10-9-2023

How to Convey Resilience: Towards A Taxonomy for Conversational Agent Breakdown Recovery Strategies

Shengjia Feng
Technical University Darmstadt, Germany, shengjia.feng@gast.tu-darmstadt.de

How to Convey Resilience: Towards A Taxonomy for Conversational Agent Breakdown Recovery Strategies

Research Paper

Shengjia Feng[1]

[1] Technical University Darmstadt, Software & Digital Business, Darmstadt, Germany
shengjia.feng@gast.tu-darmstadt.de

Abstract. *Conversational agents (CAs) have permeated our everyday lives in the past decade. Yet, the CAs we encounter today are far from perfect as they are still prone to breakdowns. Studies have shown that breakdowns have an immense impact on the user-CA relationship, user satisfaction, and retention. Therefore, it is important to investigate how to react and recover from breakdowns appropriately so that failures do not impair the CA experience lastingly. Examples for recovery strategies are the assumption of the most likely user intent (CA self-repair) or to ask for clarification (user-repair). In this paper, we iteratively develop a taxonomy to classify breakdown recovery strategies based on studies from scholarly literature and experiements with productive CA instances, and identify the current best practices described using our taxonomy. We aim to synthesize, structure and further the knowledge on breakdown handling and to provide a common language to describe recovery strategies.*

Keywords: Conversational agents, Resilience, Breakdown, Recovery Strategies.

1 Introduction

Conversational agents (CAs) went through one of the most impressive evolutions of all technological advancements. With the launch of the ChatGPT, the power of Artifical Intelligence (AI) seems infinite and tangible for everyone. Although business use cases for GPT conversational agents are yet to be explored (Chui et al., 2022), they can already be experienced in the form of personal assistants in mobile devices or as virtual customer service agents on websites. Users have high expectations as CAs oftentimes converse in a natural, human-like way. Yet, we observe limitations in their capabilities: While CAs previously often struggled with understanding or serving user inquiries (Luger and Sellen, 2016), CAs based on large language models give very persuasive answers with no factual grounding, a phenomenon called 'hallucination.' Studies show that when user encounter breakdowns in their interaction with a CA, it can lead to anger (Han et al, 2021), user dissatisfaction (Akhtar et al., 2019), and long-lasting negative adversions (Luger and Sellen, 2016). In fact, C.-H. Li et al. (2020) found that nearly

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

half of the users who abandoned a CA long-term only faced one breakdown. For CAs in productive use, breakdowns can cause severe Public Relations issues for the company (Prahl and Goh, 2021). Hence, it is crucial to consciously develop and select repairment and recovery strategies to mitigate breakdown effects.

In this work, we aim to establish a comprehensive understanding of the design space and provide a framework to accurately describe recovery strategies. This common framework is needed to efficiently organize the design space and to identify designs that are yet to be explored in research and practice. Our research question is as follows:

RQ: What are the dimensions and characteristics of a taxonomy that comprehensively and accurately classifies a CA recovery strategy?

The purpose of this work is trifold: First, we develop a framework to describe CA recovery strategies based on research studies and productive instances of CAs. In a second step, we extract recurring patterns in the CAs investigated during taxonomy development. Lastly, we identify starting points for further studies. Following the call to further investigate CA design (Rapp et al., 2019), we aim to extend the existing knowledge and provide guidance on how to deal with unwanted CA failure.

2 Related Work

A CA is "*a software program that interacts with users using natural language*" (Shawar and Atwell, 2007, p. 31) in a way as if a human was on the other side. CAs have had various roles throughout its long history: As a psychologist ('ELIZA'; Weizenbaum, 1966), counsellor (Benyon et al., 2013), virtual buddy (Fitzpatrick et al., 2017; Inkster et al., 2018), in teaching roles (Baylor and Kim, 2005; Ogan et al., 2012), and as a personal assistant (Klopfenstein et al., 2017). Today, CAs are at a developmental stage that they can support at the workplace (Feng and Buxmann, 2020; Meyer von Wolff, 2020) and workers may even consider them as collaborators rather than a tool (Bittner et al., 2019; Seeber et al., 2020). With the feasibility to train large language models, many new possibilities open up with its generative and conversational power.

But CAs can encounter breakdowns during interactions. A popular breakdown is Microsoft's *Tay* going rogue after only few hours of operation (Wakefield, 2016). Although Tay's breakdown can be attributed to abusive users, the creators should foresee such critical interactions (Suárez-Gonzalo et al., 2019; Wolf et al., 2017). Amershi et al. (2019) proposed the category 'when wrong' for their design guidelines on Human-AI interaction, indicating that breakdowns are a normal stage of interaction to be considered. A breakdown, therefore, is a situation that does not contribute to the CA's value delivery and is most likely an exceptional state in the conversation, but should not be one that is unplanned for. However, neither the CA nor the user has to be aware of the breakdown, for instance, when ChatGPT hallucinates and the user cannot validate the statements due to lacking domain knowledge himself. In this work, we focus on cases where the CA does know its breakdown, since that is the prerequisite to trigger a recovery strategy to mitigate the negative effects of the situation.

