To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System

Akshit Gupta*
a.gupta-20@student.tudelft.nl
Delft University of Technology
Netherlands

Debadeep Basu*
d.basu-1@student.tudelft.nl
Delft University of Technology
Netherlands

Ramya Ghantasala*
r.p.ghantasala@student.tudelft.nl
Delft University of Technology
Netherlands

Sihang Qiu†
s.qiu-1@tudelft.nl
Delft University of Technology
Netherlands

Ujwal Gadiraju
u.k.gadiraju@tudelft.nl
Delft University of Technology
Netherlands

ABSTRACT

Trust is an important component of human-AI relationships and plays a major role in shaping the reliance of users on online algorithmic decision support systems. With recent advances in natural language processing, text and voice-based conversational interfaces have provided users with new ways of interacting with such systems. Despite the growing applications of conversational user interfaces (CUIs), little is currently understood about the suitability of such interfaces for decision support and how CUIs inspire trust among humans engaging with decision support systems. In this work, we aim to address this gap and answer the following question: *to what extent can a conversational interface build user trust in decision support systems in comparison to a conventional graphical user interface?* To this end, we built a text-based conversational interface, and a conventional web-based graphical user interface. These served as the means for users to interact with an online decision support system to help them find housing, given a fixed set of constraints. To understand how the accuracy of the decision support system moderates user behavior and trust across the two interfaces, we considered an accurate and inaccurate system. We carried out a 2×2 between-subjects study ($N = 240$) on the Prolific crowdsourcing platform. Our findings show that the conversational interface was significantly more effective in building user trust and satisfaction in the online housing recommendation system when compared to the conventional web interface. Our results highlight the potential impact of conversational interfaces for trust development in decision support systems.

CCS CONCEPTS

• **Information systems → Decision support systems**; **Chat**; • **Human-centered computing → Empirical studies in HCI**.

*Equal contributions.
†Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WWW '22, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512248

KEYWORDS

Conversational user interface, Decision support system, Trust, AI

ACM Reference Format:
Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512248

1 INTRODUCTION

Trust is an implicit and fundamental tenet of human existence. The world is able to function the way it does because of people's trust in the government, the financial institutions, the society, and each other. Therefore, it is imperative for technology to earn and build trust among users so as to increase adoption at a societal scale, leading to more people benefiting from the technology.

Assistive technologies like decision support systems help humans in making decisions and provide the best course of action, particularly when dealing with large amounts of data and complex variables [13, 38, 43]. In spite of the numerous advantages intelligent systems have to offer, widespread acceptance of such systems is still impeded by a lack of trust [27]. Hence, it is important to better understand factors that influence user trust in decision support systems, and how trust formation can be better facilitated.

With the swift penetration of virtual digital assistants like Amazon Alexa, Apple Siri and Google Assistant, the estimated number of people using digital assistants worldwide was projected to reach 1.8 billion by 2021 [15]. Earlier forecasts from Gartner [28] suggested that by 2020 "twenty-five percent of customer service and support operations will integrate virtual customer assistant (VCA) or chatbot technology across engagement channels" and that by 2025 50% of knowledge workers will use a virtual assistant on a daily basis. Recent developments in conversational interfaces, both text and voice, have provided users with new ways to interact with machines. For instance, recent works by Mavridis et al. [29] and Huang et al. [21] have successfully deployed conversational interfaces for crowdsourcing microtasks. Others have shown that conversational interfaces can improve engagement and experience of work [35], improve memorability of information consumed [37] or facilitate training [4]. Yet, there is a lack of understanding about how conversational interfaces shape user trust. To address this

research gap, we explore the extent to which a conversational interface facilitates trust formation in a decision support system in comparison to a conventional graphical user interface.

To understand how the accuracy of the decision support system affects user behavior and trust across the two types interfaces (CUI vs GUI), we also considered accurate and inaccurate decision support systems, and therefore carried out a 2×2 between-subjects study by recruiting crowdsourced participants from Prolific[1], who were then tasked with finding the right housing based on predefined constraints. Participants could use a decision support system which suggests housing options based on configurable filters. The rationale behind choosing this context for the decision support system was the current housing crisis in many countries including the Netherlands [11]. We first created a curated dataset representing real-world houses, and generated realistic house-hunting scenarios with a varying complexity. We then presented house-hunting scenarios to participants with either a conversational interface or a conventional web-based graphical user interface. Participants were expected to find and submit the correct house, fitting all constraints laid out in their scenario (akin to real-world housing search).

We found that **users tended to trust conversational interfaces more in comparison with conventional web-based graphical user interfaces**, while interacting with identical decision support systems for housing recommendation. Interestingly, this observation was found to be independent of the accuracy of the decision support system. Our findings can inform interaction design in other domains where a decision support system is needed, for example, assistance in selecting the right university for education, or determining the appropriate selling price of a used car. Overall, this work provides insights and has broad implications for building trust in decision support systems.

2 RELATED WORK

We position our work across different realms of related literature – approaches to trust and decision support systems in HCI, the role of interfaces in building trust, and conversational crowdsourcing.

