Break, Repair, Learn, Break Less: Investigating User Preferences for Assignment of Divergent Phrasing Learning Burden in Human-Agent Interaction to Minimize Conversational Breakdowns

Mina Foosherian
BIBA - Bremer Institut für Produktion und Logistik GmbH
Bremen, Germany
fos@biba.uni-bremen.de

Samuel Kernan Freire
Delft University of Technology, Faculty of Industrial Design Engineering
Delft, Netherlands
s.kernanfreire@tudelft.nl

Evangelos Niforatos
Delft University of Technology, Faculty of Industrial Design Engineering
Delft, Netherlands
e.niforatos@tudelft.nl

Karl A. Hribernik
BIBA - Bremer Institut für Produktion und Logistik GmbH
Bremen, Germany
hri@biba.uni-bremen.de

Klaus-Dieter Thoben
University of Bremen, Faculty of Production Engineering
Bremen, Germany
tho@biba.uni-bremen.de

ABSTRACT

Conversational agents (CA) occasionally fail to understand the user's intention or respond inappropriately due to natural language complexity. These conversational breakdowns can happen because of low intent and entity prediction confidence scores. A promising repair strategy in such cases is that the CA proposes to users likely alternatives to proceed. If one of these options matches the user's intention, the breakdown is repaired successfully. We propose that successful repairs should be followed by a learning mechanism to minimize future breakdowns. After a successful repair, the CA, user, or both can learn each other's specific phrasing. This prevents similar phrasings from causing reoccurring breakdowns. We compared user preferences for these learning mechanisms in a scenario-based study with manufacturing workers (N = 26). Our result showed that users first prefer to share the learning burden with the CA (61.3%), followed by entirely outsourcing the learning burden to the CA (60.7%) as opposed to themselves.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *User studies.*

KEYWORDS

Conversational Agents, User Experience, Error Handling, Conversational Breakdown, None-progress, Learning

MUM 2022, November 27–30, 2022, Lisbon, Portugal

ACM ISBN 978-1-4503-9820-6/22/11.
https://doi.org/10.1145/3568444.3568454

ACM Reference Format:
Mina Foosherian, Samuel Kernan Freire, Evangelos Niforatos, Karl A. Hribernik, and Klaus-Dieter Thoben. 2022. Break, Repair, Learn, Break Less: Investigating User Preferences for Assignment of Divergent Phrasing Learning Burden in Human-Agent Interaction to Minimize Conversational Breakdowns. In *21th International Conference on Mobile and Ubiquitous Multimedia (MUM 2022), November 27–30, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3568444.3568454

1 INTRODUCTION

The latest advances in Artificial Intelligence (AI) made the integration of Conversational Agents (CA) increasingly popular in different areas such as smart home, e-Commerce, and customer service. Nevertheless, new application areas, such as manufacturing industry use cases, are continuously emerging. Companies that adopt new information and communication technologies in production can use CAs to empower workers improving task performance, decision-making, and interacting with machine-generated complex data [46]. The interaction with CAs is through natural conversations and could, for instance, reduce the costs of training workers in using multiple graphical user interfaces [17], among other benefits such as mobile assistance, permanent and central accessibility, and speed [46]. CAs can be text-based agents (Chatbots), Voice-User interfaces (VUI), or Embodied-dialog Agents (EDA) [19]. In manufacturing, adopting a chatbot might not be practical for situations where workers are wearing gloves or need to use their hands and eyes for their work tasks. A VUI can address this barrier and enable eyes and hands-free interaction [46]. Speech recognition, however, can be adversely affected by noise in the environment.

Essentially, a CA understands natural language, decides how to respond, and communicates its final response. This "intelligence" relies on the CA's training data and logic used to create its natural language understanding (NLU) and dialog management models [19]. A CA's ability to interpret user input accurately and return adequate responses is crucial to its user experience and trust [14, 29, 32]. In CAs, intent refers to the goal the user tries to accomplish with an

utterance. User utterances can contain one or more data fields on defined semantic types (entities) that the CA extracts to respond accurately to the user's request. Considering that users talk to the CA in a free way with their own words, the CA requires a large corpus of training data, including different forms of phrasing for every intent to interpret various user expressions and respond adequately [41]. CAs' developers can use human-human chat or call transcripts to create extensive training data containing divergent phrasings from users in some application areas, such as customer service [15]. However, this is not the case for applications such as industrial use cases where no conversation transcripts are available to create a large corpus of training data.

Despite recent technological advances, due to the immense complexity of natural language, CAs are prone to so-called "*conversational breakdown*", indicating that the CA did not correctly understand the user's utterance or responded inadequately to the user's request [16, 29]. Breakdowns are unavoidable situations in human-machine conversations [2] and occur for various reasons such as errors during intent and entity recognition, errors during task completion, errors in generating the response, and users' lack of familiarity with CA's intents [23].