Upon breakdown, the conversation needs to recover from the exceptional state in order to progress. This calls for a recovery strategy that decides on the next action. We

can generally divide recovery strategies in two main categories: The system-repairs and the user-repairs. For a system-repair, the prerequisite is for the system to recognize its breakdown without user indication, e.g., by learning to recognize breakdowns using techniques like Machine Learning (Almansor and Hussain, 2021; Kontogiorgos et al., 2020; Reinkemeier and Gnewuch, 2022). A system-repair with no user involvement means that the system resolves the issue by assuming and execution the 'best next action' (e.g., Ashktorab et al., 2019) or by proceeding the conversation with the 'best intent' (e.g., Dippold, 2023). User-repairs involve the user to different degrees to assist the CA and resolve the breakdown together. This can be conducted via asking questions to clarify a cryptic message (Müller et al., 2021) or providing options to proceed in the form of utterance templates or buttons (Benner et al., 2021). Furthermore, the perceived effect and the appropriateness of breakdown recovery strategies are influenced by many other factors, such as apologeticness (Song et al., 2023; J. Zhang et al., 2023), usage of special language elements like Emojis (Liu et al., 2023; Wang et al., 2023), other anthropomorphic features (Seeger and Heinzl, 2021), and timing of breakdown and recovery measures (Huang and Dootson, 2022; Kim et al., 2023).

3 Methodology & Taxonomy Development

Based on Nickerson et al. (2012), a taxonomy T consists of n dimensions D_i $(1 \leq i \leq n)$. Each dimension D_i is a set of k_i $(i \geq 2)$ characteristics C_{ij} $(1 \leq j \leq k_i)$. In the taxonomy, the characteristics in the same dimension should be mutually exclusive and all dimensions collective exhaustive, i.e., each object under consideration has exactly one characteristic C_{ij} of each dimension D_i. We follow the taxonomy development method proposed by Nickerson et al. (2012) to develop the taxonomy, therefore, the steps described in the following refer to their procedure.

Step 1: Determine meta-characteristic. We aim for a high-level design guide for CA creators on how to recover from a CA breakdown. Therefore, our meta-characteristic is 'recovery strategies for CA breakdowns.' Note that we treat breakdowns as single, standalone incidents. Different strategies can be applied by CA for multiple breakdowns, although consistency has shown to elevate usability (T. J.-J. Li et al., 2020).

Step 2: Determine ending conditions. After each the taxonomy needs to be assessed regarding objective and subjective ending conditions. For our taxonomy development, we use all conditions as proposed by Nickerson et al. (2012, p. 9).

3.1 Iteration 1

Step 3: Conceptual-to-empirical. For the first iteration, we choose the conceptual-to-empirical approach to include work that can set a first basis that can be validated and extended in further iterations. Ashktorab et al. (2019) already provides a variety of recovery strategies that form the first dimensions of our taxonomy. This paper was selected as the authors already performed studies to investigate users' perceptions of the strategies which we can take into consideration for the second goal of this paper, which

is to provide a best practice guideline to practitioners. The strategies provided also reflect common strategies one encounter in productive CAs.

Step 4c: Conceptualize new characteristics and dimensions of objects. Based on Ashktorab et al. (2019), we define three dimensions: 'Acknowledgment' indicates if the CA explicitly acknowledges the breakdown towards the user. The second dimension is the type of 'Explanation' to the user: the CA can provide what it understood (positive), what it did not understand (negative), or no explanation. The third dimension is the 'Coping Strategy Type:' the CA can cope on system-side, assist the user to repair, or rely on the user to handle the breakdown.

Step 5c: Examine objects for these characteristics and dimensions. For this iteration the objects can be directly taken from the studies by Ashktorab et al. (2019) as the dimensions are derived from their investigations. In their study, they investigated different customer service CAs in the fields of shopping, banking, and travel.

Step 6c: Create taxonomy. The taxonomy T_1 resulting from this first iteration is as follows: $T_1 = \{$Acknowledgment (Acknowledging, No acknowledgment), Explanation (Positive Explanation, Negative Explanation, No Explanation), Strategy Type (System-repair, Assisted User-repair, User-repair)$\}$

Step 7: Ending Conditions met? Since this is the first iteration, the objective ending conditions are naturally not fulfilled. We consider T_1 fulfilling all subjective conditions except for comprehensiveness. Due to the small number of objects examined in this iteration, we cannot claim this attribute yet.

3.2 Iteration 2

Step 3: Empirical-to-conceptual. In the second iteration, our goal is to identify more objects from literature. Our approach is similar to a systematic literature review (SLR; Kitchenham, 2004; Webster and Watson, 2002): We perform searches in the relevant databases AISeL, the ACM DL, ScienceDirect and IEEExplore. The latest search was performed in June 2023 with the search term *("conversational agent" OR "chatbot") AND ("breakdown" OR "repair" OR "resilience" OR "coping" or "recovery").* We include all publications of 2017 or later as research interest significantly increased in this timeframe (Feng and Buxmann, 2020; Rapp et al., 2021). This way, we can include a representative number of papers while limiting the effort as we do not aim to perform an exhaustive SLR. Where possible, only 'research articles' or 'peer-reviewed' publications are included. We scan the search results for original studies focusing on CA breakdowns and recovery strategies in three steps: We review the title, then the abstract and lastly, the full-text. We include publications about non-text-based agents if concepts were tested that are applicable to text-based CAs. Our process is depicted in Figure 1 and yielded a total of 22 papers in the final set (duplicates removed).

Step 4e: Identify subset of objects. The final set of 22 papers includes 45 recovery strategies in total and is listed in Table 2 with the recovery strategies investigated classified based on the final taxonomy. When there are multiple strategies presented in one publication, each row refers to one strategy.

