# ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust

Anjali Khurana
*Simon Fraser University*
*Burnaby, BC, Canada*
anjali_khurana@sfu.ca

Parsa Alamzadeh
*Simon Fraser University*
*Burnaby, BC, Canada*
palamzad@sfu.ca

Parmit K. Chilana
*Simon Fraser University*
*Burnaby, BC, Canada*
pchilana@cs.sfu.ca

***Abstract*—When breakdowns occur during a human-chatbot conversation, the lack of transparency and the "black-box" nature of task-oriented chatbots can make it difficult for end users to understand what went wrong and why. Inspired by recent HCI research on explainable AI solutions, we explored the design of in-application explainable chatbot interfaces (ChatrEx) that explain the underlying working of a chatbot during a breakdown. ChatrEx-VINC provides visual example-based step-by-step explanations in-context of the chat window whereas ChatrEx-VST provides explanations as a visual tour overlaid on the application interface. We implemented these chatbots for complex spreadsheet tasks and our comparative observational study (N=14) showed that the explanations provided by both ChatrEx-VINC and ChatrEx-VST enhanced users' understanding of the reasons for a breakdown and improved users' perceptions of usefulness, transparency, and trust. We identify several opportunities for future research to exploit explainable chatbot interfaces and better support human-chatbot interaction.**

***Index Terms*—chatbots, visual explanations, in-application help, conversational breakdowns, human-chatbot interaction**

## I. Introduction

In-application virtual assistants and task-oriented chatbots embedded inside software applications offer several opportunities to automate various tasks and support the use of complex application features. But, despite the promise of these chatbots, many users feel annoyed and even abandon these assistants after repeated unsuccessful interactions [1]. For example, *Clippy* was introduced in the Microsoft Office suite as early as 1996 [2] to assist users in performing various word processing tasks, only to be removed four years later based on negative user feedback.

Recent progress in machine learning (ML) and Natural Language Processing (NLP) has contributed to improving chatbot functionality manyfold at the underlying algorithmic level. However, complexities of natural language interactions [3, 4] and limited training sets and poor conversational understanding [5] remain to be key obstacles in fully realizing the potential of human-chatbot interaction. For example, a key challenge for users of task-oriented chatbots is dealing with conversational dead-ends or breakdowns [6, 7, 8]). In fact, during a breakdown, as many as 70% of users may opt to quit the task or completely abandon the chatbot, while others may try to rephrase their queries with little or no success [4].

A breakdown usually occurs when a chatbot fails to understand the user's intent in a query [9] and the user does not know what to do next. In fact, the chatbot often appears as a "black-box" to the user, making it difficult to understand why something did not work, what actions are actually possible, and how to recover from the breakdown. This lack of transparency, in turn, impacts the users' perceptions of usefulness and trust in the system [10, 11, 12].

In this work, we explore the design of in-application task-oriented chatbots that can explain the underlying steps of a task and where and why they failed during a conversational breakdown. We take inspiration from recent research in Explainable AI (XAI) which recognizes the need to incorporate explainability features or explanations for improving transparency and trust [11, 13, 14]. The goal of our approach was not only to acknowledge the occurrence of a breakdown (as has been explored in recent work [6]), but also to design novel mechanisms that can enhance user understanding of what caused the breakdown and where exactly the breakdown occurred. Our overarching goal was: how can we design an in-application chatbot that can explain the underlying steps of a task and indicate where and why it failed?

We propose a novel class of explainable chatbot interfaces (*ChatrEx*) that visually explain a chatbot's high-level operations and causes of a breakdown. We explore two variations of ChatrEx that either provide visual explanations in context of the chatbot (*ChatrEx-VINC*, Figure 1), or as a visual tour overlaid on the application interface (*ChatrEx-VST*, Figure 4). We implemented the design of these chatbot interfaces as an add-on for *Google Sheets*, an online spreadsheet application.

To evaluate these two explainable chatbot designs, we compared them to an existing explanation design based on keyword highlighting [6] and a baseline chatbot that provided no explanations. We conducted an observational usability study with 14 participants and assessed their perceptions of usefulness, transparency, and trust across these four chatbots. Overall, we found that participants consistently ranked ChatrEx-VST and ChatrEx-VINC higher across all of our key measures. Users indicated that the visual example-based explanations made the chatbot's functionality and decisions more transparent, and in turn, improved users' perceived trust in the chatbot.

The main contributions of this paper are: (1) the design and implementation of two novel in-application chatbot interfaces that provide visual example-based explanations to illustrate the underlying working of a chatbot and help users recognize the causes of a breakdown; (2) empirical insights into the strengths and weaknesses of explainable chatbot interfaces based on users' perceptions of usefulness, transparency and trust.

978-1-6654-4592-4/21/$31.00 ©2021 IEEE

## II. Related Work

To contextualize our research, we draw upon literature on the design and evaluation of task-oriented and in-application chatbots, and explainable AI systems.

### A. User perceptions of task-oriented chatbots

Previous studies of task-oriented chatbots, such as Siri and Cortana, have contributed insights into trust issues and the struggles that users face with a lack of appropriate feedback. For example, Luger and Sellen [9] highlighted a gap between user expectation and system operation because users found it difficult to understand the capability of the chatbot and how the chatbot could actually accomplish a task. Another study [12] raised concerns with the lack of effective system status and how chatbots, such as Alexa, were a "black box" for users when they faced an error or a breakdown. As such, users were more likely to lose trust and less likely to continue using these chatbots after experiencing a breakdown, especially when engaged in complex tasks [9].

In terms of conveying a task-oriented chatbot's understanding during a breakdown, only a few examples exist. Recently Li et. al [7] explored multi-modal strategies in the context of existing mobile apps for fixing NLU breakdowns and command disambiguations [8]. Although these solutions focused more on supporting interactive repair strategies during breakdowns, they demonstrated an effective use of app GUIs to help ground the conversation. Our work goes further to address the gap of improving users' perception of transparency and trust for chatbots that are embedded in feature-rich applications, such as spreadsheets. Our novel ChatrEx designs visually explain the underlying working of a chatbot using the UI components as referents, allowing users to learn about the chatbot's competencies and limitations even if they are not familiar with the application functionality.

