This paper might be a pre-copy-editing .pdf of an article submitted for publication. Some parts might be omitted for privacy reasons.

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 research question: *how does a conversational interface compare to a typical web-based graphical user interface for building trust in the context of decision support systems?* To this end, we built two distinct user interfaces: 1) a text-based conversational interface, and 2) a conventional web-based graphical user interface. Both of these served as interfaces to an online decision support system for suggesting housing options to the participants, given a fixed set of constraints. We carried out a 2×2 between-subjects study on the Prolific crowdsourcing platform. Our findings present clear evidence that suggests 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 web-based decision support systems.

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 *WWW '22: ACM The Web Conference, April 25–29, 2022, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

WWW '22, April 25–29, 2022, Lyon, France

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

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 its users so as to increase adoption and become an integral part of society.

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 [30, 35]. In spite of the numerous advantages intelligent systems have to offer, widespread acceptance of such systems is still impeded by a lack of trust [21]. 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 is projected to reach 1.8 billion by 2021 [10]. Further, according to Gartner [22], by 2020 "twenty-five percent of customer service and support operations will integrate virtual customer assistant (VCA) or chatbot technology across engagement channels". 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. [23] and Huang et al. [16] have successfully deployed these conversational interfaces for crowdsourcing microtasks. While recent works have explored the effect of these interfaces in terms worker engagement and quality of work [29], there is a lack of understanding regarding the effects of conversational interfaces in building trust for the users interacting with it. To address this research gap, we explore the following research question: *To what extent can a conversational interface build trust in the context of decision support systems in comparison with a typical graphical user interface?*

To address this research question, we conduct a study by asking crowd workers to use a decision support system which suggests housing options in an European city. The motivation behind choosing this context for the decision support system is familiarity of the authors with the problem as well as the current housing problems faced by new students coming to this city for study due to the housing shortage present in many countries.[8]. Our research in

this area is guided by the following hypothesis we postulate: decision support systems aided by conversational interfaces are better at building trust than typical web-based graphical user interfaces. To validate this hypothesis, we followed a two-step approach: 1) We created a curated dataset representing real-world houses, and generated realistic house-hunting scenarios. 2) We. then presented a house-hunting scenario to crowd workers with either a conversational interface or a typical web-based graphical user interface, where we expect crowd workers to submit the correct house for the scenario given to them.

We carried out crowdsourcing experiments on Prolific. We found that **online users tended to trust conversational interfaces more in comparison with typical web-based graphical user interfaces**, while interacting with an online decision support system for house recommendation. Further, we envision that our approach and results can be generalized to other domains where a decision support system is needed, like assistance in selecting the right university for education, or determining the appropriate selling price of a used car. Overall, this work provides insights for building trust needed for decision support systems of the future.

2 RELATED WORK

We look at related literature in four different viewpoints, namely Approaches to Trust, Effect of Interfaces in building Human Trust, Crowdsourcing using Conversational Interfaces, and HCI in Decision Support Systems.

2.1 Approaches to Trust in Human Computer Interaction

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 [20]. This context has lead to a number of definitions of trust. In Rotter [32], authors define trust as "expectancy held by an individual that the word, promise or written communication of another can be relied upon". Johns [17] defines trust as "willingness to rely on an exchange partner in whom one has confidence". Mayer et al. [24] 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 [14] 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. [5] 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 [20] 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. In this study, we implemented a decision support system to help online users in finding a house, and studied whether the system interfacing a conversational interface can better build user trust.

2.2 Effect of Interfaces in building Human Trust

The effect of interface design to make the human interaction more engaging has been widely researched. In a previous study [26], 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 [27] 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 [20] 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. [36], the authors propose ways to repair trust and mitigation strategies for human-robotics interaction systems. Antrobus et al. [2] 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 [33], it is found that interfaces such as conversational interface which mimic human traits can help in increasing people's trust. Similarly, Weitz et al. [37] 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.

2.3 Crowdsourcing using Conversational Interfaces

Recent works have explored the various aspects of conversational interfaces for crowdsourcing [7, 18, 19]. Huang et al. [16] propose 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. Mavridis et al. [23] explore the effectiveness of conversational interfaces for crowdsourcing microtasks and find increased worker satisfaction without any increase in task execution time or work quality 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 [28, 29]. Furthermore, Hettiachchi et al. [13] 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.

2.4 HCI in Decision Support Systems

Decision support systems are interactive systems that aid human beings 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 [34]. The wisdom has been employed in decision support systems to improve their knowledge base. Hosio et al. [15] use crowdsourcing

tasks to populate the knowledge bases in an easy and cost effective manner. Wen [38] study the effect of a conversational interface based decision support system for stock investment activities. Yuan et al. [39] explore the requirements of a decision support system in a clinical setting. The authors concluded that user interface design and implementation were key factors for the successful deployment of CDSSes (Clinical Decision Support Systems).