Conversational breakdown can lead to frustration, disappointment, mistrust, and dissatisfaction [4, 11, 26] if not addressed. To ensure an engaging conversational interaction, it is essential to minimize the number of breakdowns [28, 35] and recover from them by applying so-called "*repair*" strategies. Repair refers to various natural conversation methods to resolve troubles in speaking, hearing, or understanding [39], such as repeating or paraphrasing all or parts of the source of trouble in a prior turn [29]. Successful repair is especially critical for task-oriented CAs because it allows users to continue performing their tasks [2]. Although a successful repair increases user satisfaction, multiple breakdowns in a conversation can still lead to users abandoning the CA [22]. We argue that relying on the current technological advances, a successful repair should be followed by a learning mechanism to minimize future breakdowns. After a successful repair, the user, the CA, or both can learn the specific phrasing of each other. This prevents similar phrasings from causing reoccurring breakdowns.

Addressing learning mechanisms to minimize conversational breakdowns in industrial use cases is particularly essential for three reasons. Firstly, the CAs' developers have limited resources to create high-quality and comprehensive training data in the early stages. Secondly, they must include domain-specific terminologies to train the CA. Thirdly, workers could use diverse natural language expressions instead of domain-specific and acknowledged terms (jargon).

This paper uses three real scenarios from a task-oriented CA developed for workers in manufacturing. They feature cases where knowledge mismatch between CAs and workers results in errors during intent and entity recognition, ultimately causing conversational breakdowns. This study introduces and evaluates user preferences for three ways of distributing the burden of learning divergent phrasing from successfully repaired conversational breakdowns to minimize their recurrence, namely:

- Learning burden on the CA: The CA tries to learn the phrasing of the user and add it to its training data for the successfully matched intent and entities during the repair.
- Learning burden on the user: The CA asks the user to adapt their phrasing to the terminology suggested by the CA during the repair.
- Learning burden shared between the CA and the user: The CA tries to learn the phrasing of the user but asks the user to try and use the suggested terminology in the future.

We then debate each approach's benefits and drawbacks and suggest future research directions.

2 RELATED WORK

2.1 Breakdown and Repair in Human-Agent Interaction

Breakdown and repair refer to particular events occurring during conversations between the user and the CA. McTear et al. pointed out that in human-agent interaction, "*conversation*" is an action where a speaker utters something to achieve goals, and the addressee interprets these actions [27]. The conversation parties take turns during the conversation and use grounding to reduce the risk of misunderstanding. Grounding is the process of updating common ground and frequently occurs in conversations between humans [9]. Sometimes, situations occur where one party misses what the other said, fails to understand it, or realizes a misunderstanding [29]. Literature refers to such a situation as a "*conversational breakdown*" [27]. There are various methods in natural conversation to resolve such breakdowns. Conversation analysts use the term "*repair*" to refer to the range of practices that we have for managing troubles in speaking, hearing, or understanding [39], such as repeating or paraphrasing all or parts of the source of trouble in a prior turn [29]. Each conversation party may initiate repair to attempt to recover from the breakdown [38]. Nevertheless, the effort to repair can succeed or fail.

We consider breakdowns to be unavoidable events during human-agent conversations, similar to the human-human conversation [49]. It is common to define breakdown in CAs with an intent-based, probabilistic model as when the classifier's confidence scores for all intents are below a curtain threshold [2]. In such a case, the CA typically initiates a repair with a fallback utterance such as "*I am not sure what you mean. Could you please rephrase your request?*" (false negative - non-recognition). Another approach is to ignore the low intent recognition confidence level and respond to the most likely intent anyway (false positive, misrecognition) [2, 16, 22, 27]. Both of these strategies can potentially lead to another breakdown and user frustration [16]. Previous work [2, 3, 21, 23, 45] investigated which repair strategies users prefer CAs to adopt to repair conversational breakdown in different contexts. Ashktorab et al. [2] found that users prefer the CA to suggest likely alternatives in cases where prediction confidence falls below the threshold and let them decide which one is correct. This strategy is called "*options*". Users' preference for options repair strategy was independent of the repair outcome. As a follow-up to Ashktorab et al.'s study, Følstad et al. found that expressing uncertainty and suggesting alternatives could significantly reduce false positives but did not have a major impact on the dialogue process and outcome in dialogues with a customer service chatbot [16]. Their study concluded that such a repair mechanism does not replace the necessity of continuous training data improvement to minimize breakdowns.

2.2 Learning from Conversations in Human-Agent Communication

The performance of human-agent communication can be improved in two ways. The first one focuses on enhancing the agent's NLU capabilities, and the second one concerns how users learn to communicate with the agent to compensate for its limited understanding [3, 22].

2.2.1 CAs learning from conversations with humans. Several strategies exist to improve a CA's NLU component through machine learning. One approach is to review user conversations and add the utterances that caused breakdowns to the CA's training data. This process is called conversation-driven development (CDD) [31] and requires developers to provide an early CA version to testers. Once testers start interacting with the CA, developers review these conversations and derive new or improved training data to build an NLU model. However, reviewing conversations or generating adversarial queries [44] and manually applying improvements is highly resource-intensive. Although this strategy can be effective during the early stages of development to correct significant errors, it scales poorly. A more scalable option is for the CA to identify conversational breakdowns itself using, for instance, sentiment analysis or by asking the user for feedback [18] to automatically flag breakdowns such as false positive responses. Then the CA can attempt to repair the conversation and collect information from the user to learn from [25]. The learning process above could occur automatically or with a human-in-the-loop to supervise the process. The latter may be necessary to avoid the accumulation of biased training data.

