously evaluating users' average ratings to uncover weaknesses to derive improvement potential. Furthermore, it is necessary to consider CAs' **efficiency** because the effective performance of tasks explicates only a few insights into whether the CA also performs the tasks with a resource-based approach. Given this perspective, evaluating the average time used to complete a task (*task completion time*) and the average number of rounds of dialogue 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).

4.6. Data Privacy

Data privacy includes criteria related to the implementation and communication of data protection. In the **implementation of data protection**, a relevant criterion is that the conversations with the CA should be kept as *private* and *anonymous* as possible, especially if the CA's context is confidential and personal data are processed (Feng & Buxmann, 2020). During the interviews, it was emphasized that as little data as possible should be stored during a conversation, and anonymized data should be stored if conversational data is obligatory 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 for users (Rajaobelina et al., 2021).

5. Case-Based Instantiation

After the rigorous derivation process, the final quality criteria set including all meta-criteria was instantiated in an IT organization to evaluate its applicability and feasibility. To this end, an existing CA *(ExpertBot)* was evaluated and improved along the criteria set by using various evaluation methods. The *ExpertBot* operates within organizational boundaries, is integrated into a messenger, and facilitates employees' search for internal experts (and their skills) to help employees and staff projects. Therefore, the CA participates in chat-based conversations involving multiple employees to suggest suitable experts by accessing diverse data sources (e.g., skill database, document management systems, internal chat forums).

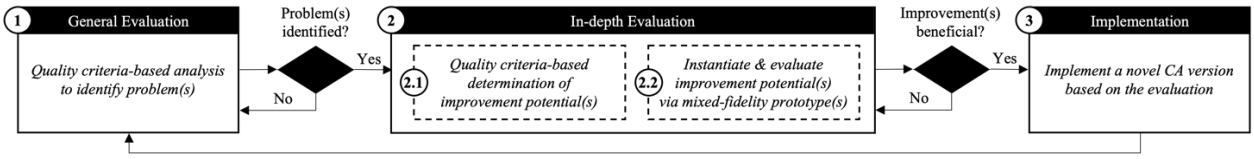

Figure 2. Procedure Model for the Evaluation and Improvement of CAs

Based on the criteria, an interdisciplinary team comprising experts from research and the IT organization conducted the evaluation of the *ExpertBot*'s quality. As there is limited substantiated knowledge on the procedure and selection of CA evaluation methods and required experts, an *explorative and iterative process* was initiated in cooperation with the IT organization. As a result, a procedure comprising three phases was completed (see **Figure 2**): In phase 1, the **General evaluation**, we performed a quality criteria-based analysis to identify problems of the current CA version in operations. The general evaluation revealed that the overall performance of *ExpertBot* was insufficient. Significant problem indicators, such as user retention and feedback rate, were considered throughout the criteria-based analysis to start an in-depth evaluation. Consequently, we initiated an improvement project to increase its performance.

As part of phase 2, the **In-depth evaluation**, we first conducted phase 2.1, an evaluation in cooperation with the organization to assess CAs' quality and determine improvement potentials based on our criteria set. In this context, 13 interdisciplinary participants (e.g., CA developers, employees responsible for staffing, and employees outside the subject area) were recruited to conduct an in-depth evaluation of the other meta-criteria to gain insights into the current *ExpertBot* quality and possible criteria interdependencies. This procedure allowed a multi-perspective in-depth evaluation of the *ExpertBot* due to the participants' varying experience levels regarding CAs and the broad discussion of criteria and weaknesses of the *ExpertBot*. Thereby, phase 2.1 was performed using a mixed-method approach. Semi-structured interviews were conducted with 7 of the 13 participants in the beginning. We presented the current *ExpertBot* version to ask the participants about the general implementation and relevance for improvement along with the individual criteria from our set. Based on the analyzed qualitative data, a survey was conducted to ask participants to rank previously determined potentials according to their relevance for improving the *ExpertBot*.