2.1 Approaches to Trust and Decision Support Systems in HCI

Trust is a multi-faceted and multi-dimensional concept. In existing literature, trust has been explored from various contexts such as interpersonal relationships, management and employees, organizational productivity, and relationship management [26]. This context has lead to a number of definitions of trust [45]. In Rotter [40], authors define trust as "expectancy held by an individual that the word, promise or written communication of another can be relied upon". Johns [22] defines trust as "willingness to rely on an exchange partner in whom one has confidence". Mayer et al. [30] defines it as "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party". Hoff and Bashir [19] models the complexities of trust in three layers of variability: dispositional trust, situational trust and learned trust. As per this model, the trust of a human in an automation is contingent upon the individual's tendency to trust automation, the context of the interaction and past experiences with the system. Specifically, Corritore et al. [8] models trust in an online environment which includes information or transactional websites on the basis of three factors: ease of use, risk, and perception of credibility. For the context of our system, we follow the definition of trust as defined by Lee and See [26] i.e. "Trust is an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". The agent here, can be any computer technology or another human and the degree of reliance of the trustor on this agent will characterize trust.

Decision support systems are interactive systems that aid humans in making decisions when there are a number of complex variables. Decisions utilizing decision support systems (DSSs) can be made more quickly and accurately than unaided decisions [42]. More recently, Erlei et al. [12] showed that humans tend to overwrite their economic self-Interest to avoid bargaining with AI systems. The wisdom of crowds has been employed in decision support systems to improve their knowledge base. Hosio et al. [20] use crowdsourcing tasks to populate the knowledge bases in an easy and cost effective manner. Wen [47] studied the effect of a conversational interface based decision support system for stock investment activities. Yuan et al. [48] explore the requirements of a decision support system in a clinical setting, where authors concluded that user interface design and implementation were key factors for the successful deployment of clinical decision support systems. Informed by such prior work, in this study, we implemented a decision support system to help users in finding housing, and explored whether a conversational interface can better build user trust in the system.

2.2 Role of Interfaces in Shaping User Trust

The effect of interface design to make the human interaction more engaging has been widely researched. In a previous study [32], the authors explored the etiquette for human computer interaction and found that the humans already share a relationship with the computer tools. Furthermore, Nass and Lee [33] explore the software acceptance by users and finds that the software, which is more similar to humans is likely to be more readily accepted. Lee and See [26] provide several guidelines for creating trustable automations ranging from showing its past performance to conveying its purpose clearly, as well as simplifying it to make it more understandable to the user. In a prior work by Tolmeijer et al. [44], the authors propose ways to repair trust and mitigation strategies for human-robotics interaction systems. Antrobus et al. [5] explore the use spoken natural language interface (NLI) to improve trust in autonomous vehicles. It is found that while the trust was similar for both the NLI and traditional touchscreen based interface, the satisfaction and confidence of the users was higher in NLI. In a similar study for autonomous vehicles [41], it is found that interfaces such as conversational interface which mimic human traits can help in increasing people's trust. Similarly, Weitz et al. [46] found that integrating virtual agents into the explainable AI interaction led to increase of trust in intelligent systems. Following this, in our system, we postulated a hypothesis that a conversational interface that has a personality close to humans is more trustworthy.

[1] https://prolific.co

2.3 Conversational Crowdsourcing

Recent works have showcased the various benefits of using conversational interfaces for crowdsourcing [10, 23–25, 36]. Huang et al. [21] proposed Evorus, an architecture for crowd-powered conversational interface to provide high quality responses with low latency and cost by leveraging past information obtained from crowd workers. Abbas et al. [2] showed how fillers can be used to improve perceived latency in crowd-powered conversational systems [3]. Mavridis et al. [29] explored the effectiveness of conversational interfaces for crowdsourcing microtasks and found that they led to an improved worker satisfaction compared to web-based interfaces. Researchers have shown that using conversational interfaces for crowdsourcing increased worker engagement as well as worker retention compared to web interfaces [34, 35]. Furthermore, Hettiachchi et al. [18] develop Crowd Tasker which uses a digital voice assistant for crowdsourcing tasks. It was found that compared to a web interface, using a voice based interface can reduce the time and effort required for initiating tasks while providing more flexibility to the workers.

3 STUDY DESIGN

We carried out a crowdsourcing study to address the aforementioned research gap. In this section, we elaborate on the crowdsourcing task design, the decision support system, the conversational interface (**`Chat`**) and the conventional web-based graphical user interface (**`Web`**), the task scenarios, and measures used in this study.

3.1 Task Design

In the tasks, the participants were provided with a housing search scenario in a situated experiment fashion. The scenario represents a student looking for a house in Delft, the Netherlands with a certain given set of preferences. The participants were expected to interact with the search system and find the house that fits all the preferences associated with the scenario. For each of the scenarios, there was only one correct house in the dataset that fit all preferences. The participants were assigned either a conversational interface, or a typical web-based graphical user interface to find the correct house. Upon submitting the preferences, the participants were provided with a house selected by the system based on the constraints entered. At this stage, the participant could either submit the house recommended by the system, or manually check all available houses and find the correct house which matched all the constraints. The actions and the decisions available to a participant were identical across both the interfaces. Figure 1 presents an overview of the interaction between the participant and the interfaces.

3.2 A Decision Support System for Housing

We built a housing recommender system to serve as a decision support system in our study. We considered the potential role of **accuracy** of the decision support system in shaping user trust and behavior. To this end, the system was configured to be either accurate (high accuracy) or inaccurate (low accuracy). In the **high accuracy** conditions, the system recommends the house that correctly fulfills all the constraints given by the user (assuming that the user enters all constraints correctly), while for **low accuracy** conditions, a random incorrect house was selected from the list of all available houses. As described earlier, the decision support system was presented to the participants using either a conventional web-based graphical user interface, or a conversational interface.