2.2.2 Humans learning from conversations with CAs. Learning from interacting with a CA results in a more subtle understanding of how to communicate with minimum breakdowns. Often, humans do not know how to structure their speech with a CA [30, 50]. This lack of knowledge can cause the user not to complete a task [47] or make several unsuccessful attempts [10]. Previous work with voice user interfaces has shown that users do not expect assistants to understand natural language [5]. Accordingly, users automatically adapt their communication strategies to align with the assistant's capabilities, for example, by using specific keywords. In the manufacturing context, the terminology used by workers can vary from person to person (e.g., they use different terms and acronyms for machines) [13]. CA repair strategies such as "*options*" could provide information to the user that facilitates the adaptation process by presenting domain-specific terminologies used to train the CA.

2.3 Adoption of Conversational Agents in manufacturing industry

The application of CAs in the manufacturing industry is an emerging topic for academia and corporate software providers. In the manufacturing industry, CAs are commonly used to assist workers during information-intensive or time-consuming tasks, providing benefits such as mobility, the delegation of tasks, and rapid data analysis [46].

Several software vendors, such as Oracle Digital [34], SAP [37], and SPIX Industry [42] are targeting use cases in manufacturing. Their solutions typically rely on a CA with connectors to existing business software. These software can possess data in several technical disciplines. One example of such software is quality control software, where manufacturers collect and analyze data about products and processes. Business software typically use domain-specific terminology. These terminologies can be internationally standardized or specific for a country, industry sector, company, factory, or department. CA developers use these terminologies to create training data and lookup tables that enable the CA to understand user utterances, validate extracted entity values, and build responses. However, workers use diverse natural language expressions to interact with the CA. Therefore, knowledge mismatch between CAs and workers is likely a cause for a manufacturing CA's breakdown. An important factor contributing to this behavior is that CA developers cannot write training data extracted from human-human conversations as they can in customer service, sales, and banking. A CA for workers in manufacturing does not replace a worker but provides a tool to assist or enhance workers.

3 APPROACH

The scope of this paper is on **learning mechanisms** applied after **successfully** repaired conversational breakdowns in human-agent interaction. The learning mechanisms are introduced only for this situation because an unsuccessful outcome indicates that the conversation parties could not restore a common understanding. While humans can learn from their failures, we argue that a CA cannot learn from unsuccessful repairs. However, a CA can be programmed to learn upon a successful repair. For example, a CA with an intent-based, probabilistic model can add the user utterance, which caused the breakdown, to its training data for the matched intent.

This paper demonstrates three real scenarios from manufacturing where divergent phrasing and knowledge mismatch between the CA and the workers regularly cause breakdowns. In such cases, we propose a CA that attempts to repair the breakdown by acknowledging understanding problems and suggesting its best matches to user utterance, so called "*options*" strategy. The user can either select one of the suggested options or deny them by saying something else. If the user does not select any of the suggested options and provides another utterance, the CA attempts to understand it and respond. If the intent or entity recognition has low confidence, the agent applies the "*options*" strategy again. If one of the options offered by the CA is a correct match - i.e. the user selects it - the breakdown is successfully repaired. The reason for selecting the"*options*" strategy is that it provides the user information that facilitates the learning process by presenting phrasings and domain-specific terminologies used to train the CA. We argue this successful repair should be followed by a learning mechanism to minimize the breakdown caused by a similar user phrasing in the future. We evaluate user preferences for three learning mechanisms where the CA assigns the burden (responsibility) of learning to (a) itself, (b) the user, or (c) both (shared) (Fig. 1). In addition, the CA explains the implications for the user based on who receives the burden. The party responsible for learning has to adjust their knowledge to understand or be understood by the other party.

A continuously learning CA will inevitably outperform a CA that does not learn [24], even if it starts with a significantly smaller training set and knowledge base. A continuously learning CA can learn