### B. Design and evaluation of in-application chatbots

Early versions of in-application task-oriented chatbots that were designed to help end users be more efficient with software tasks, unfortunately, saw high rates of user abandonment [15]. Perhaps the most well-known failure has been that of the Office Assistant named "Clippy" that received widespread negative user feedback and was later removed by Microsoft [15, 16, 17]. Since then, there have been many research efforts to advance the work in creating more helpful and efficient automated chatbots in applications. For example, *Calendar.help*, was introduced as a personal assistant to provide fast and efficient scheduling via email, but, ultimately, it was unable to handle a lot of the complex calendaring tasks on its own [18]. The opacity of these systems is known to be a key challenge as users struggle to understand what inputs and outputs are actually possible. More recently, Glass et al. [11] assessed the factors impacting the trust and understandability of CALO, a personalized assistant for office-related tasks, and similarly found that users perceived the system to be too "opaque" and difficult to comprehend. In fact, the lack of transparency was mentioned to be one of most crucial reasons responsible for affecting trust among users. While some works suggest using explanation-based systems [11] to augment chatbots and make them easier to understand, it is yet to be explored how to structure and design such explanations. Our paper complements these existing works by designing novel explainable interfaces for in-application task-oriented chatbots that can improve transparency and trust among users.

### C. Explainable AI (XAI) systems

Recently, there has been a big push in AI and HCI research to design XAI solutions to make complex ML algorithms more understandable for end users [19, 20, 21, 22, 23, 24]. Notably, many of these studies have focused on explaining the underlying algorithms through different explanation methods such as *global explanations* to explain the model, *local explanations* to explain a prediction or *inspect counterfactual* to explain the features influencing the prediction [13]. However, end users who do not have any knowledge or experience with ML struggle to understand these in-depth algorithm-specific explanations [25, 26]. Still, it has been shown that such explanations can play a key role in enhancing transparency and trust for AI systems [21, 22, 23]. Our paper complements these existing works by exploring the potential of XAI design solutions for improving user interaction with in-application task-oriented chatbots.

The closest work to ours is perhaps the recent work on keyword highlighting [6] that tries to explain the underlying intent of the user's input in a query and highlights parts that the chatbot did and did not understand. Although this level of highlighting was useful as a repair strategy, we argue that for more complex tasks and applications, these highlighting-based explanations are not sufficient enough to explain the underlying working of the chatbot. To provide transparency and more in-depth reasons of the breakdown, it is equally important to expose the chatbot's inner workings and provide users a window into the chatbot's competencies and limitations [27]. ChatrEx takes inspiration from these existing works to expand and explore the design space of explainable chatbots while contributing novel visual explanation designs for chatbots embedded in feature-rich applications.

## III. Motivation and Design Goals

In this paper, we explore the design of in-application explainable chatbot interfaces (ChatrEx) that can explain a chatbot's underlying functionality during a breakdown. Our main goal is to improve users' perceptions of transparency, trust and usefulness when working with in-application chatbots. Based on the related work and current state-of-the-art in task-oriented chatbots, we considered how to structure the explanation and how enhance the explanation with examples and visuals, culminating in five key design goals.

### A. Structuring the explanation with intent and entity

To perform the action requested in a user's query, a typical task-oriented chatbot first identifies the *intent* and the *entity*. The intent refers to the final objective of the user's query, while

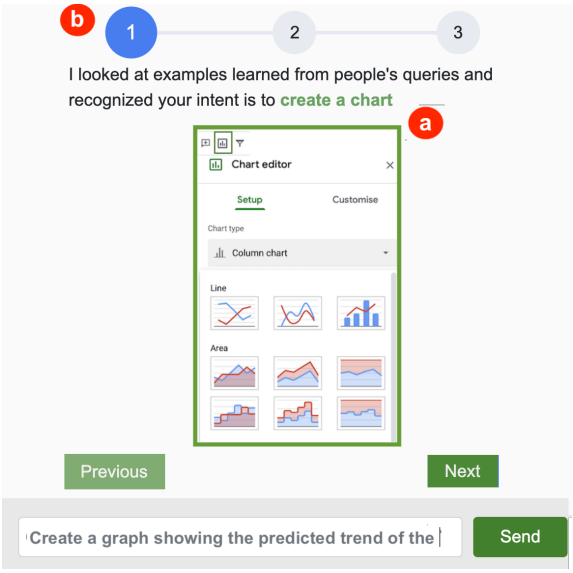


Fig. 1. An example of **ChatrEx-VINC** displaying normative visual training examples (a), highlighted in green, to convey chatbot's competencies .

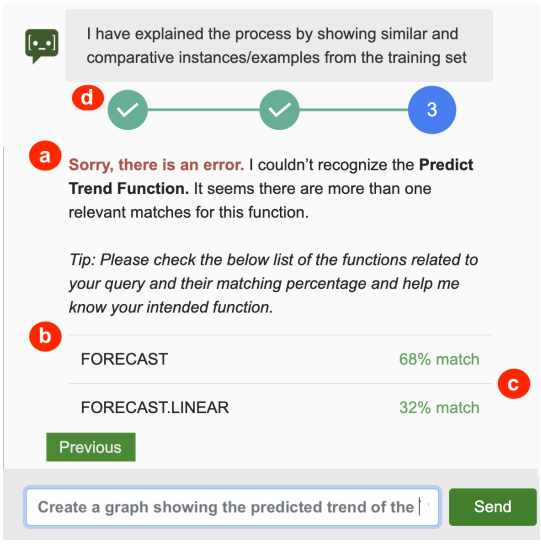


Fig. 2. An example of **ChatrEx-VINC** explaining a breakdown (disambiguation task). Comparative visual examples are shown for most similar visual training examples (a), the potential matches (b), and match percentages (c).

the entity includes the remaining information from the query to add parameters and to make the objective more specific [6, 28, 29]. For example, consider this chatbot query in a spreadsheet application: "Create a graph that shows the square root of column C data." While the intent would be *creating a graph*, the entity would be the functions or operations such as square root and data (i.e., Column C). The critical conversational breakdown occurs when the chatbot fails to correctly comprehend the intended meaning of the user's query. In explaining the internal working of a chatbot, it is imperative to structure the explanation such that it provides clear and concise information about the intent and the entity.