3.2.1 Web-Based Graphical User Interface. The web-based graphical user interface (**`Web`**) is a website designed as a portal for searching houses. The web GUI task and its workflow is shown in Figure 2. In the web GUI task, participants are directed to a screen displaying the scenario, and encounter an attention check question asking for the name of the persona described in the scenario (**w1**). Only when participants submit the correct name, they are directed to a page to specify the constraints (**w2**) given in the scenario. After submitting the constraints (**w3**), they are shown the house recommended by the DSS. Participants can either choose to submit the house recommended by the DSS (**w4**), or check the list of available houses (**w5**). If a participant chooses to view all the available houses, a list of houses is retrieved from the database and is displayed to the user along with the house recommended by the DSS. On submitting a house, participants are asked to confirm their house selection (**w6**) or reset filters. If a participant chooses to reset filters, the constraints specified previously are cleared and they are redirected to the constraints submission page. The participant also has the option to view the DSS recommended house (**w7**) after choosing to view all available houses. The web GUI task ends when the participant submits and confirms a selected house, after which they are directed to the next step in the workflow as shown in Figure 4.

The web interface was built using React. All user actions in the interface were logged using Node.js and Express, and stored in a MongoDB database. The interface, including the APIs, is hosted on a Heroku server (https://www.heroku.com/).

3.2.2 Conversational Interface. The conversational interface (**`Chat`**) features a text-based conversational agent with which participants can engage to specify their constraints. Figure 3 gives an overview of the interface. In this task, the participant is provided with a scenario text eliciting the housing constraints of a student in the situated experiment (**c1**). The participants are expected to converse with the conversational agent to provide their housing constraints, unlike the web graphical user interface where participants are provided a list of preferences to enter. The conversational agent initiates a conversation by greeting the participants and asking for the names assigned to them in the scenario (**c2**). This first prompt also serves as an attention check for the participant; in that the conversational agent does not proceed until the correct name associated with the scenario is entered. The conversational agent then proceeds to engage in open-ended conversations with the participants and the conversation can be steered by either the participants or the agent. The participants are free to input either free text or choose one of the suggestions buttons presented. The conversation proceeds until the participants convey that they do not have more preferences to convey, after which they are presented with a housing recommendation (**c3**). At this stage, the participants have the option to either submit the recommended housing option or look at all the houses in the system and select one of them (**c4**). Participants can also reset all the constraints if they think they may have committed a mistake. Once the participants are satisfied with the housing option and submit the house, the

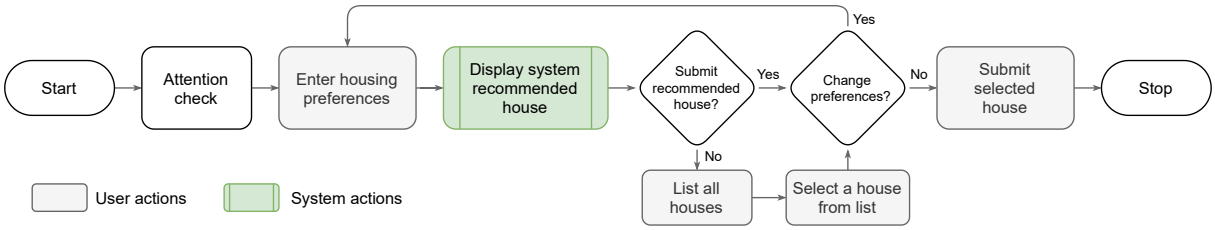

Figure 1: Overview of interaction between the participant and the house search interfaces.

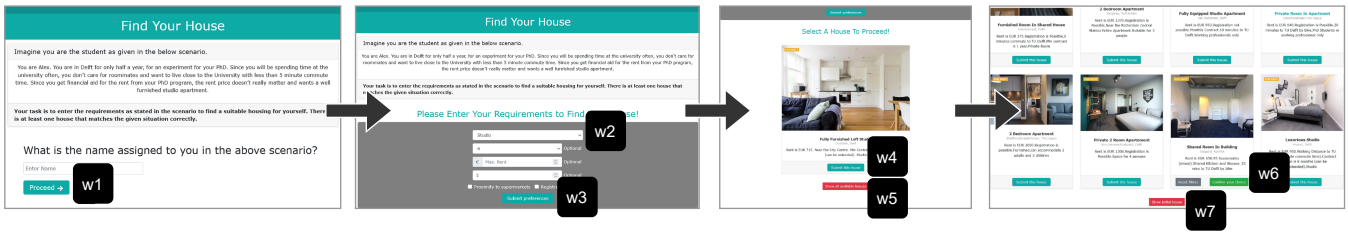

Figure 2: The web-based graphical user interface task and its workflow.

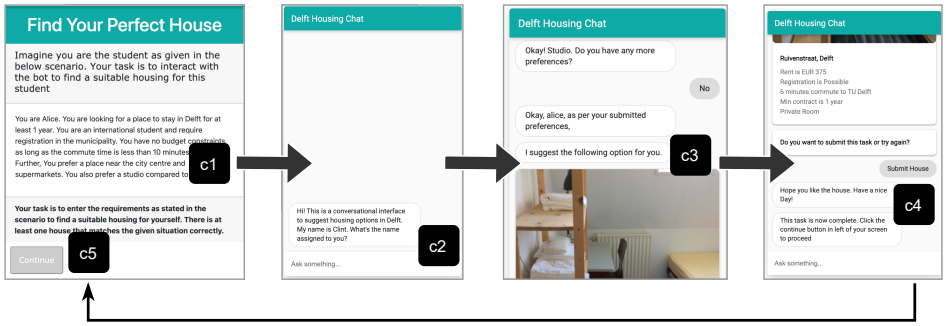

Figure 3: The conversational interface and its workflow.

continue button is activated to allow participants to move towards the next step of the workflow (c5).