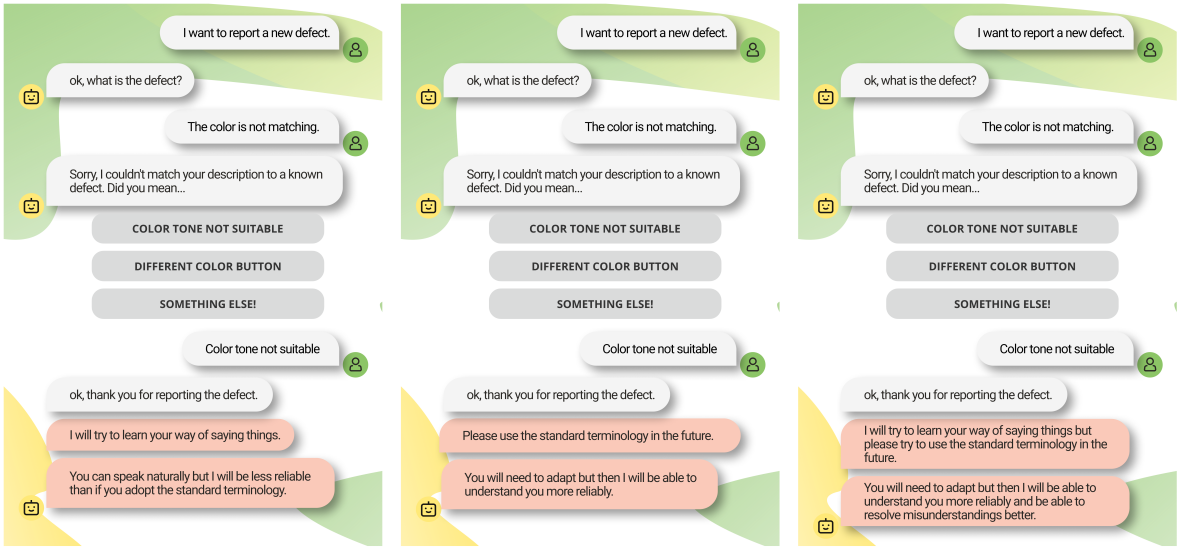


Figure 1: Assigning the burden of learning divergent phrasing which caused the breakdown to (a) the conversational agent, (b) the user, or (c) both the user and the conversational agent upon a successful repair - white-goods factory scenario

automatically or use a human-in-the-loop approach. Whichever learning strategy a CA uses, it is advisable to include a validation step as it may otherwise learn the user's bias, false information, and conflicts that need to be resolved [6]. Users will likely prefer to communicate naturally and rely on the continuously learning CA to adapt to them, as this requires less immediate effort. However, if the user also learns to adapt to the CA, they will benefit from even fewer conversational breakdowns. Furthermore, the development effort will be reduced. Ultimately, we expect the user experience to be best if both parties (CA and user) continuously learn. However, this may not be obvious to the users, or they may be unwilling to adapt.

We addressed the following research question in our scenario-based study:

- RQ: How do workers in the manufacturing sector prefer to treat the burden of learning divergent phrasing and domain-specific keywords from a repaired conversational breakdown to minimize its recurrence when interacting with a text-based conversational agent (CA)? (a) the CA should learn, (b) the user should learn, or (c) both the CA and the user should learn, and why?

We expect that users might prefer the CA to take over the learning burden as in the context of breakdown repair, users preferred CAs that contribute more to the repair process [2]. However, the unified theory of acceptance and use of technology (UTAUT) [43] states that the perceived usefulness of a technology must outweigh the perceived effort [48]. Therefore, if users perceive the benefits of adapting to the CA as higher than the perceived effort, they may be willing to take the learning responsibility. A CA with frequent unresolved breakdowns will likely cause the user to abandon the conversation due to poor user experience [22]. Furthermore, correctly interpreting users' intents and responding adequately is key to user experience [33]. Therefore, the users may prefer the shared learning burden scenario as it may result in the least breakdown occurrences.

4 STUDY

4.1 Design

To answer our research question, we took three examples of real conversational breakdowns during workers' interactions with the CA caused by knowledge mismatch between the conversation parties. The CA is implemented using Rasa[1], an open-source conversational AI framework. The agent has three skills to help workers perform tasks in three factories. Workers in each factory interact with the skill developed solely for their use case. For the purpose of this study, we focused on one task per factory: recording a product defect during quality control at a white-goods factory in Italy, reporting a machine issue at a detergent factory in the Netherlands, and reporting a product-quality issue at a detergent factory in Italy. We used three breakdown examples from actual conversations and combined them with learning mechanisms to create screenshots of this user study's scenarios.

In all scenarios, the CA recognized the intent correctly but had low confidence in entity extraction. The CA attempts to repair the conversation by expressing an understanding issue and suggesting the top two matches to user utterance from its training data and lookup tables. One of the suggested options is a match to user message in all scenarios to demonstrate a successful repair. At this point, the scenarios differ, namely in who is supposed to learn the

[1]https://rasa.com/

other's phrasing to prevent recurrences of breakdowns caused by a similar source in the future: the CA, the user, or both. To reduce the effect of conversation length on the dependent variables, we kept the number of conversation turns constant between scenarios; since the longer it takes to get to an answer, the less satisfied the user will be [1]. Figure 1 shows the screenshots created to demonstrate the learning mechanisms after a successful repair for the scenario related to the white-goods factory. The learning mechanisms for the other two factories' scenarios were similar.

4.2 Participants, Task, and Procedure

The participant recruitment process began with disseminating the study invitation through factories' research & development departments. Interested participants followed the link to the online study and proceeded on their own. Participation in this study was voluntary and uncompensated. A total of (N = 26) workers completed the survey.