A common challenge for XAI solutions is to reconcile the significance of explaining decisions versus competencies of the AI system [30]. Typically, XAI systems are expected to explain the decision process (i.e., reasons for the system's action), especially when something goes wrong [31]. While it is helpful for users when a system acknowledges the decision (i.e., breakdown) [6], it is equally significant to help users comprehend the competencies or capabilities of AI systems [30]. We hypothesize that explaining the competencies and limitations of the chatbot using the identified intent and entity will not only aid users to recognize the breakdown but also improve transparency. Furthermore, within the breakdown decision, an indication of where the problem occurred and its possible causes would help the users more clearly understand the cause of the breakdown and repair their queries [7].

### *B. Enhancing the explanation with examples and visuals*

Examples have been shown to be effective [32, 33] for explaining AI predictions without overwhelming users with internal algorithmic logic [34, 35, 36]. Particularly, users find high level and simple explanations to be more useful and easier to interpret [37, 38, 39, 40]. In fact, the XAI Taxonomy recommends the method of "example-based explanations" [13] that provide normative or comparative examples of the instance [19, 41, 42]. Normative explanations display the most similar training examples from the target classes for enhancing system understanding. In contrast, comparative explanations highlight similarities or differences between a user's input and the alternative classes as limitations, which can be useful for representing breakdowns related to disambiguation and infeasibility [19]. When normative and comparative explanations are used to demonstrate the capability and limitations of complex systems, they are found to be more effective in improving users' trust [19, 20].

Another consideration for designing explanations is whether to present them verbally [26] or visually [43]. Recent studies suggest that verbal prompts tend to become "visually unappealing" and "difficult to read" [6] whereas visual explanations increase transparency and users' trust in automated systems [19, 20, 44]. Moreover, the internal working of the chatbot and how it processes the user's query should be shown step-by-step [29]. Finally, chatbots that only appear when called upon by users can be less intrusive [15] and may be perceived to be more useful [14].

### *C. Design Goals for Explainable Chatbot Interfaces*

Based on the above considerations, we derived five key design goals for building in-application chatbots that can explain their underlying functionality during a breakdown:

1) **DG1:** Explain the **chatbot's functionality in terms of intent and entity** so that users can better understand the high level underlying working of the chatbot.
2) **DG2:** Illustrate **competencies of the chatbot and reasons why a breakdown occurred** so that users can understand what the chatbot could and could not comprehend from their query. The explanation should indicate the exact reason of the chatbot's failure by elucidating

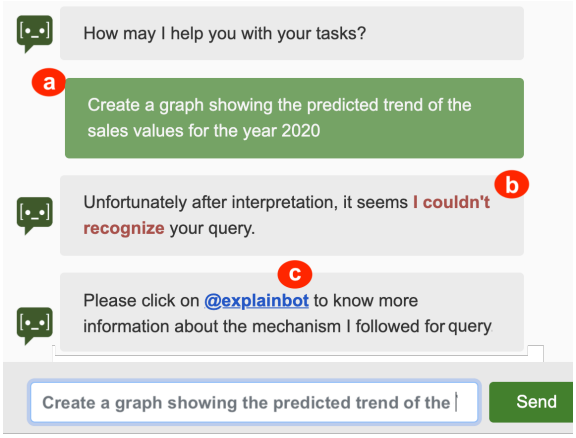


Fig. 3. The common entry point for ChatrEx: (a) users submit a query, (b) error message is shown, (c) @explainbot feature can be invoked in response.

"where" and "what caused" the breakdown with respect to the identified intent and entity.

3) **DG3:** Provide **normative and comparative example-based explanations** for explaining both the competencies and limitations, respectively.
4) **DG4:** Provide **visual step-by-step explanations** to make them appealing, relatable, and easy to comprehend.
5) **DG5:** Allow users to have **freedom and control in accessing and navigating the explanations** by including UI controls such as, "next", "previous", or "exit".

## IV. ChatrEx: System Design

Based on the above goals and iterative design approaches [45], we designed and implemented novel web-based chatbot interfaces that simulate breakdowns and their corresponding explanations. We first created several low-fidelity paper prototypes, followed by mock-ups using *PowerPoint*, and medium-fidelity prototypes using the *Axure* prototyping software, soliciting informal user feedback at each stage to help us iterate on our ideas. Our final two designs for ChatrEx were *ChatrEx-VST*(Figure 4) and *ChatrEx-VINC* (Figure 1, 2) that provide two different types of explanations about a chatbot's underlying functionality during a breakdown and why it failed, including reasons related to disambiguation and infeasibility. We selected Google Sheets as the underlying application as it has several complex spreadsheet features, allowing us to devise a range of tasks for chatbot assistance. [46].

Our ChatrEx-VINC and ChatrEx-VST chatbots represent two different kinds of visual explanations, as described below. In both cases, users can issue text-based queries to initiate a conversation about automating spreadsheet tasks (Figure 3.a). If a user sees an error message (Figure 3.b) after issuing the query, they can invoke the @explainbot feature (Figure 3.c) to see an explanation about what the chatbot understood and why the breakdown occurred.

### A. ChatrEx-VINC: Visual in-context explanations

ChatrEx-VINC provides in-context visual example-based step-by-step explanations (DG4). Similar to the idea of example-based explanations based on the training set for a classifier [13, 19], ChatrEx-VINC shows examples from the training set of each keyword in the query (i.e., intent and entity) recognized by the chatbot. Fulfilling DG1, DG2 and DG3, ChatrEx-VINC distinctly explains (Figure 1) the intent/entity that the chatbot comprehended successfully through training examples from the target class (normative explanations). Similarly, ChatrEx-VINC further explains the breakdown decision through the most similar or different examples from the alternative training classes (i.e., comparative explanations). In particular, when the breakdown occurs due to disambiguation, the explanation provides similar examples which matched the user's intent or entity and were possibly misrecognized (Figure 2). In contrast, when the breakdown occurs due to a task being infeasible for the chatbot, the explanations provide feasible alternative examples which users can follow instead of the original intent or entity that the chatbot is not trained for.