The conversation agent follows a frame based architecture [17] which is built on top of Dialogflow Messenger [1]. The backend of the agent is built on a Node.js web app deployed on a Heroku server. This web app provides appropriate responses for each of the intents and slots conveyed by the participant in the form of responses to each POST request originating from the front end. Other parts of the user interface are built using vanilla HTML and CSS.

3.3 Dataset and Scenarios

The dataset for housing options was manually created by scraping housing options from real online housing sites (such as `housinganywhere.com` and `kamernet.nl`). The houses were chosen such that each one of them has the the following properties:

(1) **Housing type**: The housing type had four options - studio, apartment, private room or sharing.
(2) **Duration**: The amount of time in months, that the user needs the house.
(3) **Rent**: The maximum rent of the housing option.
(4) **Proximity to the supermarket**: Whether or not a house is close to the supermarket.
(5) **Registration**: Whether or not a resident at the housing can be registered at the municipality.

We manually created six scenarios representing students looking for housing options with specific preferences. These scenarios had two different levels of complexity - **easy** and **hard**. In the easy scenarios, participants were supposed to find a house according to 3 given preferences. The hard scenarios had 5 preferences each. Table 1 gives an example of an easy scenario and a hard scenario. MongoDB was used both for storing data pertaining to the houses and scenarios, and for logging user interactions in the tasks.

3.4 Measures

3.4.1 Measuring Affinity for Technology. Attig et al. [6] showed that the affinity a user has towards technology interaction could be seen as a subset of the user's personality, and can be useful in helping them cope with technology successfully. In order to understand

Table 1: Example easy and hard scenarios given to users in each task. Preferences in each scenario are highlighted in bold.

Complexity	Scenario
Easy	Your name is Cece. You are looking for a student house in Delft for a duration for at least **6 months**. You are an international student and **need to be registered** at the Delft municipality. You have a **maximum budget of 550 euros**. You don't mind sharing a flat with others as long as she has her **own room**. You also prefer to stay **near supermarkets** so that you can shop for groceries easily.
Hard	You are Alice. You are looking for a place to stay in Delft for at least **1 year**. You are an international student and **require registration** in the municipality. You have **no budget constraints** as long as the **commute time is less than 10 minutes** by bike. Further, You prefer a place **near the city centre and supermarkets**. You also prefer a **studio** compared to sharing.

the tendency of the participants of our study to actively engage in interacting with either web or conversational interfaces, we used the 9-item 'Affinity for Technology Interaction' (ATI) questionnaire based on 6-point Likert scales ranging from *Completely Disagree* to *Completely Agree* [14]. The questionnaire is presented to the participants prior to the housing search task and is tailored to the interface they are expected to interact with.

3.4.2 User Behavior. We analyse behavior of participants based on 3 aspects: *the correctness of the submission*, *the time spent during the task*, *and whether all the available houses were browsed*. Since each of the scenarios contains a set of constraints which are satisfied by only one particular house in the database, we check the correctness of user's submission to investigate whether the accuracy of the decision support system (either low or high) affects user behavior. Furthermore, we measure users' active task execution time to understand how different interfaces can affect their behavior.

3.4.3 Measuring Trust in the System. To measure the trust that a participant emulates in the interface used to complete the scenario, we use a shortened version of the widely used "Recommender systems' Quality of user experience" questionnaire [39] which consists of the four main components of recommender systems useful in modeling user trust. We use a subset of the questionnaire, consisting of 26 questions divided into 8 categories. The questions are answered using a 5-point Likert scale ranging from *1: Completely Disagree* to *5: Completely Agree*. Negatively worded questions are reverse coded to maintain uniformity. A 'trust score' is obtained for each of the responses provided by a participant by averaging over the scores of all the components of the questionnaire.

3.4.4 Measuring Satisfaction towards the System. For measuring the satisfaction of the users towards the interfaces, we use a subset of the shortened "Recommender systems' Quality of user experience" questionnaire [39] used in the measurement of trust. The subsets included in the measurement of satisfaction towards an interface were the Quality of Recommendations, Interface Adequacy, Interaction Adequacy, Ease of Use, Usefulness of the interface, and Control and Transparency. The 'satisfaction score' is obtained for each response by computing the mean scores of the aforementioned parameters.

4 EXPERIMENTAL SETUP

4.1 Experimental Conditions

We carried out a controlled crowdsourcing experiment with a 2×2 between-subject design. The independent variables were the user interface (**`Chat`** vs **`Web`**) and the accuracy of the housing recommender system (**high** accuracy vs **low** accuracy), resulting in four experimental conditions:

1) Web-Low represents the condition in which participants were asked to use the web-based graphical user interface to find a suitable house, with the recommender system configured to provide inaccurate suggestions.

2) Web-High represents the condition in which participants were asked to use the web-based graphical user interface to find a suitable house, with the recommender system configured to provide accurate suggestions.

3) Chat-Low represents the condition in which participants were asked to find a suitable house through a conversation with the conversational interface, and featuring the recommender system configured to provide inaccurate suggestions.

4) Chat-High represents the condition that participants are asked to find a suitable house through a conversation with the conversational interface, and featuring a recommender system configured to provide accurate suggestions.

In each condition, to maximize the chance of interaction between the participant and the user interface, we ask each participant to complete two house finding tasks (one relatively easy scenario and one relatively hard scenario, as shown in Table 1). The order of performing the two difficulty-level tasks is evenly distributed, meaning in each condition, 50% of workers first perform the house finding task in an easy scenario followed by a hard scenario, while the other 50% perform the two tasks in reverse order. Through such counterbalancing we avoid potential order effects.