The independent variable in this study is the assignment of divergent phrasing learning burden to avoid future conversational breakdowns with three possibilities: (a) on CA itself, (b) on the user, or (c) on both (shared). Each participant was exposed to all conditions via three pairwise comparisons (one pair at a time - Fig. 2) for the context related to their work factory: CA vs. User, Shared vs. User, and Shared vs. CA. This pairwise comparison method has already been used to determine the preferred repair strategy for chatbots [2]. As explained by Ashktorab et al., pairwise comparisons, compared to the Likert scale, can capitalize on simple judgments and a small set of stimuli [8, 12, 20]. To control the negative impact of learning effect, Latin Square design [36] was used to rotate the sequence of pairs and test scenarios. To draw participants' attention to the assignment of the learning burden, we highlighted relevant conversation turns (i.e., chat bubbles).

At the beginning of the study, we informed the participants that they would see three pairs of alternative conversations with a CA for the same manufacturing scenario, and in all conversations, there is a misunderstanding between the user and the CA, which is then resolved. We asked them to read the conversations, pay attention to the highlighted differences and select their preferred scenario.

We measured two key dependent variables: subjective satisfaction and the reason for the preference. After each comparison, participants could substantiate their decisions of one learning-burden category over another by selecting one or more of predefined reasons, such as (more) "supportive", "efficient", "easy", "clear", "exciting", "interesting", "inventive", "leading edge", and "other," where they could specify their own reason(s) (Fig. 3). The predefined characteristics were based on the short version of the User Experience Questionnaire (UEQ-S) [40].

5 RESULTS

To answer our research question, we applied the Bradley-Terry model [7] for ranking the preferences of the participants in terms of paired comparisons. We used the XLSTAT[2] statistical software to fit the model. We selected the best fit based on minimizing the

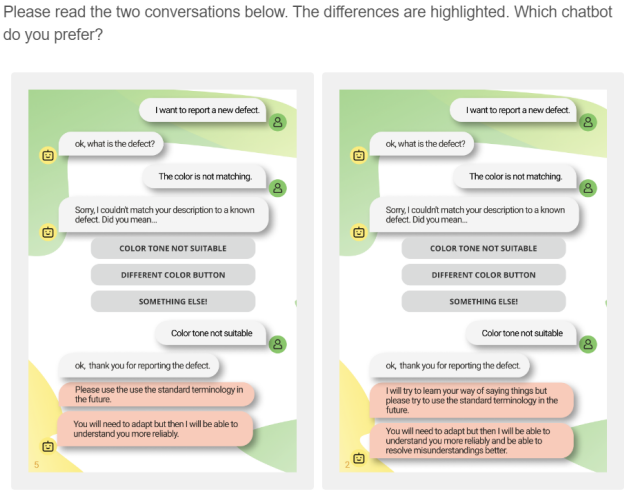


Figure 2: One of the three pairwise comparisons for the white-goods factory scenario to select the preferred learning mechanism

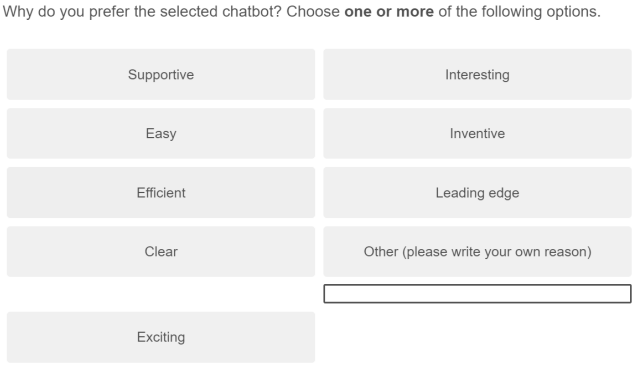


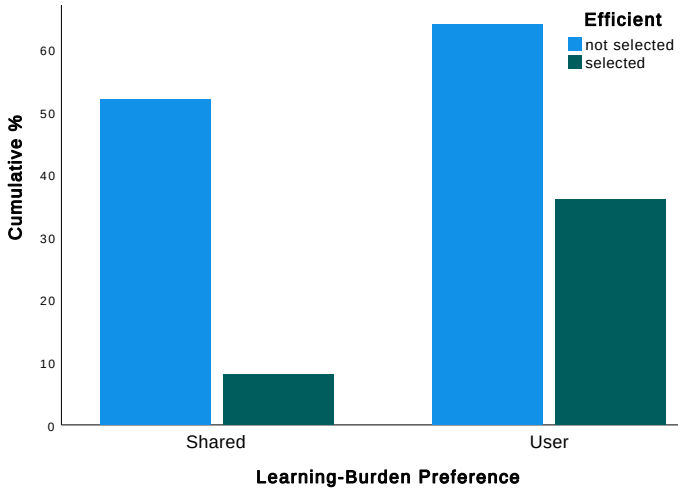
Figure 3: After each pairwise comparison, the participants are asked to provide the reason for their preference

Akaike Information Criterion (AIC) using (a) the Bayesian EM algorithm (AIC = 113.232), (b) the numerical inference method (AIC = 113.720), and (c) the sampling inference method (AIC = 111.401). Here, the recorded preference for each learning-burden comparison (CA vs. User, Shared vs. User, and Shared vs. CA) is considered a between-subjects factor for each pair, since the participants had to pick one of two options (e.g., "Shared" vs. "User"). Thus, we ran post-hoc chi-square tests of goodness of fit, while correcting for continuity, to determine if the comparison results for each learning-burden pair differed significantly. In other words, we tested if there were significant differences in the number of times participants selected one learning-burden preference over the other (e.g., number of "Shared" vs. number of "User"). Our results showcase that participants (1^{st}) prefer to share the learning burden with the CA (61.3 %), followed by (2^{nd}) entirely outsourcing the learning burden to the CA (60.7 %) as opposed to themselves, and last (3^{rd}) sharing

[2]https://help.xlstat.com/6467-bradley-terry-model-excel-tutorial last accessed on October 21, 2022.