To provide a better understanding of the chatbot during the breakdown, each example (Figure 2.b) is accompanied with the corresponding match percentages (Figure 2.c). This is analogous to confidence scores within an intent-based model [47] that represents the similarities between the user's intent and examples in the training set. As shown in Figures 1.a and 2.a, the explanations also highlight the competencies of the chatbot in green and breakdowns in red along with a dialog message. To show the real-time system status more interactively (as suggested in [9]), we adopted a design similar to the "Status Tracker" UI [48] and show the step-by-step explanations in the form of latest status and updates in a chronological order (Figure 1.b). When each of these steps are visited by the user, they are updated with GREEN check marks (Figure 2.d) allowing users to follow the explanation steps intuitively. Addressing DG5, users can use the *next* and *previous* buttons to control the navigation of these explanations.

### B. ChatrEx-VST: Visual step-through explanations

In contrast to ChatrEx-VINC, ChatrEx-VST presents a step-by-step visual tour with examples overlaid (DG4) directly on the application user interface (Figure 4). We draw inspiration from in-application software walkthroughs or onboarding tours that explain features and functionality in a way that is relatable and engaging for users [49].

When a user invokes the @explainbot feature (Figure 3.c), they can seek more information about their query and see what the chatbot understood. ChatrEx-VST first minimizes the chat window and overlays a transparent background atop the UI. Next, it highlights (Figure 4.a, 4.d) the visual examples in the UI (e.g., menu items, data items, functions, etc.) corresponding to the intent or entity recognized from the user's query (DG1) along with a descriptive message. Similar to ChatrEx-VINC, ChatrEx-VST fulfills DG3 and DG4 by distinctly explaining the chatbot's competencies through normative visual explanations highlighted in green boxes (Figure 4.a, 4.b) and breakdown decisions through comparative explanations highlighted in red boxes along with match percentages (Figure 4.d, 4.e). Addressing DG5, the user can easily navigate to the next or

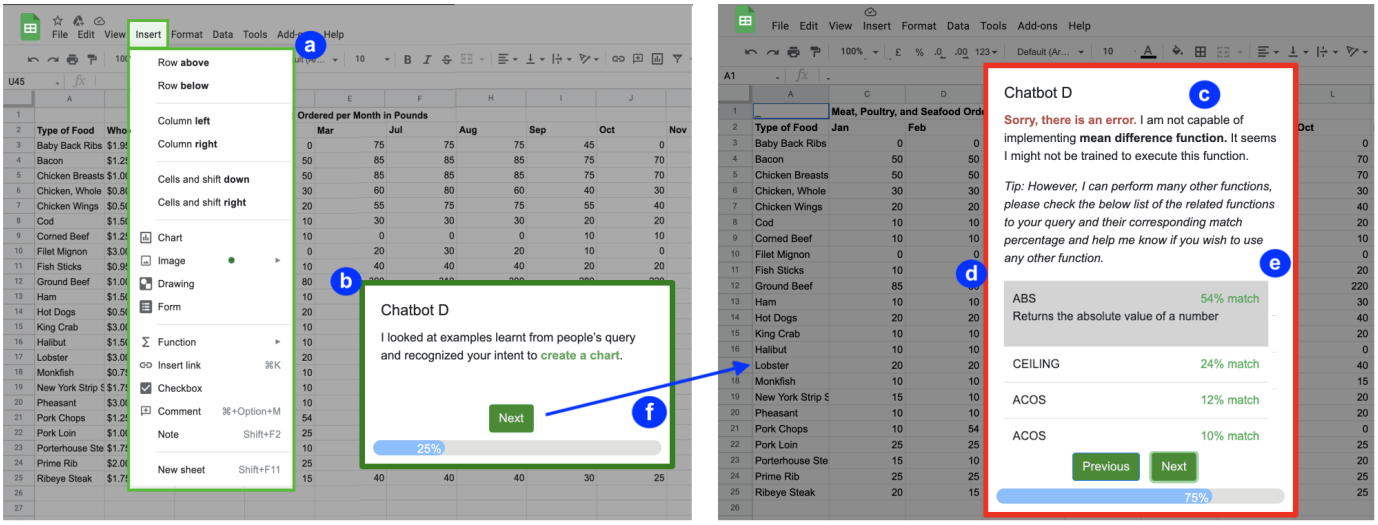


Fig. 4. ***(Left)* ChatrEx-VST Competencies:** By clicking @explainbot, ChatrEx-VST presents a visual tour overlaid on the application UI, highlighting normative visual training examples on the interface in green (a,b). ***(Right)* ChatrEx-VST Breakdown decision:** provides comparative visual example-based explanations (c) through alternative visual training examples (for an infeasible query) along with match percentages (d,e) *[Note: both ChatrEx designs support disambiguation and infeasible queries, here we showed disambiguation for ChatrEx-VINC and infeasible query for ChatrEx-VST]*

previous step on their own (Figure 4.f) or use the *Finish* button to end the overlaid tour and return to the chat window.

### C. Implementation details

Since our main contribution is in exploring the design of visual explainable chatbots, our implementation focused on developing interactive proof-of-concept prototypes rather than innovating on the underlying NLP or ML algorithms. We created web-based prototypes to demonstrate key chatbot functionality so that users could evaluate the different explanation designs for various statistical and visualization-related spreadsheet tasks. We took inspiration from chatbots that rely on intent-based models [47] where multi-classifiers can predict the intent in the user's query and calculate confidence scores with respect to all predefined intents. A breakdown occurs if all of these confidence scores are below a certain threshold. The highlighting of the breakdown in red is inspired by the typical ML approach used to identify keywords in the query having the highest weight on predicted intent. Aspects of the visual examples are inspired from recent work [19] where the training set included visual examples for predefined intent.