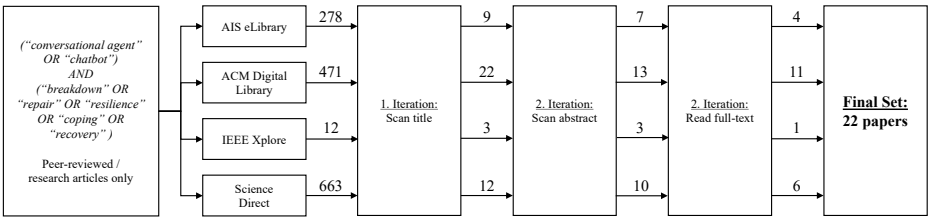

Figure 1. Overview of the Literature Review Process

Step 5e: Identify common characteristics and group objects. Among the recovery strategies of the objects investigated in this iteration, we observe that there are different levels of acknowledgment. Many recovery strategies make use of anthropomorphism in the form of an apology to improve the CA's likeability and to retain trust in the user-CA relationship (e.g., Law et al., 2022; Mahmood et al., 2022; Song et al., 2023; X. Zhang et al., 2023). We observe that button options are often used to continue with limited functional range (e.g., Ashktorab et al., 2019; Law et al., 2022; C.-H. Li et al., 2020). This aligns with findings that user expectations need to be managed to prevent (further) breakdowns (Luger and Sellen, 2016). In addition to providing an explanation, we identify different types of explanations: the user's utterance was not understood correctly ("Unfortunately, I do not understand your request." from Diederich et al., 2020, p. 8) or the intent was not recognized ("Did you mean? need *help* *thank you* clinic's *location* *None* of these" from Dippold, 2023, p. 27).

Step 6e: Group characteristics into dimensions to create taxonomy. Based on our observations, we extended 'Acknowledgment' with the apologetic acknowledgment. We refine the assisted user-repair to differentiate between the 'open' assistance and the assistance with functional limitation. Lastly, we add a new dimension to account for the explanation type. The taxonomy T_2 is as follows: $T_2 = \{$Acknowledgment (Apologetic, Neutral, None), Explanation (Positive, Negative, None), Type of Explanation (Utterance-based, Intent-based, None), Strategy Type (System-repair, Limited Assisted User-Repair, Assisted User-repair, User-repair)$\}$

Step 7: Ending Conditions met? Due to the addition and modification of dimensions, the objective ending conditions are not met. For the subjective ending conditions, this taxonomy is concise, extendible and explanatory and robust. However, we need at least one more iteration to prove comprehensiveness.

3.3 Iteration 3

Step 3: Empirical-to-conceptual. In our third iteration, we investigate CAs in productive use. In order to examine their recovery strategies, we need to provoke breakdowns first. For that, we develop a testing scheme to systematically evaluate based on different breakdown causes identified in Section 2. We define three utterances per breakdown cause, twelve in total. The complete testing scheme is provided in Table 1. Each utterance is entered in a new conversation session to avoid side effects from consecutive breakdowns. We search for CAs on landing pages as well as on 'Customer Service' and 'Contact' pages, since CAs are most often available on these pages and most likely

allow for free text input (i.e., not a hard-wired decision tree logic), so the same testing scheme can be applied to all CAs tested. Hence, CAs in this iteration are mainly focusing on customer service management. In the utterances, placeholders are used for subjects to select appropriate in-scope and out-of-scope products. For example, asking for 'van' on an automotive company's website would be in-scope, but asking about a 'banana' is out-of-scope, whereas the opposite would be valid for a grocery store.

Table 1. CA Testing Scheme - Overview of Utterances

Trigger	Utterance	Description
Language / Vocabulary	*Können Sie mir nähere Informationen zu Ihren Produkten zusenden?*	Unknown language (German)
	Pouvez-vous m'envoyer des informations détaillées sur vos produits?	Unknown language (French)
	Send me information you're products	Language errors
Out-of-scope	*What is the weather like today?*	Off-topic request
	Please turn off the lights in the living room.	Off-topic request
	Why does my [product] make weird noises?	Very specific request
Multiple in-scope intents	*I bought a [product] last week, but there is something wrong with it and I would like to return it. How is the process?*	Multiple intents: Complaint, return process
	I have an account, but forgot my credentials. Now I want to change my address.	Multiple intents: Restore credentials, change address
	I ordered [product] but I need to change the billing address. If that is not possible, I will need to cancel my order.	Multiple intents with if-condition: Change address, cancel order
Low Semantic Value	*Hello*	Utterances with no/low semantic value
	Thank you	
	Lol	

Step 4e: Identify new subset of objects. For our tests, we include publicy accessible CAs based on the following criteria: It is not only an interface to chat with a human, understands English, allows free text input, and is available for multiple conversations. Since CAs are oftentimes used to automate low-effort processes, CAs mostly used by larger companies with a high number of reoccurring (customer) requests. This lead to our strategy to scan the websites of Fortune 500 companies fron different industries for suitable chatbot instances. Despite a variety of CAs used in customer service, only few allow for free text input, many CAs built using low-code development platforms are more similar to a conversational display of a decision tree and therefore eliminate the possibility to provoke breakdowns. We examined five CAs from companies in the industries retail, automotive retail, consulting, transportation and information technology.

Step 5e: Identify common characteristics and group objects. In our tests, we identify a new exerted type of acknowledgement: Unapologetic, but with willingness

to help ("*I might be able to help…*"). Another characteristic we noticed is the transparent communication of the limits of the its capabilities. This constitutes a type of explanation that is neither evidence- nor example-based. Lastly, we found that almost all of the CAs offer the option to transfer to a human associate.