3 METHODOLOGY

We created crowdsourcing experiments to represent scenarios in which a decision support system suggests housing options. In this section, we elaborate upon the crowdsourcing task design, the decision support system interfacing the conversational interface (Chat) and the typical web-based graphical user interface (Web), the dataset and scenarios, and the measures used in this study.

3.1 Crowdsourcing Task Design

In the tasks, the participants were provided with a house searching scenario in a situated experiment fashion. The scenario represents a student looking for a house in the European city of Delft, Netherlands with a certain given set of preferences. The participants were expected to interact with the provided system and enter the preferences correctly associated with the scenario. For each of the scenarios, there was only one correct house in the dataset that fit all the provided preferences. The participants were assigned either a conversational agent, 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 kept identical across both the interfaces. Figure 1 illustrates a general overview of the interaction between the participant and the interfaces. The specific details of design and structure of the two interfaces are provided in the respective subsections below.

3.2 A Decision Support System for House Recommendation

A house recommendation system acted as a decision support system in our study. The **accuracy** of the decision support system is configurable, which can be either accurate (high accuracy) or inaccurate (low accuracy). For **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 house is selected from the list of all available houses.

The decision support system in this study was presented to the participants using either a typical 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, the participant is directed to a screen displaying the scenario, and an attention check question that asks for the name of the persona described in the scenario (w1). Only if the worker submits the correct name, they are directed to a page to fill out the constraints (w2) given in the scenario. After submitting the constraints (w3), the participant is shown the house recommended by the DSS. They can either choose to submit the house recommended by the DSS (w4), or check the list of available houses (w5). If the 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 DSS recommended house. After the user clicks on submit the house, they are asked to confirm their house selection (w6) or reset filters. If the participant chooses to reset filters, the constraints they had 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 is built using React. All actions in the interface are logged using Node.js and Express, and are sent to 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 which elicits the participants to provide 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 the participant is provided a list of preferences to enter. The conversational agent initiates a conversation by greeting the participant and asking for the name assigned to them in the scenario (c2). This first prompt also serves as attention check for the participant in the sense that the conversational agent does not proceed until the correct name associated with the scenario is entered. It then proceeds to have an open ended conversation with the participant where the floor of the conversation can be taken either by the participant or the agent. Also, the participant is free to input either free text or choose one of the suggestions buttons presented. The conversation proceeds until the participant conveys that he does any have more preferences to convey and is presented with a suggested housing option (c3). At this stage, the participant has the option to either submit the suggested housing option or look at all the houses in the system and select one of them (c4). The user also has the option to reset all the constraints in case he thinks that he has made a mistake . Once the participant is satisfied with the housing option and submits the house, the continue button is activated to allow the participant to move towards the next step of the workflow (c5).

The conversation agent follows a frame based architecture [12] which is built on top of Dialogflow Messenger [1]. The backend of the agent is built on Node.js web app deployed on a Heroku server (https://www.heroku.com/). 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

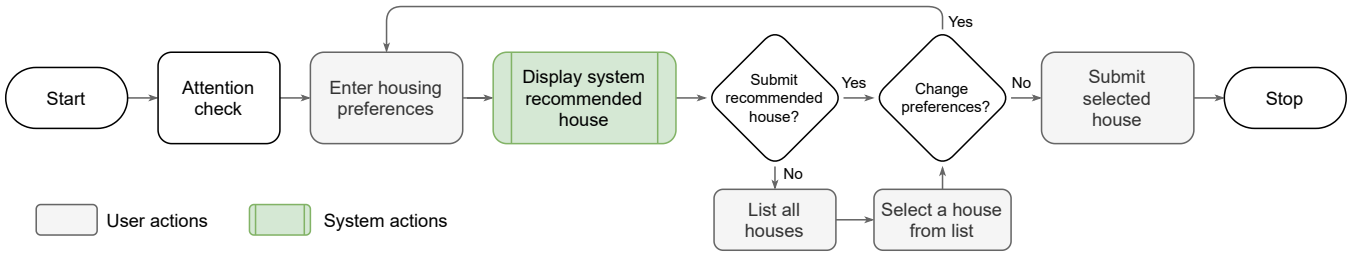

Figure 1: Overview of interaction between the participant and the house search interfaces.