ChatrEx consists of two main modules: the UI module and the Natural Language Understanding (NLU) module. The UI module lays out the various user interface components and receives the user's query. Next, this input query is sent to the NLU module, which uses regex and keyword extraction to detect a user's intentions and runs the query through another model to extract semantic information about the task. The intent and semantics extracted from the user queries are then mapped against our pre-existing database to retrieve the corresponding series of screenshots and context. The retrieved data are then used to fill our predefined templates for the different chatbot types (ChatrEx-VST and ChatrEx-VINC, and two other implementations used for comparison in the user study). To generate the explanation responses, these templates are then rendered using our UI module within each chatbot. The UI module is built using ReactJS and migrated to Chrome as an extension by adapting a boilerplate template [50].

## V. User Study

To evaluate the extent to which the explanations provided by ChatrEx-VINC and ChatrEx-VST help users understand a breakdown, we ran a usability study with 14 participants. To compare these designs, we implemented two other chatbot prototypes: (1) KEYHT, which was adapted from recent work on verbal keyword highlighting and confirmation explanations [6] (Figure 5) where keywords that are understood are highlighted in green and the misunderstood keywords are shown in orange; (2) BASELINE, which was our implementation of commonly used in-application chatbots that do not provide any explanations, but often recommend related search results [9]. The goal of this study was to assess the strengths and weaknesses of the different explanation designs and how users perceive them in terms of usefulness, transparency, and trust.

### A. Participants

We recruited participants mainly from our university's mailing lists and found additional participants through snowball sampling. We ended up with a diverse pool of 14 participants (7F/7M) who came from different backgrounds (CS, Sciences, Arts) and professions (client services, lab technicians, medical photographers, information designers, students and researchers). Our participants were all between the ages of 18–34 and had different levels of education (1 Bachelor's, 1 Diploma, 8 Master's, 4 PhD). The participants were familiar with a range of chatbots, including Google Assistant (12/14), Alexa (11/14), Siri (13/14). But, most participants (9/14) did not use these chatbots frequently (at most 3 times/week).

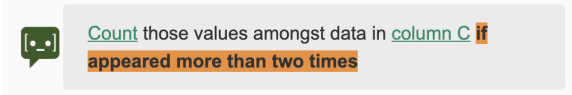


Fig. 5. **KEYHT:** Keyword Highlighting Verbal explanations.

### B. Study Design and Tasks

We used a within-subject design to minimize the impact of known high variation among participants. Each participant interacted with four web-based chatbot prototypes that represented one of the explanation designs (ChatrEx-VST, ChatrEx-VINC, KEYHT, BASELINE) in a random order. For each chatbot, we asked users to try two distinct spreadsheet tasks each (8 in total) that represented two breakdown situations:

1) *Infeasible tasks*: these spreadsheet tasks resulted in a breakdown because the chatbots were not trained to recognize and perform them. For example, for the task "Create a graph showing Euclidean distance between Column C and D", our chatbots were not trained to recognize and execute the Euclidean distance function, making the task infeasible.

2) *Disambiguation tasks*: these spreadsheet tasks were feasible but misunderstood by the chatbot. They resulted in a breakdown because there could be multiple relevant matches for the identified intent or entity. For example, for the task "Create a graph showing the predicted trend of sales values for the year 2020", the chatbot would not be able to recognize the intent for "predicted trend" because there were multiple matches (e.g., FORECAST, FORECAST.LINEAR, etc) and it would need more specific information to process the query.

We explored a range of complex statistical functions as we considered different aspects of feasibility and disambiguation. We explained to the users that the goal of our study was not to complete the actual tasks in Google Sheets, but to assess the explanations that they saw during breakdowns in their interaction with different chatbot designs. We conducted pilot tests and iterated on the phrasing of the queries several times to strike a balance between appropriate challenge, allotted time, and comprehensibility.

### C. Procedure

We conducted the study remotely through Zoom and participants were each given a $15 Amazon gift card in appreciation of their time. Participants were asked to install our prototypes via a Chrome extension that would make our chatbot designs functional on Google Sheets (an example spreadsheet was provided). Next, participants filled out a pre-test questionnaire that captured demographics and information about prior experiences with virtual assistants and spreadsheet applications.

We presented each of the 4 chatbots and spreadsheet tasks one-by-one in a random order. For each task, we asked participants to phrase an appropriate query and use the @explainbot feature to seek an explanation and we encouraged them to think aloud. When necessary, we also provided hints for constructing an appropriate query as the purpose of our study was not to test the user's understanding of spreadsheet features. After interacting with each of the 4 chatbots, users filled out post-task questionnaires (via *SurveyMonkey*) to assess their overall experience and ability to improve their query along with their perceptions of usefulness, transparency, and trust. To assess how well the explanations aid transparency, participants were asked to explain their understanding of each chatbot's underlying working and reason for a breakdown in their own words.

Lastly, we carried out follow-up interviews to further probe into the strengths and weaknesses of each chatbot's explanation design. We asked users to rank the four prototypes they interacted with in terms of explainability and trust. Sessions were video and audio-recorded for transcription, and the participants were asked to share their screen through Zoom (only during the usability test). The usability test and follow-up interview took approximately one hour.

### D. Data Analysis

We used a combination of statistical tests and an inductive analysis approach [51] to explore our study data about users' perceptions of usefulness, transparency, and trust. We ran Pearson's Chi-square test for independence with nominal variable "explanation type" (having four levels: ChatrEx-VST, ChatrEx-VINC, KEYHT, BASELINE) and ordinal variable (having three collapsed levels: Agree, Neutral and Disagree) to quantitatively determine the significance of the results. We also qualitatively observed and analyzed the participant's approach for breakdown recovery. We then created affinity diagrams using the gathered data from the task observations and interviews. Through discussions with our research team, we categorized our findings and identified key recurring themes.

## VI. Results

Overall, all of our participants ranked either ChatrEx-VINC (8/14) or the ChatrEx-VST (6/14) as the most explainable chatbot. We next present users' perceptions of usefulness, transparency, and trust as they interacted with the different chatbots in our study.