4.2 Procedure

Participants for the study were recruited from the Prolific crowdsourcing platform. The crowd workers were invited to participate in a study called "*Test a Housing Recommendation System*". A total of four single session studies were created according to the setup outlined in Section 4.1. To ensure reliable and unique participation for the experiments, only workers with a minimum approval rate of 90% were allowed to participate, and workers were allowed to participate in only a single experimental condition. We recruited 60 workers for each condition (30 workers first complete an easy task, followed by hard task and 30 workers first complete a hard task, followed by an easy task). Thus, $60 \times 4 = 240$ unique crowd workers from Prolific participated in our study. Participants in our study received 1.25 GBP upon successful completion of the task. According to the report from Prolific, the actual average hourly

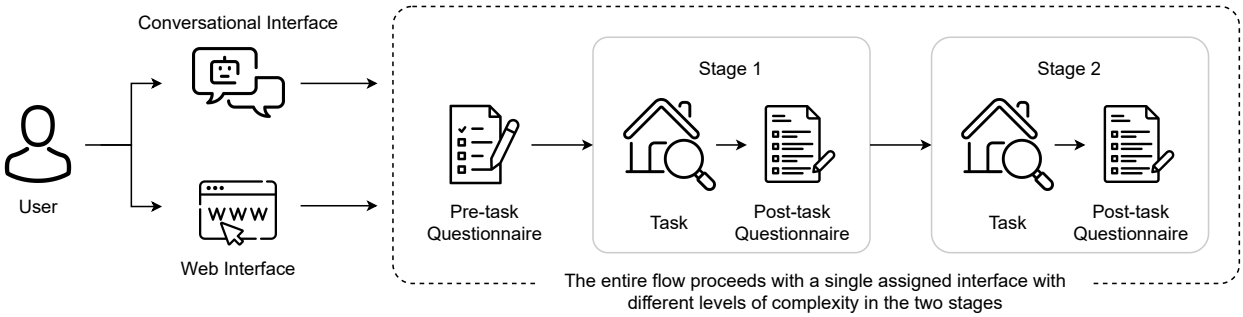

Figure 4: An overview of the study workflow.

reward of our study was nearly 7.5 GBP (which is considered a good payment on the Prolific platform).

The participants of the study were provided with a set of instructions, and on their consent, were redirected to the appropriate interface based on the experimental condition. An overview of the procedure involved in the study is shown in Figure 4.

In the first stage of the study, the participants were asked to answer a set of pre-task questionnaires related to the interface they were going to use in the experiment. The questions were based on the ATI Scale. The participants were then directed to the task consisting of a housing search scenario using either the chat interface or the web graphical user interface. They were then directed to the post-task questionnaire based on ResQue, regarding the recommendation provided by the system. On completing the questionnaire, the participants were redirected to a transition page, from which they could continue to the second stage of the study.

The second stage consisted of a second housing search task with a different scenario but with the same interface used in the first stage. The complexity of the scenario was either hard or easy based on the complexity in the first stage; each participant received one easy and one hard scenario. Upon completing the task, the participants were asked to fill in another post-task questionnaire regarding the recommendation provided by the system in the second stage. Finally, the participants were provided with a completion code which they could enter on Prolific to claim their compensation.

5 RESULTS AND ANALYSIS

We filtered out participants with incomplete submissions before carrying out our analysis. A few participants had completed only one scenario and some had submitted the same task multiple times by interacting with the conversational agent again after submitting a house. A total of 222 valid submissions (111 unique participants) were obtained for the conversational interface (**`Chat`**), and 234 valid submissions (117 unique participants) for the web-based graphical user interface (**`Web`**). We publicly shared all the data and the code (including the analysis of affinity to technology interaction) for the benefit of the community.[2]

5.1 User Behavior with the CUI

The analysis of user behavior with the conversational interface is shown in Table 2. From the analysis, it is clear that the user performance in terms of finding the correct house was higher for the condition with high accuracy than low-accuracy condition. It also can be seen that user performance was mostly similar for easy and complex scenarios. Further, for the condition with low accuracy, participants were more inclined to distrust the suggestion given by the systems, and instead, looked at the complete list of houses in the system. Moreover, the time spent by participants was longer on complex scenarios, and in the conditions with low accuracy.

Table 2: Conversational interface user behaviour analysis.

		Correct Submissions (%)	Time Spent (mins)	Submissions looking at all houses (%)
System accuracy	High (N = 116)	65%	3.5 ± 2.25	38%
	Low (N = 106)	42%	3.91 ± 2.86	65%
Scenario complexity	Easy (N = 111)	55%	3.24 ± 2.35	50%
	Hard (N = 111)	53%	4.17 ± 2.69	52%
Overall (N = 222)		54%	3.70 ± 2.57	51%

5.2 User Behavior with the GUI

The analysis of user behavior with the web interface is shown in Table 3. We find that around 75% of the submissions were manually selected by the participants (the house submitted by the user was not the one recommended by the system), which may allude to a distrust in the system. Comparing conditions with high accuracy and conditions with low accuracy, it can be seen that there were marginally more correct submissions in the high-accuracy condition (62.931%) than in the low-accuracy condition (52.542%). Evidently, participants spent almost a minute more in the condition with low accuracy than the condition with high accuracy. While contrasting the easy and complex conditions, we observed that the split between correct and incorrect submissions for the complex scenarios is almost even, while the percentages are in favour of correct submissions for the easy scenarios (64.957%). Interestingly, the time spent for complex scenarios was less than that of easy scenarios. This can be explained by the fact that there were more submissions of the system recommended house for complex scenarios than in the case of easy scenarios.