In phase 2.2, we instantiated prototypes illustrating the highest ranked improvement potentials uncovered with the criteria set to test their feasibility and demonstrate an CA improvement. In this context, the prototypes provide a well-founded comparison of the current and novel CA version(s) and allow to involve real users by providing a basis for decision-making on whether the identified improvement potentials are beneficial when they are implemented or need to be revised. Therefore, the results of phase 2.1 were used to create mixed-fidelity prototypes, present them to participants, and compare them to the current CA version. Prototypes were designed based on the prioritized improvement potentials corresponding to the analyzed *ExpertBot*. For this purpose, the Figma Inc. (2022) design tool was used. The improvements were arranged into several scenarios, assembling suitable criteria (e.g., one scenario focused on the meta-criterion output with selected sub-criteria) to visualize and evaluate them with individual prototypes. Thereby, we presented all participants two prototypes for each scenario during semi-structured interviews. The first prototype contained the assumed improvements, while the other prototype represented the current CA state. During the presentation, three questions were asked regarding each scenario. First, participants were asked to evaluate which of the two prototypes was more effective at first glance and which aspects were crucial to this impression. Second, scenario criteria were individually addressed, and the participants were asked to determine which criterion is conceivable for increasing CA quality. Third, we asked which implemented criteria were the most important in elevating CA quality.

Finally, in phase 3, the improvement potentials identified in the evaluation to be particularly effective for increasing CAs' quality were implemented in a novel CA version by the CA team. After phase 3, the procedure should again start with phase 1 to examine the quality of the new CA version, which, however, was not part of the instantiation.

6. Discussion and Conclusion

Contemporary CAs have attracted considerable attention in organizations and academic research, introducing a paradigm shift in how users interact with IS (Zierau, Wambsganss, et al., 2020). However, CAs have a high discontinuation rate (Gnewuch et al., 2017; Janssen et al., 2021). In this context, a holistic overview of how to evaluate and improve CA quality throughout its lifecycle is lacking. From a research perspective, primary contributions exist in evaluating

CA design (e.g., Seeger et al., 2021). However, existing scientific knowledge is segregated and does not yet address how CAs can be continuously evaluated and improved. Moreover, structured knowledge on how to conduct an improvement process is so far lacking. This is unsatisfactory, since CA development comprises several novel and effortful activities that should be systematically orchestrated.

To close this knowledge delta, we conducted a rigorous DSR project by aggregating insights from the literature supplemented by experiences from the practice-based, real-life environment to derive a systemized and synthesized set of CA quality criteria. In addition, we developed a procedure model in the context of the instantiation of the criteria set.

We contribute the presented criteria set, serving organizations as an overview of relevant aspects to evaluate and improve the quality of CAs as part of their operations. In combination with the application of the prototype method, the instantiation of the criteria set can pave the way to systematically evaluate and improve CAs by comparing different CA versions. First, the application of the criteria set enables organizations and CA teams to check whether a new CA version possesses better quality than the current version. Consequently, it can be ensured that a new CA version will be deployed only if its quality is at least as high as the previous version. Second, comparing the quality of the two CA versions may reveal improvement potentials before going live. Third, against the backdrop of moderate CA success, determining proper criteria can help CA teams (even beforehand) to design better CAs and evaluate them with users to confirm their intended use.

In addition to the criteria set, we contribute a procedure model, serving as a blueprint to apply the criteria set. This allows to structure the evaluation of CAs and discover areas for systematic improvement. Regarding required experts, we discovered that the involvement of people from different departments is beneficial for the evaluation process, as CA development is highly interdisciplinary and demands the combination of technical and design-oriented aspects (e.g., intent recognition, dialog design, CAs' front channel). For the instantiation, people from the IT, business and data protection departments were involved. Furthermore, people outside the subject area can significantly contribute to CA evaluation and improvement. In general, CAs' quality criteria evaluation should be conducted as naturally and quickly as possible to identify actual user behavior.