Step 6e: Group characteristics into dimensions to create taxonomy. In this iteration, we add two characteristics, the 'exerted' acknowledgement and the 'capability-based' type of explanation. In addition, we add the dimension 'delegation to human,' to cover the human back-up solution. The taxonomy after this iteration is therefore as follows: $T_3 = \{$Acknowledgment (Apologetic, Exerted, Neutral, None), Explanation (Positive, Negative, None), Type of Explanation (Utterance-based, Intent-based, Capability-Based, None), Strategy Type (System-repair, Limited Assist. User-Repair, Assist. User-repair, User-repair), Delegation to Human (Yes, No)$\}$

Step 7: Ending Conditions met? The objective ending conditions are not met as we added new characteristics and a dimension. After this iteration, we consider this taxonomy to be comprehensible as it now covers both research and practice instances. Nonetheless, we need at least one more iteration due to the objective ending conditions.

3.4 Iteration 4

Step 3: Conceptual-to-Empirical. We take this approach since we have now examined various objects and have a deeper knowledge of the recovery strategies.

Step 4c: Conceptualize new characteristics and dimensions of objects. To clarify the dimensions 'explanation' and 'type of explanation,' we rename the latter to 'reason.' Respectively, the characteristics are modified to match the naming. The 'strategy type' dimension is renamed to 'repair.' Lastly, we reorder the dimensions to reflect the chronological order of occurrence during a breakdown.

Step 5c: Examine objects for these characteristics and dimensions. We examined all objects from the previous iterations and they all fit in the new taxonomy.

Step 6c: Create taxonomy. The taxonomy is as follows: $T_4 = \{$Acknowledgment (Apologetic, Exerted, Neutral, None), Reason (Utterance, Intent, Capability, None), Explanation (Positive, Negative, None), Repair (System-repair, Limited Assist. User-Repair, Assist. User-repair, User-repair), Delegation to Human (Yes, No)$\}$

Step 7: Ending Conditions met? Since we only renamed/reordered, the objective ending conditions are met. As the subjective ending conditions were already met after the last iteration, the development process ends here.

4 Taxonomy

Our final taxonomy consists of five dimensions with two to four characteristics each. In this section, we present the dimensions and characteristics in more detail. A classification of all objects from the development process is provided in Table 2 where each row represents one recovery strategy identified in the source.

Table 2. Final Taxonomy and Classification of All Objects Examined

Conversational Agent	ACK				REA				EXP			REP				DH	
	Apologtic	Exerted	Neutral	None	Utterance	Intent	Capability	None	Positive	Negative	None	System	Lim. Ass. User	Ass. User	User	Yes	No
Count	23	4	21	12	12	9	6	33	10	7	43	10	12	13	25	9	51
Ashktorab et al., 2019				X				X			X	X					X
			X					X			X				X		X
			X			X			X				X				X
			X			X			X				X				X
	X							X			X	X				X	
			X		X				X					X			X
	X				X					X				X			X
Beneteau et al., 2019				X				X			X	X					X
	X							X			X				X		X
			X			X			X				X				X
Cheng et al., 2018				X				X			X				X		X
Cuadra et al., 2021	X							X			X				X		X
Diederich et al., 2020			X		X						X				X		X
	X				X						X				X		X
Dippold, 2023				X				X			X	X					X
			X			X			X				X				X
	X				X					X					X		X
			X		X				X					X			X
	X						X			X				X		X	
Do et al., 2023	X						X		X				X				X
Han et al., 2021			X		X						X				X		X
			X		X						X		X				X
Huang and Dootson, 2022			X					X			X				X	X	
Khurana et al., 2021			X					X			X			X			X
	X					X				X				X			X
Law et al., 2022	X							X			X			X			X
	X							X			X				X		X
			X					X			X			X			X
			X					X			X				X		X
C.-H. Li et al., 2020	X					X					X				X		X
T. J.-J. Li et al., 2020			X			X			X				X				X
Liu et al., 2023	X				X						X				X		X
	X						X				X				X		X
Mahmood et al., 2022	X				X					X					X		X
Müller et al., 2021				X	X				X				X				X
				X	X						X				X		X
Poser et al., 2021	X					X				X					X	X	
Seeger and Heinzl, 2021			X					X			X		X				X
	X							X			X				X		X
Song et al., 2023	X							X			X				X		X
		X						X			X				X		X
Wu et al., 2021			X					X			X			X			X
Zargham et al., 2022				X				X			X	X					X

X. Zhang et al., 2023	X							X			X				X		X
				X				X			X				X		X
Retail Company	X							X			X				X		X
		X				X			X				X			X	
		X						X			X	X				X	
Transportation Company		X						X			X				X		X
			X					X			X				X		X
	X						X			X				X			X
				X				X			X	X					X
Consulting Company				X				X			X		X				X
	X						X				X			X			X
Automotive Retail Company			X				X				X		X			X	
				X				X			X	X					X
IT Company			X					X			X			X			X
				X				X			X	X					X
			X					X			X	X				X	
	X							X			X			X		X	

Legend (Dimensions): ACK: Acknowledgment • **EXP**: Explanation • **REA**: Reason • **REP**: Repair • **DH**: Delegation to human

Acknowledgment (ACK). The dimension 'acknowledgment' covers how the CA communicates the breakdown towards the user. If the CA acknowledges the breakdown, it can be done in different tones, such as being apologetic (Lee et al., 2010; Mahmood et al., 2022) or neutral (Law et al., 2022). Apologetic acknowledgments usually start with "*I'm sorry*" and have a friendly tone in the remainder of the response. In productive CAs, we often encountered the 'exerted' tone which conveys the CA's willingness to help in a friendly, but unapologetic tone. An exerted tone also includes any efforts from the CA to maintain a human-like, trustful relationship with the user (Song et al., 2023). The choice of acknowledgment can express different levels of anthropomorphism and convey a certain the CA personality.