5.3 Analysing Trust Across Interfaces

The trust scores of the interfaces were obtained by computing the mean scores of the post-task questionnaire provided to the participants. In Table 4, we show the descriptive statistics for the trust

[2]https://sites.google.com/view/www22trust

Table 3: Web interface user behaviour analysis.

		Correct Submissions (%)	Time Spent (mins)	Submissions looking at all houses (%)
System accuracy	High (N = 116)	63%	4.60 ± 2.42	72%
	Low (N = 118)	53%	5.30 ± 2.70	86%
Scenario complexity	Easy (N = 117)	65%	4.96 ± 2.57	82%
	Hard (N = 117)	50$	4.94 ± 2.61	75%
Overall (N = 234)		58%	4.95 ± 2.58	79%

scores of the two interfaces, in two levels of system accuracy. For conditions with low accuracy, the conversational interface obtained a mean trust score of 3.445 ± 0.795 from 106 responses, while the web interface obtained a mean trust score of 2.371 ± 0.6, from 118 responses. For conditions with high accuracy, the conversational interface obtained a mean trust score of 3.870 ± 0.595 from 116 responses, while the web interface obtained a mean score of 2.353 ± 0.642 from 116 responses.

Table 4: Descriptive statistics for trust score and satisfaction score grouped by interface type and accuracy level.

User interface	System accuracy	Trust score ($M \pm SD$)	Satisfaction score ($M \pm SD$)
Chat Interface	Low accuracy (N = 106)	3.445 ± 0.795	3.511 ± 0.810
	High accuracy (N = 116)	3.870 ± 0.596	3.945 ± 0.613
Web Interface	Low accuracy (N = 118)	2.371 ± 0.600	2.254 ± 0.505
	High accuracy (N = 116)	2.353 ± 0.642	2.208 ± 0.578

A two-way ANOVA test was performed to analyse the effect of interface type and accuracy of the system on the trust score. The results (Table 5) showed significant effects of both interface type and system accuracy, and a significant interaction effect of the interface type and accuracy on the trust score. A post-hoc Tukey test (Table 6) showed that the trust score did not differ significantly for the web interface with low-accuracy condition against the web interface with high-accuracy condition. The comparisons of conversational interface with web interface, with both high-accuracy and low-accuracy conditions, as well as the web interface with low-accuracy against the conversational interface with high-accuracy showed significant difference in trust scores.

This suggests that despite the differences in level of accuracy, the participants tended to trust the conversational interface more, in comparison with the web interface.

Table 5: Results of a two-way ANOVA on trust score against interface type and accuracy.

Cases	Sum of Squares	df	Mean Square	F	p	VS-MPR*
Interface (Chat vs Web)	182.829	1	182.829	420.623	< .001	3.264e+62
Accuracy (Low vs High)	4.529	1	4.529	10.420	0.001	41.467
Interface * Accuracy	5.353	1	5.353	12.316	< .001	97.473

5.4 User Satisfaction Across Interfaces

The satisfaction scores of the interfaces were obtained by computing the mean scores of the interface quality and usability parameters of the post-task questionnaire provided to the workers. These included the Quality of Recommendations, Interface Adequacy, Interaction Adequacy, Ease of Use, Usefulness of the interface, and Control and Transparency. In Table 4, we see the descriptive statistics for the satisfaction scores for the two interfaces, moderated by the system accuracy. For low-accuracy conditions, the conversational interface obtained a mean trust score of 3.511 ± 0.810 from 106 responses, while the web interface obtained a mean trust score of 2.254 ± 0.505, from 118 responses. For the high accuracy condition, the conversational interface obtained a mean trust score of 3.945 ± 0.613 from 116 responses, while the web interface obtained a mean score of 2.208 ± 0.578 from 116 responses.

Table 6: Post-hoc comparisons of interface moderated by accuracy on trust score.

		Mean Difference	SE	t	p_{tukey}
Chat-Low *vs.*	Web-Low	1.074	0.090	11.982	< .001
	Chat-High	-0.426	0.092	-4.608	< .001
	Web-High	1.091	0.090	12.132	< .001
Web-Low *vs.*	Chat-High	-1.499	0.089	-16.907	< .001
	Web-High	0.018	0.086	0.206	0.997
Chat-High *vs.*	Web-High	1.517	0.089	17.039	< .001

Table 7: Results of a two-way ANOVA on the satisfaction score against interface type and accuracy.

Cases	Sum of Squares	df	Mean Square	F	p	VS-MPR*
Interface (Chat vs Web)	244.245	1	244.245	616.662	< .001	7.948e+81
Accuracy (Low vs High)	4.099	1	4.099	10.350	0.001	40.194
Interface * Accuracy	6.257	1	6.257	15.798	< .001	474.041

Table 8: Post-hoc comparisons of interface moderated by accuracy on satisfaction score.

		Mean Difference	SE	t	p_{tukey}
Chat-Low *vs.*	Web-Low	1.258	0.086	14.701	< .001
	Chat-High	-0.434	0.088	-4.919	< .001
	Web-High	1.303	0.086	15.175	< .001
Web-Low *vs.*	Chat-High	-1.691	0.085	-19.979	< .001
	Web-High	0.046	0.082	0.555	0.945
Chat-High *vs.*	Web-High	1.737	0.085	20.436	< .001

A two-way ANOVA test was performed to analyse the effect of interface type and accuracy of the system on the satisfaction score. Similar to the results of trust scores, the results of user satisfaction (Table 7) show significant effects of the interface type and system accuracy, and a significant interaction effect of the interface type and accuracy on the satisfaction score. A post-hoc Tukey test (Table 8) showed that the satisfaction score did not differ significantly for the web-based graphical user interface with low accuracy condition against the web interface with high accuracy condition. The comparisons between conversational interface and web interface across both high-accuracy and low-accuracy conditions, and the web interface with low accuracy against the conversational interface with high accuracy showed significant differences in satisfaction score.