### A. Usefulness

Users found the visual explanations by ChatrEx-VST (12/14) and ChatrEx-VINC (11/14) to be more useful than KEYHT (8/14) and BASELINE (0/14) and these differences in perceived usefulness were significant ($\chi^2(6, N=56) = 51.51$, $p<0.0001$). Participants' comments indicated that ChatrEx's in-context visual representations were more "intuitive" and "more clear than words." ChatrEx-VST's step-by-step tour highlighting the visual representations directly on the application UI was particularly useful for locating specific functions corresponding to the query: *"Highlighting the menubars and data columns in the worksheet itself makes the chatbot [ChatrEx-VST] looks more organic because that is also what a human would do, so I can relate to how it thinks (P10)."*

For ChatrEx-VINC, participants found it useful to have the instructions condensed within the chat window and felt that they had more freedom to go back-and-forth between the application UI and the chatbot UI. In contrast to ChatrEx-VST where participants said the overlay and visual tour "took over"

the screen, ChatrEx-VINC offered more recognition than recall as the instructions could be used as a reference within the same screen: *"I liked [that] it [ChatrEx-VINC] was kept within the chat window...I could scroll back to the top to see what exactly I have said in case I needed to recall any information. The bot [ChatrEx-VINC] didn't expect me to remember it all. All the information just stayed there for me (P09)."*

As expected, participants did not find it useful to see the web links offered by BASELINE in response to a breakdown: *"It [Baseline] was lot more frustrating because it didn't tell me anything (P04)."* Participants mentioned that KEYHT was somewhat useful in that the chatbot acknowledged *where* it went wrong, but it was not exactly clear *why* the breakdown occurred. KEYHT's compressed and verbose explanations lacking the suggestions were perceived as "vague", less indicative of how to resolve the problem: *"KEYHT gives rough areas of the problem...didn't give me any suggestion showing [a] gap between me and the software's [process] (P12)."*

### B. Transparency

To assess users' perceptions of transparency, we considered how well the users were able to: (i) understand how the chatbot works, (ii) follow the reasons explained by the chatbot during a breakdown and, (iii) understand how to take the next step to recover from a breakdown.

In terms of understanding how the chatbot works, all participants ranked ChatrEx-VST (14/14) as their first choice, followed by ChatrEx-VINC(11/14), KEYHT (8/14), BASELINE (4/14). These differences between explanation type and users' perceptions of how the chatbot works were significant ($\chi^2$(6, N=56) = 17.72, $p$<0.01). All participants found ChatrEx-VST to be intuitive as the visual step-by-step tour showed them exactly how the chatbot processed their query.

In terms of recognizing the reasons for a breakdown, users ranked ChatrEx-VINC (13/14) and ChatrEx-VST (12/14) as being more helpful than KEYHT (9/14) and BASELINE(4/14). These differences were significant ($\chi^2$(6, N=56) = 20.76, $p$<0.01). Participants commented that the red highlights and corresponding comparative explanations in ChatrEx helped them to know where and why the failure occurred for both disambiguation and infeasible tasks: *"It [ChatrEx-VST] failed the first time [disambiguation task] because there were multiple past functions that were used for the same query. The second time [infeasible task], it failed because it wasn't capable of performing the mean difference (P02)."*

Users ranked ChatrEx-VST (11/14) and ChatrEx-VINC (9/14) higher than KEYHT (4/14) and BASELINE (4/14) in helping them to take the appropriate next step for breakdown recovery. These differences were significant ($\chi^2$(6, N=56) = 13.08, $p$<0.05). Participants mentioned that the alternative or similar function list and match percentages (Figure 4.e) served as helpful cues to see the relevant functions and improve their query: *"Because it's 45% match and return values of normal distribution function...[it's] something I want to accomplish, so I will probably use NORM.DIST command as the function name (P11)."* ChatrEx's explanations also helped users to understand infeasible tasks that the chatbot was not programmed to perform and that they could explore alternatives: *"the available functions list gave me a hint on what is/isn't available on Google Sheets and I realize that I asked it to execute or run a nonexistent function (P10)."*

Although many participants (9/14) could see some of the query-specific breakdowns with KEYHT's highlights, they struggled to understand "why" the problems occurred: *"It [KEYHT] gives me a rough idea, but [it's] not clear enough...I [had to] guess on why it failed to understand (P09)."* Since KEYHT and BASELINE overall did not provide any guidance on how to resolve the breakdown, participants said that they would rely on "trial and error" to revise their queries.

### C. Trust

Users ranked ChatrEx-VINC(7/14) and ChatrEx-VST(6/14) as more trustworthy than KEYHT(1/14) and BASELINE (0/14). This difference between explanation type and users' perceptions of trust was significant ($\chi^2$(6, N=56) = 29.43, $p$<0.0001). A recurring sentiment among participants was that the visual feedback and explanation from ChatrEx designs gave them more confidence about how the chatbot works and they could trust it more for their task: *"I trust the mechanism of ChatrEx-VINC and ChatrEx-VST...I would probably rely on that a bit better just because it at least explains and provides the suggestions I could use (P06)."*

Users also appreciated seeing visual confirmations directly within the application UI and not having to struggle to find an appropriate mapping on their own: *"I would trust CharEx-VST because...it was really highlighting the column right where it is on the worksheet...as a user I kind of recognized the location (P10)."* This allowed participants to place their trust in the chatbot as it least it was trying to understand them and be helpful. In contrast, since most participants failed to figure out the breakdown reason with KEYHT and BASELINE, they were hesitant to trust these chatbots.

Overall, we found that since participants could trust ChatrEx, they were more enthusiastic about using these chatbots for their future spreadsheet tasks. Even beyond spreadsheets, many participants expressed an interest in seeking explanations using ChatrEx in other complex applications and indicated that they would even enjoy the experience: *"It's kind [of] like pair programming with the bot. It's nice to have something to bounce ideas back and gather information from within the [ChatrEx] bot instead [of] Google search (P09)."*

## VII. Discussion

We have contributed the design and evaluation of two novel explainable chatbot interfaces (ChatrEx-VINC and ChatrEx-VST) that visually explain a chatbot's underlying functionality and decisions during a breakdown. Our findings indicate that users found these explainable chatbots to be more useful, transparent, and trustworthy compared to chatbots based on verbal keyword highlighting [6] and chatbots that provide no explanations. We now reflect on our key insights and highlight opportunities in HCI for designing explainable chatbots.