Reason (REA). The dimension 'reason' classifies the breakdown reason that is communicated to the user. Therefore, it is a high-level classification rather than a breakdown of all possible failures to make the taxonomy explanatory. *Utterance* includes all breakdowns that are triggered because the utterance from the user could not be parsed because of reasons like a foreign language or unknown vocabulary. *Intent* covers breakdowns that happen when an utterance cannot be assigned to an intent with sufficient confidence. It can also happen that an intent was matched with a high confidence which is not in-scope of the CA, so the reason for the breakdown is the *capability*. This happens in goal-oriented CA (e.g., customer service) built on general purpose frameworks that contain a predefined set of unused intents or when a CA is not performing its designated tasks correctly. The communicated reason does not need to match the real reason for the breakdown to conceal quality issues that might have other negative effects. For instance, the CA can communicate that the intent is out of scope (capability) although the issue is that the user has entered an utterance in a non-supported language.

Explanation (EXP). The CA can optionally add an explanation on why the breakdown occurred. A positive explanation explains what the it *has* understood, a negative explanation shows why it was *not* able to understand the user input utterance. Explanation means that the reason of failure needs to be elaborated in a way that is helpful to

the user. For instance, "*I am sorry that I was not able to understand your question*" (Law et al., 2022, p. 5) does not qualify as an explanation, but "*I couldn't recognize the Predict Trend Function*" (Khurana et al., 202, p. 3) does.

Repair (REP). The repair dimension classifies recovery strategies based on the type of their repair mechanism. That is, how the CA recovers after a breakdown. We differentiate between a system-repair, a limited assisted user-repair, an assisted user-repair, and the complete user-repair. A *system-repair* is a repair mechanism that does not involve the user in the process. For instance, the CA assumes the best matching intent despite a low confidence and executes consequential actions without user involvement or confirmation. The *limited assisted user-repair* has a low involvement of the user by limiting the functional range. Examples are providing options that are most likely to match the user's intent or asking the user to confirm an intent. In these cases, the options provided by the CA are the only way moving forward in the conversation, the user is not allowed to 'change the topic.' Hence, it is a 'limited' assisted user-repair. *Assisted user-repair*, in contrast, allows for the user to repair the breakdown more freely with the assistance of the CA in the form of hints, examples, and other guidance. The user can rephrase or correct the understanding with no limitations in functionality. The assistance for the user-repair can be transported in the form of an explanation of the breakdown, an example utterance, or other information that helps the user to repair. *User-repair* includes all recovery strategies that rely on the user to repair the breakdown. That is, the CA does not perform any valuable action, and does not propose any way to help the user achieve their intents. Yet, the CA can restart or continue the conversation in a way that the user has the opportunity to send a different utterance.

Delegation to Human (DH). Some CAs are backed by a human associate who can takeover upon breakdown. This dimension indicates whether the recovery strategies includes handing over the request to a human (*yes*/*no*). The handover does not need to be automatic, but can also be offered as one of the options to the user.

5 Discussion

5.1 Findings & Recurring Recovery Strategies

The most part of the recovery strategies (48/60) prescribe to admit a breakdown and interact with the user to resolve the issue instead of making assumptions. However, only a third of the strategies include apologetic behavior (23/60) although a CA has anthropomorphic traits by its conversational nature. We observe a discrepancy between research and practice regarding the explanation of the breakdown. While half of the strategies in research (23/45) provide reasons and a third even detailed explanations (15/45), we rarely encounter this in productive CAs (4/15 and 2/15, respectively). As productive CAs tend not to explain the breakdown, it is intransparent to the user what exactly failed, providing no assistance to for the user to repair. Productive CAs often refer to a human as a fallback solution, while in research, they are rarely examined. Our tests reveal that the productive CAs often cannot perform other tasks outside the purpose the were built for, thus have very limited functional ranges that are tailored to

specific customer support queries. Regardless, most of the productive CA are still capable of handling utterances with low semantic value and smalltalk.

The most prevalent three patterns we observe among the recovery strategies in the CAs examined are rather 'simple:' With seven mentions, one of the highest recurring is the 'ignore the breakdown and move on' strategy. This means that the CA simply ignores that a breakdown has occurred, does not provide any kind of acknowledgment or even explanation towards the user, and performs the next action that it considers as the best suitable for this case. This pattern reoccurred in both literature (Ashktorab et al., 2019; Beneteau et al., 2019; Dippold, 2023; Zargham et al., 2022) and in the productive instances. For the productive CAs, we cannot know for sure if the breakdown was actively detected or if the action was performed with no special mitigation strategy. Tied in occurrence with the ignorance strategy is the recovery strategy to 'apologize and let the user fix it.' The CA does acknowledge the breakdown, but provides no further explanation or assistance. The user is left to resolve the issue on their own, e.g. by rephrasing their request. This recovery strategy was mostly prevalent in literature with six out of seven times (e.g., Beneteau et al., 2019, Cuadra et al., 2021; Law et al., 2022). Third ranks the recovery strategy to 'neutrally acknowledge the breakdown counterquestion/narrow down the space.' Examples for this recovery strategy is the introduction of buttons for the user to specify their request, or to ask a specific counterquestion ("Do you mean …?"). This recovery strategy was only found in literature (e.g., Ashktorab et al, 2019; Dippold, 2023; T. J.-J. Li et al., 2020).