This analysis shows that the inaccurate recommendations caused participants to be less satisfied with the conversational interface when compared with accurate recommendations. However, for the web interface there is no significant difference in the levels of satisfaction. It is also interesting to note that the participants were more satisfied with the conversational interface than with the web interface irrespective of the accuracy of the system.

6 DISCUSSION

It is clear that users tended to trust the decision support system while using a conversational interface more than they did while using a web-based graphical user interfaces. This result was found to be independent of the accuracy of the underlying system they were interacting with. Having said that, we found a significant difference in user trust between the system configured with low-accuracy condition (3.45 ± 0.80) and system with high-accuracy condition (3.87 ± 0.60) for the conversational interface conditions. This suggests that the system accuracy can play a role in regulating user trust in decsision support systems when using conversational interfaces. Whereas, for the web-based graphical user interface, we did not find any significant difference across trust scores in conditions with low or high system accuracy respectively (2.37 ± 0.60 and 2.35 ± 0.64).

Similarly, it was also clear from the results that users were more satisfied with using the conversational interface over the web-based graphical user interface, irrespective of the accuracy of the condition. We also noted a similar significant difference in the satisfaction score between the conversational interface with low-accuracy and high-accuracy conditions (3.51 ± 0.81 and 3.95 ± 0.61 respectively), while the web-based graphical user interface did not show any significant differences between low-accuracy and high-accuracy conditions (2.25 ± 0.51 and 2.21 ± 0.58 respectively).

6.1 Trust vs. Performance and the Role of Time

Interestingly, we found that the overall task completion time for conversational interfaces was significantly shorter than web-based graphical user interfaces (as can be seen in Table 2 and Table 3). A possible reason for this might be that since users trust the conversational interface more, they were less inclined to change the constraints once entered, after a recommendation was given. Whereas for the the web interfaces, due to less trust in the system, the users may have tended to be more careful in cross-checking the constraints, thus increasing the completion time. This is further substantiated by looking at the percentage of correct submissions in case of low accuracy conditions for both the interfaces. We found that for the conversational interfaces with low-accuracy condition, 43% of the submissions were correct, whereas, for the web interfaces with low-accuracy condition, 53% of the submissions were still correct. Although researchers have paid attention to the trust and work performance [16, 31], most previous studies mainly focused on the output quality and the time [7, 9, 34, 35]. Our work suggests a three-way trade-off between trust as a result of the interface, active completion time, and user performance.

6.2 Implications for Designing DSS

The results suggest that for decision support systems of the future, the choice of interface can have a major impact in the development of both trust and user satisfaction. Decision support system designers should be aware that conversational interfaces can potentially be more trustworthy in general. This also suggests that conversational interfaces should not be abused, since the goal of designing a proper user interface is to elicit appropriate system reliance by building appropriate trust between the user and the system, rather than over-trust or under-trust.

Furthermore, for conversational interfaces, the accuracy of the system has an impact on the satisfaction of the user, whereas the user satisfaction is not as affected by the accuracy of suggestions on a web-based graphical user interface. The results on trust and satisfaction across both the decision support systems with low-accuracy and high-accuracy conditions convey that, the accuracy of a decision support system in case on a conversational interface is more detrimental in developing trust and user satisfaction, than on a web-based graphical user interface.

6.3 Limitations and Future Work

Our experimental study was limited to a single domain (i.e., a decision support system for housing search). Although one may argue that our findings can generalize to similar contexts with conversational interfaces and decision support systems, further explorations are required to corroborate these insights. This may range from the control interfaces in cars to robotics as well as in the e-commerce domain. Further, it would be interesting to see how the trust in conversational interfaces evolves in a long term as the system accuracy improves or deteriorates. In the imminent future, we aim to study trust formation in decision support systems with voice-based conversational user interfaces.

7 CONCLUSIONS

In this work, we investigated the role of conversational interfaces in shaping user trust in a decision support system and explored whether accuracy of the system moderates the role of the interface. We designed novel conversational interfaces and used conventional web-based graphical user interfaces for a decision support system in the form of a housing recommender system. We recruited 240 participants from a crowdsourcing platform and carried out a between-subjects study. We found that the conversational interface was significantly more effective in building user trust and satisfaction in the decision support system when compared to the web-based interface. We found that this was consistent across conditions with varying accuracy of the housing recommender system. Our study highlights the impact of conversational interfaces in human computer interaction for trust development. These findings have valuable insights and implications for system designers to build trustworthy decision support systems in the future.

ACKNOWLEDGMENTS

This work was partially supported by the Delft Design@Scale AI Lab, and the 4TU.CEE UNCAGE project.

REFERENCES

[1] [n.d.]. Dialogflow Messenger | Dialogflow ES | Google Cloud. https://cloud.google.com/dialogflow/es/docs/integrations/dialogflow-messenger

[2] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 2–14.

[3] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia Barakova, and Panos Markopoulos. 2020. Crowd of oz: a crowd-powered social robotics system for stress management. *Sensors* 20, 2 (2020), 569.

[4] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 3–12.