Rank	$i > j$	P_B	$\chi^2(1)$	p
1	**Shared > User**	**.613**	21.947	$< .001^*$
2	**CA > User**	**.607**	21.947	$< .001^*$
3	**Shared > CA**	**.506**	22.132	$< .001^*$
4	CA > Shared	.494		
5	User > CA	.393		
6	User > Shared	.387		

(a) User preferences for learning-burden assignment, where $P_B(i > j)$ is the probability of assigning the learning burden to i instead of j.



(b) Participants indicated that it is more efficient if the User handles the learning burden instead of sharing it with the CA.

Figure 4: (a) Participants preferred sharing the learning burden with the Conversational Agent (CA) with a probability of 61.3 %. (b) However, participants indicated that it is more efficient if the user handles the learning burden instead of sharing it with the CA.

the learning burden with the CA instead of entirely outsourcing it to the CA (50.6 %) (see Table 4a).

Next, we inquired into the reasons why our participants selected one learning burden over the other, for the top 3 learning-burden preferences (see Table 4a). Similar to the analysis above, we ran multiple chi-square tests of goodness of fit while correcting for continuity. Again, the learning burden was treated as a between-subjects factor in each comparison, and the indicated reasons were encoded as dichotomous variables (e.g., "1" if a learning burden type was indicated as "supportive" and "0" if it was not). For the "Shared > User" comparison, a series of chi-square tests displayed no significant difference over the reasons of being more supportive ($\chi^2(1) = .007, p = .934$), easier ($\chi^2(1) = .174, p = .677$), more clear ($\chi^2(1) = .260, p = .610$), and for "other" reason ($\chi^2(1) = 1.500, p = .221$), but a significant difference for being more efficient ($\chi^2(1) = 6.084, p < .05, V = .578$). Interestingly, this significant difference emerged for the exact opposite selection of learning-burden preference (i.e., "User > Shared"). In other words, 77.8 % of the participants indicated that it is more efficient if the User handles the learning burden, as opposed to 22.2 % who indicated that Sharing the learning burden with a CA is a more efficient approach (see Fig. 4b).

For the "CA > User" comparison, a series of chi-square tests displayed no significant difference over the reasons of being more supportive ($\chi^2(1) = .356, p = .551$), more efficient ($\chi^2(1) = .048, p = .826$), easier ($\chi^2(1) = .000, p = 1.000$), more clear ($\chi^2(1) = .000, p = 1.000$), or "other" reason ($\chi^2(1) = 1.346, p = .246$). Finally, for the "Shared > CA" comparison, a series of chi-square tests displayed no significant difference over the reasons of being more supportive ($\chi^2(1) = .000, p = 1.000$), more efficient ($\chi^2(1) = 1.284, p = .257$), easier ($\chi^2(1) = .155, p = .694$), more clear ($\chi^2(1) = .057, p = .811$), or "other" reason ($\chi^2(1) = .390, p = .532$). Additional reasons, such as "exciting, "interesting", "inventive", and "leading edge" were excluded due to low occurrence.

6 DISCUSSION AND CONCLUSION

Our results indicate that users in the manufacturing domain prefer CAs that *learn users' divergent phrasings to some extent* from successfully repaired breakdowns. Based on UTAUT theory [43], users seem to perceive that the effort in learning what the CA can understand is significantly higher than the benefit of adapting to it. Interesting to note, participants who preferred to take responsibility for learning domain-specific keywords rather than sharing it with the CA stated that they found it more efficient. A potential reason could be that these participants do not trust the CA, i.e., due to negative prior experience, and would instead rely on their own domain knowledge.

A continuously learning CA will improve its NLU and knowledge base over time and reduce the number of conversational breakdowns. However, there are several disadvantages and barriers to this strategy: (1) it initially requires more effort from developers, (2) it is more costly, and (3) the new training data provided by the users may introduce conflicts (e.g., faults, bias). Hence, it is advisable to introduce a validation mechanism that can either be automated or involves a human in the loop.

The main advantages of relying on users to learn from breakdowns to minimize their recurrence are the following: (1) less development is required, (2) users can generalize what they learned to reduce breakdowns in other scenarios, and (3) users do not expect the CA to improve and are therefore less disappointed when it fails.