5.2 Implications for Research and Practice

We developed a taxonomy for recovery strategies based on CAs in productive use and studies in literature. A taxonomy provides a common language to describe recovery strategies and helps to identify research gaps, which we will discuss in the following.

Based on the prevalence of studies on the effect of anthropomorphism in research (Diederich et al., 2021; Seeger and Heinzl, 2021; X. Zhang et al., 2023), it is surprising that for breakdown recovery, the usage of apologetic behavior is that low. This can be in line with the observation that a chatbot's apology is not perceived as a sincere apology (J. Zhang et al., 2023) and therefore omitted for efficiency reasons.

While the recovery strategies in this taxonomy focus on single breakdowns, the repair may not be successful, and lead to subsequent breakdowns. Based on previous findings, three consecutive breakdowns can already lead to users abandoning the CA (C.-H. Li et al., 2020). However, the combination of different recovery strategies for consecutive breakdowns have not yet been tested to our knowledge. This raises the question: Which recovery strategy *pattern* is the most effective to counteract the impact of breakdowns? Additionally, what impact does the choice of recovery strategies have on the relationship between the CA and the user? We suspect that, similar to human-human interaction, the answers to the questions have many influencing factors, including the CA personality, the user personality, and the purpose of the CA. Human-like communication has shown to increase the perceived enjoyment (de Sá Siqueira et al., 2023), so it is probable that anthropomorphism can help in mitigating breakdowns.

Before the background of the recovery strategies grounding our taxonomy, we can observe a discrepancy in the strategies investigated in literature and in productive CAs, as described in Section 5.1. This calls for the question to investigate if the recovery strategies tested in theory and in experimental setups are not feasible or suitable for real-world use cases, or if external influencing factors participate in the design decision that do not exist or are not replicable in research experiment settings.

During the development of our taxonomy, we discover clear potentials in productive CA. As current CAs provide little transparency on the reason of a breakdown, it is not clear to the user what had caused it. However, as research as shown, users prefer to have evidence of why the CA was not able to process their request appropriately (Ashktorab et al., 2019). Therefore, we call for more transparency in productive CA.

Although automation is one of the reasons for a CA, the CAs' authority to perform tasks needs to be carefully chosen. Executing the 'best matching action' despite low confidence on the intent recognition without user confirmation can lead to negative consequences like users giving up (Luger and Sellen, 2016). It is crucial to choose the appropriate recovery strategy not only when a breakdown occurs, but to also adjust timings of potential breakdown points. A breakdown at a later stage as well as a handover to a human processor can lead to user aggression (Huang and Dootson, 2022; Kim et al., 2023). We recommend to make more use of explicit user confirmations, which can be tedious due to the extra conversation turn, but are perceived as polite and meaningful by the user (Ashktorab et al., 2019). A superfluous repair attempt when there is nothing to repair may put the CA capabilities in question. However, the "*positive impact of self-repair in the wake of an error outweighs the negative impact of overcorrection.*" (Cuadra et al., 2021, p.27)

6 Limitations and Outlook

Although we included recovery strategies that originate in research with voice-based and/or embodied CAs, we only considered properties that are applicable to text-based CAs. It is clear that the taxonomy can be extended to include different modalities contained in the recovery, such as gestures (Mori et al., 2020) and physical design (Kontogiorgos et al., 2020). However, our taxonomy can be applied and is extensible as per the qualitative requirement so further dimensions can be added as needed. This can be investigated in future research in the context of robotic service agents. Second limitation of this work is that in the literature review underlying this taxonomy development, we only include studies that focus on breakdown recovery. Due to the large body of knowledge around CAs in general, it is probable that there are many more studies that contain recovery strategies not yet included in this taxonomy which can be examined for potential future development of this taxonomy. Lastly, we have only examined a small number of productive CAs due to the reasons given in Section 3.3. With the emergence of business use cases for large language models, there will be further evaluation needed to update and extend this taxonomy including more CAs in productive use, with a particular focus on new types of CAs like ChatGPT.

References

Akhtar, M., Neidhardt, J. & Werthner, H. (2019), 'The Potential of Conversational agents: Analysis of Conversational agent Conversations,' *in* IEEE Conference on Business Informatics, pp. 397-404.

Almansor, E. H. & Hussain, F. K. (2021), 'Fuzzy Prediction Model to Measure Chatbot Quality of Service,' *in* Proceedings of the 2021 IEEE International Conference on Fuzzy Systems.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R. & Horvitz, E. (2019), 'Guidelines for Human-AI Interaction,' *in* CHI Conference on Human Factors in Computing Systems.

Ashktorab, Z., Jain, M., Liao, Q.V. & Weisz, J.D. (2019), 'Resilient Conversational agents: Repair Strategy Preferences for Conversational Breakdowns,' *in* CHI Human Factors in Computing Systems, no. 254.

Baylor, A. L. & Kim, Y. (2005), 'Simulating Instructional Roles Through Pedagogical Agents,' *International Journal of Artificial Intelligence in Education* 15(2), pp. 95-115.