[5] Vicki Antrobus, Gary Burnett, and David Large. 2018. 'Trust me – I'm AutoCAB': Using natural language interfaces to improve the trust and acceptance of level 4/5 autonomous vehicles.

[6] Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing Personality Differences in Human-Technology Interaction: An Overview of Key Self-report Scales to Predict Successful Interaction. 19–29. https://doi.org/10.1007/978-3-319-58750-9_3

[7] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Andrea Mauri. 2013. Reactive Crowdsourcing. In *Proceedings of the 22Nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. ACM, New York, NY, USA, 153–164.

[8] Cynthia L. Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58, 6 (2003), 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7 Trust and Technology.

[9] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

[10] Vincenzo Della Mea, Eddy Maddalena, and Stefano Mizzaro. 2015. Mobile Crowdsourcing: Four Experiments on Platforms and Tasks. *Distrib. Parallel Databases* 33, 1 (March 2015), 123–141.

[11] TU Delta. 2020. Room Shortage in Delft. Retrieved October 31, 2021 from https://www.delta.tudelft.nl/article/room-shortage-delft-will-continue-increase.

[12] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-Interest to Avoid Bargaining With AI Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

[13] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.

[14] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150

[15] Go-Gulf. 2018. The Rise of Virtual Digital Assistants Usage – Statistics and Trends. Retrieved June 12, 2020 from https://www.go-gulf.com/virtual-digital-assistants/.

[16] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[17] Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan 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 Green Open Access added to TU Delft Institutional Repository 'You share, we take care!' – Taverne project https://www.openaccess.nl/en/you-share-we-take-care Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

[18] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I Am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376320

[19] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 arXiv:https://doi.org/10.1177/0018720814547570 PMID: 25875432.

[20] Simo Hosio, Jorge Goncalves, Theodoros Anagnostopoulos, and Vassilis Kostakos. 2016. Leveraging Wisdom of the Crowd for Decision Support. (Jan 2016). https://doi.org/10.14236/ewic/hci2016.38

[21] Ting-Hao (Kenneth) Huang, Joseph Chee Chang, and Jeffrey P. Bigham. 2018. Evorus: A Crowd-Powered Conversational Assistant Built to Automate Itself Over Time. 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, 1–13. https://doi.org/10.1145/3173574.3173869

[22] J L Johns. 1996. A concept analysis of trust. *Journal of advanced nursing* 24 1 (1996), 76–83.

[23] Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

[24] Pavel Kucherbaev, Azad Abad, Stefano Tranquillini, Florian Daniel, Maurizio Marchese, and Fabio Casati. 2016. CrowdCafe-Mobile Crowdsourcing Platform. *arXiv preprint arXiv:1607.01752* (2016).

[25] Abhishek Kumar, Kuldeep Yadav, Suhas Dev, Shailesh Vaya, and G. Michael Youngblood. 2014. Wallah: Design and Evaluation of a Task-centric Mobile-based Crowdsourcing Platform. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (London, United Kingdom) *(MOBIQUITOUS '14)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 188–197.

[26] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80.

[27] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[28] Technology Magazine. 2018. Gartner Says 25 Percent of Customer Service Operations Will Use Virtual Customer Assistants by 2020. https://www.technologymagazine.com/ai/gartner-virtual-assistants-feature-25-customer-services-2020

[29] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 243–251. https://doi.org/10.1145/3320435.3320439

[30] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[31] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2271–2282.

[32] Christopher A Miller. 2002. Definitions and dimensions of etiquette. In *Proc. AAAI Fall Symposium on Etiquette and Human-Computer Work*.

[33] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 171–181. https://doi.org/10.1037/1076-898x.7.3.171

[34] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[35] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. Association for Computing Machinery, New York, NY, USA. https://dl.acm.org/doi/fullHtml/10.1145/3313831.3376403

[36] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Ticktalkturk: Conversational crowdsourcing made easy. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 53–57.

[37] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Towards memorable information retrieval. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 69–76.

[38] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.

[39] Li Chen Rong Hu and Pearl Pu. 2010. ResQue. Retrieved June 11, 2020 from https://hci.epfl.ch/research-projects/resque/.

[40] Julian B. Rotter. 1967. A new scale for the measurement of interpersonal trust1. *Journal of Personality* 35, 4 (1967), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

[41] Peter A. M. Ruijten, Jacques M. B. Terken, and Sanjeev N. Chandramouli. 2018. Enhancing Trust in Autonomous Vehicles through Intelligent User Interfaces That Mimic Human Behavior. *Multimodal Technologies and Interaction* 2, 4 (2018). https://doi.org/10.3390/mti2040062

[42] Peter Todd and Izak Benbasat. 2000. Inducing compensatory information processing through decision aids that facilitate effort reduction: an experimental assessment. *Journal of Behavioral Decision Making* 13, 1 (2000), 91–106. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<91::AID-BDM345>3.0.CO;2-A

[43] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87.

[44] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. 2020. Taxonomy of Trust-Relevant Failures and Mitigation Strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/3319502.3374793

[45] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[46] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3308532.3329441

[47] M. Wen. 2018. A conversational user interface for supporting individual and group decision-making in stock investment activities. In *2018 IEEE International Conference on Applied System Invention (ICASI)*. 216–219.

[48] Michael Juntao Yuan, George Mike Finley, Ju Long, Christy Mills, and Ron Kim Johnson. 2013. Evaluation of User Interface and Workflow Design of a Bedside Nursing Clinical Decision Support System. *Interactive Journal of Medical Research* 2, 1 (2013). https://doi.org/10.2196/ijmr.2402