### A. Leveraging Explainable AI for breakdown recovery

Our research provides initial evidence that it can be useful for users to see *where* a breakdown occurred and *what caused* the breakdown when they are working with in-application chatbots. In particular, we demonstrated that by leveraging XAI approaches and offering visual explanations within the UI, users were able to better understand the chatbot's capabilities. Even for tasks that were infeasible for the chatbot to perform, users still found it helpful to learn about the chatbot's limitations instead of wasting time and effort in using trial-and-error strategies. Interestingly, one participant expressed the desire to have an explanation option not only during a breakdown but also during successful interactions as it could be reassuring to see that a task was completed properly. Given the promise and importance of XAI solutions explored in other contexts [19, 20], future chatbots should incorporate explanation strategies similar to the ones we have introduced in ChatrEx to allow people to learn how chatbots work. Instead of focusing on algorithmic-level explanations of the chatbot's functionality, it may be more important to focus on explaining the application UI-level functionality so that even users who are not trained in AI or ML can still find the chatbot to be transparent and trustworthy.

### B. Designing a hybrid of visual tour and non-tour mode

Both ChatrEx-VST and ChatrEx-VINC exhibit some unique strengths through their explanation designs. Although in this study we did not consider users' familiarity with spreadsheet GUIs as a factor, we informally observed that the more experienced spreadsheet users found ChatrEx-VINC's condensed within-chat explanations to be more useful. ChatrEx-VINC provided users with more control and freedom to access the information when required and allowed them to try to improve their query without leaving the screen. On the other hand, the tour mode of ChatrEx-VST that highlighted each step directly on the application UI was more intuitive for the less experienced users and helped them become aware of unfamiliar functions. One participant described the step-by-step feature of ChatrEx-VST as if somebody was "holding their hand" in helping them work through a breakdown.

Feature-rich applications such as Google Sheets support many complicated tasks, so it is likely that even experienced users may be unfamiliar with several features and functions and could benefit from ChatrEx-VST's explanations. Future chatbots should leverage the strengths of both ChatrEx-VST and ChatrEx-VINC and allow users to toggle between the 'tour mode' and the 'non-tour mode.' Future research could also build upon recent work [7] that maps user intents to specific portions of GUIs and interaction examples from other users. There are many opportunities at the intersection of ML and HCI to further investigate what level of automation and guidance may be appropriate for explainable chatbots.

### C. Empirically understanding human-chatbot interaction

With the rapid innovations in the field of AI, there is more need for HCI-oriented research that actually tries to understand human behavior and user perceptions of AI solutions [52, 53]. Our study provides various insights into how to make in-application chatbots more transparent (and even trustworthy) by leveraging visual explainable designs. Our results provide initial evidence that such explanations can enhance users' mental models of these chatbots, particularly during situations of breakdowns. There are many ways to build on these findings in future research to investigate other automated ways for increasing transparency and making these "black box" AI systems more comprehensible.

The results from our study also showed that most of our participants who had little to no experience in ML were still able to understand the visual step-by-step explanations and found them to be useful. While much of the early focus of explainable AI solutions has been on explaining algorithms [19, 20], we decided to focus on designing more high-level visual example-based explanations. We believe that such explanations can be a starting point for further understanding and improving human-chatbot interaction, particularly for complex, application-specific chatbots. Lastly, we observed some interesting individual differences among novice and expert spreadsheet users regarding their preferences for explanation designs. There is an opportunity for future empirical work to investigate these differences further. Future work can also explore how explanations would be perceived by a larger number of users in a field study in the context of even more diverse tasks and breakdowns, especially if these tasks are chained together.

## VIII. Limitations

Although our proof-of-concept interactive prototypes were useful for assessing users' initial perceptions and reactions when using explainable chatbots, more work is needed to fully understand how users would interact with such explanations when using dialog-based chatbots (e.g., [54]) and other sophisticated NLP/ML-based chatbot implementations. Although our design and implementation was limited as it was based on a single spreadsheet application, the general design of our visual explanations can be used to map user intents to specific portions of any similar graphical user interface. Our designs can be adapted to any feature-rich application that has similar UIs and menu structures and allows for clear and distinct one-to-one mappings between intents/entities and GUI interfaces and components.

## IX. Conclusions

In this paper, we have introduced two novel designs of ChatrEx that provide visual example-based step-by-step explanations to illustrate the underlying working of a chatbot during a conversational breakdown. Users found such explanations of a chatbot's competencies and reasons for breakdown to be useful, transparent, and trustworthy. Our empirical findings have several implications for leveraging and adapting Explainable AI solutions to design in-application explainable chatbots and improve overall human-chatbot interaction.

## Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this research.

## References

[1] J. Xiao, J. Stasko, and R. Catrambone, "An empirical study of the effect of agent competence on user performance and perception," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, ser. AAMAS '04. USA: IEEE Computer Society, 2004, p. 178–185.

[2] R. Meyer, "Even early focus groups hated clippy," Jun. 2015. [Online]. Available: https://www.theatlantic.com/technology/archive/2015/06/clippy-the-microsoft-office-assistant-is-the-patriarchys-fault/396653/ [Accessed: 10-May-2020].

[3] D. Cem, "8 Epic Chatbot / Conversational Bot Failures [2020 update]," Aug. 2017. [Online]. Available: https://research.aimultiple.com/chatbot-fail/ [Accessed: 24-May-2020].

[4] F. Radlinski and N. Craswell, "A theoretical framework for conversational search," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, ser. CHIIR '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 117–126.

[5] "Why Chatbots Fail: Limitations of Chatbots," Dec. 2019. [Online]. Available: https://medium.com/voice-tech-podcast/why-chatbots-fail-limitations-of-chatbots-7f291c4df83f [Accessed: 15-May-2020].

[6] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz, "Resilient chatbots: Repair strategy preferences for conversational breakdowns," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12.

[7] T. J.-J. Li, J. Chen, H. Xia, T. M. Mitchell, and B. A. Myers, "Multi-modal repairs of conversational breakdowns in task-oriented dialogs," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1094–1107.

[8] T. Li, I. Labutov, X. Li, X. Zhang, W. Shi, W. Ding, T. M. Mitchell, and B. A. Myers, "Appinite: A multi-modal interface for specifying data descriptions in programming by demonstration using natural language instructions," in *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Los Alamitos, CA, USA: IEEE Computer Society, 2018, pp. 105–114.