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J. & Hiniker, A. (2019), 'Communication Breakdowns Between Families and Alexa,' *in* CHI Conference on Human Factors in Computing Systems.

Benner, D., Elshan, E., Schöbel, S. & Jansen, A. (2021), 'What Do You Mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents,' *in* Proceedings of the 2021 International Conference on Information Systems, no. 13.

Benyon, D., Gambäck, B., Hansen, P., Mival, O. & Webb, N. (2013), 'How Was Your Day? Evaluating a Conversational Companion,' *IEEE Transactions on Affective Computing* 4(3), pp. 299-311.

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,' *in* Hawaii International Conference on System Sciences, pp. 284-293.

Cheng, Y., Yen, K, Chen, Y., Chen, S. & Hiniker, A. (2018), 'Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies,' *in* Proceedings of the 17th ACM Conference on Interaction Design and Children, pp. 337-348.

Chui, M., Roberts, R. & Yee, L. (2022), 'Generative AI is Here: How Tools Like ChatGPT Could Change Your Business,' *Quantum Black AI by McKinsey*.

Cuadra, A., Li, S., Lee, H., Cho, J. & Ju, W. (2021), 'My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction,' *Proceedings of the ACM Human-Computer Interaction* 5(CSCW1).

de Sá Siqueria, M. A., Müller, B. C. N. & Bosse, T. (2023), 'When Do We Accept Mistakes From Conversational agents? The Impact of Human-Like Communication on User Experience in Conversational agents That Make Mistakes,' *International Journal of Human-Computer Interaction*.

Diederich, S., Lembcke, T.-B., Brendel, A. B. & Kolbe, L. M. (2020), 'Not Human After All: Exploring the Impact of Response Failure on User Perception of Anthropomorphic Conversational Service Agents,' *in* Proceedings of the 2020 European Conference on Information Systems, no. 110.

Diederich, S., Lembcke, T.-B., Brendel, A. B. & Kolbe, L. M. (2021), 'Understanding the Impact that Response Failure Has on How Users Perceive Anthropomorphic Conversational Service Agents: Insights From an Online Experiment,' *AIS Transactions on Human-Computer Interaction* 13(1), pp. 82-103.

Dippold, D. (2023), '"Can I Have the Scan on Tuesday?" User Repair in Interaction with a Task-Oriented Chatbot and the Question of Communication Skills for AI,' *Journal of Pragmatics* 204, pp. 21-32.

Do, H. J., Kong, H.-K., Tetali, P., Lee, J. & Bailey, B. P. (2023), 'To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions,' *ACM Human-Computer Interaction* 7(CSCW1), no. 99.

Feng, S. & Buxmann, P. (2020), 'My Virtual Colleague: A State-of-the-Art Analysis of Conversational Agents for the Workplace,' *in* Hawaii International Conference on System Sciences, pp. 156-165.

Fitzpatrick, K. K., Darcy, A. & Vierhile, M. (2017), 'Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial,' *JMIR Mental Health* 4(2), e19.

Han, E., Yin, D. & Zhang, H. (2021), 'Interruptions During a Service Encounter: Dealing with Imperfect Chatbots,' *in* Proceedings of the 2021 International Conference on Information Systems.

Huang, Y.-S. S. & Dootson, P. (2022), 'Conversational agents and Service Failure: When Does it Lead to Customer Aggression,' *Journal of Retailing and Consumer Services* 68, no. 103044.

Inkster, B., Sarda, S. & Subramanian, V. (2018), 'An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study,' *JMIR mHealth and uHealth* 6(11), no. e12106.

Khurana, A., Alamzadeh, P. & Chilana, P. K. (2021), 'ChatrEx: Designing Explainable Conversational agent Interfaces for Enhancing Usefulness, Transparency, and Trust,' *in* IEEE Symposium on Visual Languages and Human-Centric Computing.

Kim, A., Yang, M. & Zhang, J. (2023), 'When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms,' *ACM Transactions on Computer-Human Interaction* 30, no. 14.

Kitchenham, B. (2004), *Procedures for Performing Systematic Reviews*. Keele University.

Klopfenstein, L. C., Delpriori, S., Malatini, S. & Boglioli, A. (2017), 'The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms,' *in* Designing Interactive Systems Conference, pp. 555-565.

Kontogiorgos, D., Pereira, A., Sahindal, B., van Waveren, S. & Gustafson, J. (2020), 'Behavourial Responses to Robot Conversational Failures,' *in* Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 53-62.

Law, E. L.-C., Følstad, A. & van As, N. (2022), 'Effects of Humanlikeness and Conversational Breakdown on Trust in Conversational agents for Customer Service,' *in* NordiCHI Nordic Conference on Human-Computer Interaction.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S. & Rybski, P. (2010), 'Gracefully Mitigating Breakdowns in Robotic Services,' *in* ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Li, C.-H., Ye, 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 Conversational agent,' *in* CHI Conference on Human Factors in Computing Systems, no. 82.

Li, T. J.-J., Chen, J., Xia, H., Mitchel, T. M. & Myers, B.A. (2020), 'Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs,' *in* ACM Symposium on User Interface Software and Technology, pp. 1094-1107.

Liu, D., Lv, Y. & Huang, W. (2023), 'How Do Consumers React to Chatbots' Humorous Emojis in Service Failures,' *Technoloy in Society* 73, no. 102244.