In contrast, it has the following disadvantages: (1) there is a limit to how many new terms the user can remember, (2) the user may resist adapting to the CA, and (3) the same breakdown may occur multiple times before the user learns to avoid it (especially if that specific breakdown occurs infrequently).

As a starting point, we believe that our findings can be helpful for other use cases where (1) CAs' training data intensely uses domain-specific terminologies, or (2) building a comprehensive training data set for the early production stages is not feasible. Furthermore, businesses having use cases with similar characteristics should consider a setup where the CA continuously learns. The potential lack of trust in the CA is a serious constraint for user acceptance. Therefore, businesses should focus on methods to build trust. These could include training sessions to teach workers about the capabilities, risks, and limitations of a CA.

We encourage future work to investigate the possible impact of personal factors on user preferences. Since the perceived benefits of the CA depend on the presented conversations in this study, future studies should investigate under which circumstances the users are willing to adapt. Such studies could focus on scenarios featuring tasks with various complexity, importance, or time-criticality. Besides, sharing the learning burden between the user and the CA can be conceptualized in further details. For instance, the CA and the user could collaborate and dynamically assign the burden of learning to each other depending on the reason behind the breakdown, task sensitivity, or user interaction history.

ACKNOWLEDGMENTS

This work is funded by the European Union's Horizon 2020 research and innovation program via the project COALA "COgnitive Assisted agile manufacturing for a LAbor force supported by trustworthy Artificial Intelligence" (Grant agreement No 957296).

REFERENCES

[1] M. Akhtar, J. Neidhardt, and H. Werthner. 2019. The Potential of Chatbots: Analysis of Chatbot Conversations. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, Vol. 1. IEEE Computer Society, Los Alamitos, CA, USA, 397–404. https://doi.org/10.1109/CBI.2019.00052

[2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[3] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 243, 13 pages. https://doi.org/10.1145/3290605.3300473

[4] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[5] Maresa Biermann, Evelyn Schweiger, and Martin Jentsch. 2019. Talking to Stupid?!? Improving Voice User Interfaces. In *Mensch und Computer 2019 - Usability Professionals*, Holger Fischer and Steffen Hess (Eds.). Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn, 53–61. https://doi.org/10.18420/muc2019-up-0253

[6] Luka Bradeško, Janez Starc, Dunja Mladenic, Marko Grobelnik, and Michael Witbrock. 2016. Curious Cat Conversational Crowd Based and Context Aware Knowledge Acquisition Chat Bot. In *2016 IEEE 8th International Conference on Intelligent Systems (IS)* (Sofia, Bulgaria). IEEE Press, New York, 239–252. https://doi.org/10.1109/IS.2016.7737428

[7] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[8] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. 2009. A Crowdsourceable QoE Evaluation Framework for Multimedia Content. In *Proceedings of the 17th ACM International Conference on Multimedia* (Beijing, China) *(MM '09)*. Association for Computing Machinery, New York, NY, USA, 491–500. https://doi.org/10.1145/1631272.1631339

[9] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*, L.B. Resnick, J.M. Levine, and S.D. Teasley (Eds.). American Psychological Association, Washington, D.C., 127–149. http://psych.stanford.edu/~herb/

[10] Eric Corbett and Astrid Weber. 2016. What Can I Say? Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 72–82. https://doi.org/10.1145/2935334.2935386

[11] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can i Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[12] H. A. David. 1963. *The method of paired comparisons*. Hafner Pub. Co, New York.

[13] Brett Edwards, Michael Zatorsky, and Richi Nayak. 2008. Clustering and Classification of Maintenance Logs Using Text Data Mining. In *Proceedings of the 7th Australasian Data Mining Conference - Volume 87* (Glenelg, Australia) *(AusDM '08)*. Australian Computer Society, Inc., AUS, 193–199.

[14] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience* 5, 1 (2020), 3. https://doi.org/10.1007/s41233-020-00033-2

[15] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for Customer Service: User Experience and Motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3342775.3342784

[16] Asbjørn Følstad and Cameron Taylor. 2020. Conversational Repair in Chatbots for Customer Service: The Effect of Expressing Uncertainty and Suggesting Alternatives. In *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Springer International Publishing, Cham, 201–214.

[17] Dominic Gorecky, Mathias Schmitt, Matthias Loskyll, and Detlef Zühlke. 2014. Human-machine-interaction in the industry 4.0 era. In *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*. IEEE Press, New York, 289–294. https://doi.org/10.1109/INDIN.2014.6945523

[18] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazaré, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! *CoRR* abs/1901.05415 (2019), 3667–3684. arXiv:1901.05415 http://arxiv.org/abs/1901.05415

[19] J. Harms, P. Kucherbaev, A. Bozzon, and G. Houben. 2019. Approaches for Dialog Management in Conversational Agents. *IEEE Internet Computing* 23, 2 (2019), 13–22. https://doi.org/10.1109/MIC.2018.2881519

[20] Nancy Larson-Powers and Rose Marie Pangborn. 1978. Paired comparison and time-intensity measurements of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. *Journal of Food Science* 43 (1978), 41–46. https://doi.org/10.1111/j.1365-2621.1978.tb09732.x

[21] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully Mitigating Breakdowns in Robotic Services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, New York, 203–210.