Overall, the combination of the applied criteria set and procedure model in the IT organization helps to address the knowledge gap on how to reduce the discontinuation rate of CAs and evaluate and improve CAs' quality throughout their lifecycle to sustain their operation.

However, the instantiation of the criteria set also revealed three challenges and aspects that need further research. First, as there is limited substantiated knowledge on the procedure and selection of CA evaluation methods and required experts in general, an explorative process was selected, which was time-consuming in terms of both the actual activities as well as the application of the methods. Further research is needed to explore alternative or faster ways of performing activities and methods for criteria-based evaluation of CAs' quality. There may also be automation potentials with tool support. Closely related, a guideline is needed to determine when such an evaluation should be performed, in general, and who must be involved during the process. Second, in phase 1 of our instantiation (see Chapter 5), we determined the need for a quality criteria-based in-depth evaluation of CAs' performance in their natural context. The performance criterion proved to be a valid indicator for the improvement of *ExpertBot*. Nevertheless, further investigation is required to determine whether there are additional indicators to start the in-depth evaluation. For example, with increasing CA progress and more team expertise, other aspects of the criteria set could trigger an in-depth evaluation. Furthermore, there may be criteria that need to be more or less frequently evaluated, designed, or technically improved. Third, concerning the criteria set, we observed that different criteria have varying levels of influence on CAs' quality. Additionally, we discovered that specific criteria from our set differed in their importance depending on the expertise of our interview partners. For instance, anthropomorphism was less significant in the interviews and ranked low in phase 2.1 compared to the findings in the literature. Although the cue design of the CA was dominant in the reviewed literature, several experts stated that CA humanization was not as relevant for the use case considered in the project, which can be attributed to the fact that many employees in the IT organization have an IT versed background. A classification or assessment ranking of the criteria's influence and importance combined with a more in-depth procedure offers additional potential for future research.

Despite these valuable insights, there are a few methodological limitations, which provide further avenues for future research. First, concerning DSR, one objective was to apply our final quality criteria set in a naturalistic evaluation setting to verify whether the set could serve CA teams in evaluating and revealing potential improvements in a procedural and structured way. Although our set was applicable and could meet those objectives, the instantiation referred

only to a single CA team in an IT organization. Further studies need to identify whether the criteria set can be applied to other organizations or if it needs to be extended or reorganized based on further perspectives. Second, the experts in this study and their domain-specific experiences influenced the study's external validity. In particular, our derived knowledge is dependent on their experiences. Finally, we recognize that the results depend on the authors' literature selection, aggregation, and judgment. Further studies could modify or prioritize the quality criteria set and reveal significant interdependencies for CA teams.

6. Acknowledgments

The research was financed by the German Federal Ministry of Education and Research and the European Social Fund (funding ref.: INSTANT, 02L18A111).

7. 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.

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.

Brink, A. (2013). *Anfertigung wissenschaftlicher Arbeiten* (5th ed.). Springer Gabler.

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 HCI*, *37*(8), 729–758.

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, New York, NY, United States.

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

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.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019). *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. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering*, *62*(3), 193–209.

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

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 Inc. (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.

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). *What makes users trust a chatbot for customer service? An exploratory interview study.* International Conference on Internet Science (INSCI),

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2018). *Different chatbots for different purposes: towards a typology of chatbots to understand interaction design.* International Conference on Internet Science (INSCI),

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

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.

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.

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

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, *19*(2), 4.

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,

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.

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.

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.

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*, *13*(3), 26.

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

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., Heuer, M., Vogel, P., & Böhmann, T. (2022). *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),

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.

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.

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-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.

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

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).

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*.

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.

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.

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.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information systems*, *22*(4), 8.

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*, 1–35.

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),

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.

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.

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.

Zierau, N., Elshan, E., Visini, C., & Janson, A. (2020). *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. (2020). *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. (2020). *The anatomy of user experience with conversational agents: A taxonomy and propositions of service clues.* International Conference on Information Systems (ICIS), India.