Luger, E. & Sellen, A. (2016), '"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents,' *in* CHI Conference on Human Factors in Computing Systems, pp. 5286-5297.

Mahmood, A., Fung, J. W., Won, I. & Huang, C.-M. (2022), 'Owning Mistakes Sincerely: Strategies for Mitigating AI Errors,' *in* CHI Conference on Human Factors in Computing Systems.

Meyer von Wolff, R., Hobert, S., Masuch, K. & Schumann, M. (2020), 'Conversational agents at Digital Workplaces – A Grounded-Theory Approach for Surveying Application Areas and Objectives,' *Pacific Asia Journal of the Association for Information Systems* 12(2), pp. 64-102.

Mori, T., Jokinen, K. & Den, Y. (2021), 'On the Use of Gestures in Dialogue Breakdown Detection,' *in* Proceedings of the 24th Conference on the Oriental COCOSDA Internation Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques.

Müller, R., Paul, D. & Li, Y. (2021), 'Reformulation of Symptom Decsriptions in Dialogue Systems for Fault Diagnosis: How to Ask for Clarification,' *International Journal of Human-Computer Studies* 145, no. 102516.

Nickerson, R. C., Varshney, U. & Muntermann, J. (2012), 'A Method for Taxonomy Development and its Application in Information Systems,' *European Journal of Information Systems* 22, pp. 336-359.

Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N. & Cassell, J. (2012), '"Oh, dear Stacy!" Social Interaction, Elaboration, and Learning with Teachable Agents,' *in* CHI Conference on Human Factors in Computing Systems.

Poser, M., Singh, S. & Bittner, E. A. C. (2021), 'Hybrid Service Recovery: Design for Seamless Inquiry Handovers Between Conversational Agents and Human Service Agents,' in *Proceedings of the 54th Hawaii Conference on System Sciences*, pp. 1181-1190.

Prahl, A & Goh, W. W. P. (2021), '"Rogue Machines" and Crisis Communication: When AI Fails, How Do Companies Publicly Respond?' *Public Relations Review* 47(4), no. 102077.

Rapp, A., Curti, L. & Boldi, A. (2021), 'The Human Side of Human-Conversational agent Interaction: A Systematic Literature Review of Ten Years of Research on Text-Based Conversational agents,' *International Journal of Human-Computer Studies* 151, no. 102630.

Reinkemeier, F. & Gnewuch, U. (2022), 'Designing Effective Conversational Repair Strategies for Conversational agents,' *in* European Conference on Information Systems.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G. & Söllner, M. (2020), 'Machines as Teammates: A Research Agenda on AI in Team Collaboration,' *Information & Management* 57(2), no. 103174.

Seeger, A.-M. & Heinzl, A. (2021), 'Chatbots Often Fail! Can Anthropomorphic Design Mitigate Trust Loss in Conversational Agents for Customer Service?' *in* Proceedings of the 2021 European Conference on Information Systems, no. 12.

Shawar, B. A. & Atwell, E. (2007), 'Conversational agents: Are They Really Useful?,' *LDV-Forum* 22(1), pp. 31-50.

Song, M., Zhang, H., Xing, X. & Duan, Y. (2023), 'Appreciation vs. Apology: Research on the Influence Mechanism of Chatbot Service Recovery Based on Politeness Theory,' *Journal of Retailing and Consumer Services* 73, no. 103323.

Suárez-Gonzalo, S., Mas-Manchón, L. & Guerrero-Solé, F. (2019), 'Tay is You. The Attribution of Responsibility in the Algorithmic Culture,' *Observatorio* 13(2), pp. 1-14.

Wakefield, J. (2016), 'Microsoft conversational agent is taught to swear on Twitter,' *BBC*. URL: https://www.bbc.com/news/technology-35890188/ (visited on November 08, 2022).

Wang, K.-Y., Chih, W.-H. & Honora, A. (2023), 'How Emoji Use in Apology Messages Influences Customers' Responses in Online Service Recoveries: The Moderating Role of Communication Style,' *International Journal of Information Management* 69, no. 102618.

Webster, J. & Watson, R.T. (2002), 'Analyzing the Past to Prepare for the Future: Writing a Literature Review,' *Management Information Systems Quarterly* 26(2), pp. xiii-xxiii.

Weizenbaum, J. (1966), 'ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine,' *Computational Linguistics* 9(1), pp. 36-45.

Wolf, M. J., Miller, K. W. & Grodzinsky, F. S. (2017), 'Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications,' *The ORBIT Journal* 1(2), pp. 1-12.

Wu, M.-H., Yeh, S. F., Chang, X. & Chang, Y.-J. (2021), 'Exploring User's Preferences for Conversational agent's Guidance Type and Timing,' *in* ACM Conference On Computer-Supported Cooperative Work and Social Computing, pp. 191-194.

Zargham, N., Pfau, J., Schnackenberg, Z. & Malaka, R. (2022), '"I Didn't Catch That, But I'll Try My Best': Anticipatory Error Handling in a Voice Controlled Game,' *in* Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, no. 153.

Zhang, J., Zhu, Y., Wu, J. & Yu-Buck, G. F. (2023), 'A Natural Apology is Sincere: Understanding Chatbots' Performance in Symbolic Recovery,' *International Journal of Hospitality Management* 108, no. 103387.

Zhang, Y., Lee, S. K., Kim, W. & Hahn, S. (2023), '"Sorry, it was my Fault": Repairing Trust in Human-Robot Interactions,' *International Journal of Human-Computer Studies* 175, no. 103031.