[9] E. Luger and A. Sellen, ""like having a really bad pa": The gulf between user expectation and experience of conversational agents," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 5286–5297.

[10] W. Xu, "Toward human-centered ai: A perspective from human-computer interaction," *Interactions*, vol. 26, no. 4, p. 42–46, Jun. 2019.

[11] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, ser. IUI '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 227–236.

[12] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–12.

[13] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15.

[14] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important? a study of deployed, low-cost intelligent interactive systems," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, ser. IUI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 169–178.

[15] B. Feldman, "Clippy Didn't Just Annoy You — He Changed the World," Oct. 2016. [Online]. Available: https://nymag.com/vindicated/2016/10/clippy-didnt-just-annoy-you-he-changed-the-world.html [Accessed: 01-June-2020].

[16] N. Baym, L. Shifman, C. Persaud, and K. Wagman, "Intelligent failures: Clippy memes and the limits of digital assistants," *AoIR Selected Papers of Internet Research*, vol. 2019, Oct. 2019.

[17] A. Maedche, S. Morana, S. Schacht, D. Werth, and J. Krumeich, "Advanced user assistance systems," *Business & Information Systems Engineering*, vol. 58, pp. 367–370, 2016.

[18] J. Cranshaw, E. Elwany, T. Newman, R. Kocielnik, B. Yu, S. Soni, J. Teevan, and A. Monroy-Hernández, "Calendar.help: Designing a workflow-based scheduling agent with humans in the loop," *ACM CHI Conference on Human Factors in Computing Systems*, January 2017.

[19] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 258–262.

[20] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?" in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 189–201.

[21] R. F. Kizilcec, "How much information? effects of

transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 2390–2395.

[22] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? the effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 1–10.

[23] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 2119–2128.

[24] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12.

[25] J. Krause, A. Perer, and E. Bertini, "A user study on the effect of aggregating explanations for interpreting machine learning models," *In Proceedings of KDD 2018 Workshop on Interactive Data Exploration and Analytics (IDEA'18)*, Aug 2018.

[26] S. Stumpf, S. Skrebe, G. Aymer, and J. Hobson, "Explaining smart heating systems to discourage fiddling with optimized behavior," in *IUI Workshops*, 2018.

[27] J. Cho and E. Rader, "The role of conversational grounding in supporting symbiosis between people and digital assistants," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW1, May 2020.

[28] M. Jain, R. Kota, P. Kumar, and S. N. Patel, *Convey: Exploring the Use of a Context View for Chatbots*. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–6.

[29] B. Galitsky, *Chatbot Components and Architectures*. Cham: Springer International Publishing, 2019, pp. 13–51.

[30] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[31] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13.

[32] A. Renkl, "Toward an instructionally oriented theory of example-based learning," *Cognitive science*, vol. 38, 09 2013.

[33] A. Renkl, T. S. Hilbert, and S. Schworm, "Example-based learning in heuristic domains: A cognitive load theory account," *Educational Psychology Review*, vol. 21, pp. 67–78, 2009.

[34] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS Q.*, vol. 38, no. 1, p. 73–100, Mar. 2014.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations." in *AAAI*, vol. 18, 2018, pp. 1527–1535.

[36] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," 2017.

[37] M. Colombo, L. Bucher, and J. Sprenger, "Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance," *Frontiers in Psychology*, vol. 8, p. 1430, 2017.

[38] S. Read and A. Marcus-Newhall, "Explanatory coherence in social explanations: A parallel distributed processing account." 1993.

[39] D. Norman, "The Design of Everyday Things." in *The Design of Everyday Things.*, 2013, p. p.39.

[40] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *CoRR*, vol. abs/1706.07269, 2017.

[41] S. Coppers, J. Van den Bergh, K. Luyten, K. Coninx, I. van der Lek-Ciudin, T. Vanallemeersch, and V. Vandeghinste, "Intellingo: An intelligible translation environment," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–13.

[42] B. J. Dietvorst, J. Simmons, and C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of experimental psychology. General*, vol. 144 1, pp. 114–26, 2015.

[43] A. Dasgupta, J.-Y. Lee, R. Wilson, R. Lafrance, N. Cramer, K. Cook, and S. Payne, "Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, pp. 1–1, 08 2016.

[44] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82 – 115, 2020.

[45] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in hci," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 493–502.

[46] J. Taylor, "Is Microsoft Excel Hard To Learn?" Jan. 2019. [Online]. Available: https://basiccompute

rtips.com/is-microsoft-excel-hard-to-learn/ [Accessed: 01-June-2020].

[47] J. Williams, N. B. Niraula, P. Dasigi, A. Lakshmiratan, C. Garcia Jurado Suarez, M. Reddy, and G. Zweig, "Rapidly scaling dialog systems with interactive learning," January 2015.

[48] M. Rosala, "Status Trackers and Progress Updates: 16 Design Guidelines," Feb. 2019. [Online]. Available: https://www.nngroup.com/articles/status-tracker-progress-update/ [Accessed: 21-June-2020].

[49] A. Mura, "Why, How, And When To Use Walkthroughs To Enhance UX." [Online]. Available: https://usabilitygeek.com/use-walkthroughs-enhance-ux/ [Accessed: 21-June-2020].

[50] L. Michael, Xieyang, "lxieyang/chrome-extension-boilerplate-react." [Online]. Available: https://github.com/lxieyang/chrome-extension-boilerplate-react [Accessed: 10-Jan-2021].

[51] J. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative Sociology*, vol. 13, pp. 3–21, 1990.

[52] J. Grudin and R. Jacques, "Chatbots, humbots, and the quest for artificial general intelligence," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–11.

[53] D. Jovic, "The Future is Now – 37 Fascinating Chatbot Statistics," Aug. 2020. [Online]. Available: https://www.smallbizgenius.net/by-the-numbers/chatbot-statistics/#gref [Accessed: 21-Sept-2020].

[54] E. Fast, B. Chen, J. Mendelsohn, J. Bassen, and M. S. Bernstein, "Iris: A conversational agent for complex tasks," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–12.