[22] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. *A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376209

[23] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. *Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs*. Association for Computing Machinery, New York, NY, USA, 1094–1107. https://doi.org/10.1145/3379337.3415820

[24] Bing Liu and Sahisnu Mazumder. 2021. Lifelong and Continual Learning Dialogue Systems: Learning during Conversation. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15058–15063. https://ojs.aaai.org/index.php/AAAI/article/view/17768

[25] Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2060–2069. https://doi.org/10.18653/v1/N18-1187

[26] Ewa Luger and Abigail Sellen. 2016. *"Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents.* Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[27] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices.* Springer, Cham. https://doi.org/10.1007/978-3-319-32967-3

[28] Joanne Meredith. 2017. Analysing technological affordances of online interactions using conversation analysis. *Journal of Pragmatics* 115 (2017), 42–55.

[29] Robert J. Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework.* Association for Computing Machinery, New York, NY, USA.

[30] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173580

[31] Alan Nichol. 2020. Conversation-Driven Development. https://rasa.com/blog/conversation-driven-development-a-better-approach-to-building-ai-assistants/

[32] Cecilie Bertinussen Nordheim, Asbjørn Følstad, and Cato Alexander Bjørkli. 2019. An Initial Model of Trust in Chatbots for Customer Service—Findings from a Questionnaire Study. *Interacting with Computers* 31, 3 (2019), 317–335. https://doi.org/10.1093/iwc/iwz022

[33] Cecilie Bertinussen Nordheim, Asbjørn Følstad, and Cato Alexander Bjørkli. 2019. An Initial Model of Trust in Chatbots for Customer Service — Findings from a Questionnaire Study. *Interacting with Computers* 31, 3 (2019), 317–335. https://doi.org/10.1093/iwc/iwz022

[34] Oracle. 2022. Oracle Digital Assistant for ERP and SCM. https://www.oracle.com/chatbots/digital-assistant-for-erp-scm/

[35] Xin Rong, Adam Fourney, Robin N. Brewer, Meredith Ringel Morris, and Paul N. Bennett. 2017. Managing Uncertainty in Time Expressions for Virtual Assistants. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17).* Association for Computing Machinery, New York, NY, USA, 568–579. https://doi.org/10.1145/3025453.3025674

[36] R. Rosenthal and R. L. Rosnow. 2008. Essentials of Behavioral Research: Methods and Data Analysis (3rd edition). https://doi.org/10.34944/dspace/66 The only comprehensive treatment of methods and data analysis, this classic advanced undergraduate/graduate text in research methods requires statistics as a prerequisite. The first half of the text concentrates on research methods and the second half introduces students to advanced statistical procedures..

[37] SAP. 2022. SAP Conversational AI. https://www.sap.com/germany/products/conversational-ai.html

[38] Emanuel A. Schegloff. 1992. Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *Amer. J. Sociology* 97, 5 (1992), 1295–1345. http://www.jstor.org/stable/2781417

[39] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language* 53, 2 (1977), 361–382. https://doi.org/10.2307/413107

[40] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (01 2017), 103. https://doi.org/10.9781/ijimai.2017.09.001

[41] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide. 2000. The thoughtful elephant: strategies for spoken dialog systems. *IEEE Transactions on Speech and Audio Processing* 8, 1 (2000), 51–62. https://doi.org/10.1109/89.817453

[42] Spix Industry. 2022. Spix: Industry-Ready Chatbot Technology. https://www.spix-industry.com/en/products-chatbot-for-industrial-software-development-spix/

[43] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478. http://www.jstor.org/stable/30036540

[44] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics* 7 (July 2019), 387–401. https://doi.org/10.1162/tacl_a_00279

[45] Justin D. Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19).* Association for Computing Machinery, New York, NY, USA, 448–459. https://doi.org/10.1145/3301275.3302290

[46] Stefan Wellsandt, Karl Hribernik, and Klaus-Dieter Thoben. 2021. Anatomy of a Digital Assistant. In *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*, Alexandre Dolgui, Alain Bernard, David Lemoine, Gregor von Cieminski, and David Romero (Eds.). Springer International Publishing, Cham, 321–330. https://doi.org/10.1007/978-3-030-85910-7_34

[47] J Wilke, Fergus McInnes, Mervyn A Jack, and P Littlewood. 2007. Hidden menu options in automated human–computer telephone dialogues: dissonance in the user's mental model. *Behaviour & Information Technology* 26, 6 (2007), 517–534.

[48] Jason D. Williams, Nobal B. Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. *Rapidly Scaling Dialog Systems with Interactive Learning.* Springer International Publishing, Cham, 1–13. https://doi.org/10.1007/978-3-319-19291-8_1

[49] Terry Winograd and Fernando Flores. 1987. *Understanding Computers and Cognition: A New Foundation for Design.* Addison-Wesley Longman Publishing Co., Inc., USA.

[50] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95).* ACM Press/Addison-Wesley Publishing Co., USA, 369–376. https://doi.org/10.1145/223904.223952