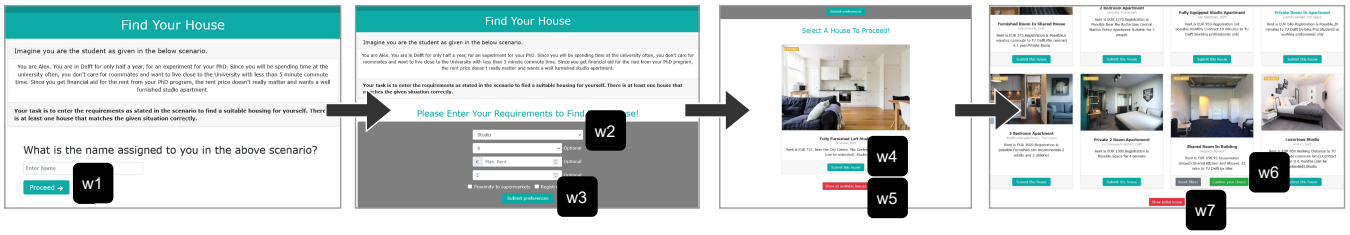

Figure 2: The Web-based graphical user interface task and its workflow.

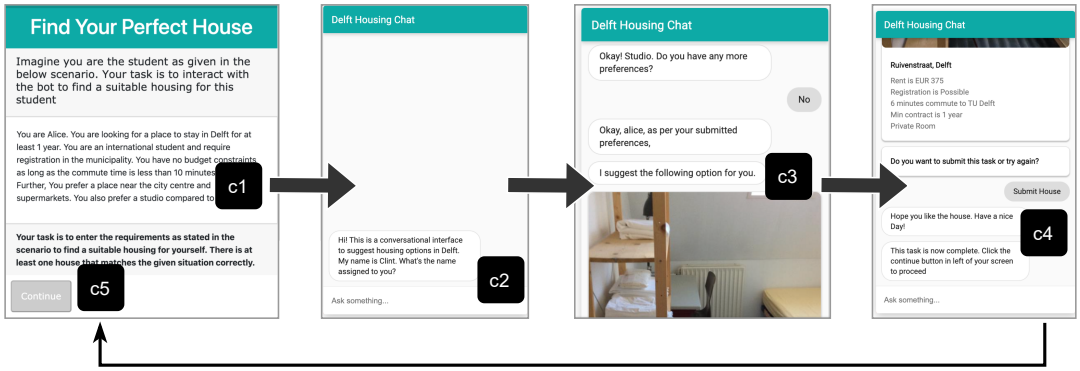

Figure 3: The conversational interface and its workflow.

the front end. The 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) **House type**: The house 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 place.
(4) **Proximity to the supermarket**: If a house is close to the supermarket or not.
(5) **Registration**: If resident at the house can be registered at the municipality or not.

Simultaneously, we created six scenarios representing students looking for housing options with particular 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 & scenarios and logging user interactions in the tasks.

3.4 Measures

3.4.1 Measuring Affinity for Technology. Attig et al. [3] 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 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

Table 1: Example of an easy and a hard scenario given to the user in each task. The 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.

based on 6-point Likert scales ranging from *Completely Disagree* to *Completely Agree* [9]. The questionnaire is presented to the participant prior to the house search task and is tailored to the interface they are expected to interact with.

3.4.2 User Behavior. We measure the user behavior in three aspects: *the correctness of the submission*, *the time spent during the task*, *and whether all the available houses are 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 (could be either low or high) can affect 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 a user 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 [31] 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 *Completely Agree* to *Completely Disagree*. The responses are assigned scores from 1 to 5 with *Completely Disagree* being 1, and *Completely Agree* being 5. 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 [31] 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 are the user interface (conversational interface vs web-based graphical user interface) and the accuracy of recommendation (high accuracy vs low accuracy), resulting in four experimental conditions:

1) Web-Low represents the condition that participants are asked to use the web-based graphical user interface to find a suitable house with a recommender system providing low-accuracy suggestions.

2) Web-High represents the condition that participants are asked to use the web-based graphical user interface to find a suitable house with a recommender system providing high-accuracy suggestions.

3) Chat-Low represents the condition that participants are asked to find a suitable house through a conversation with the conversational interface featuring a recommender system providing low-accuracy suggestions.

4) Chat-High represents the condition that participants are asked to find a suitable house through a conversation with the conversational interface featuring a recommender system providing high-accuracy 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.

4.2 Procedure

Participants for the study were recruited from the crowdsourcing platform Prolific. The crowd workers were invited to participate in a study called "*Test a House 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 participants with a minimum approval rate of 90% were selected, and participants taking part in a particular experimental condition were excluded from any subsequent experiments performed. The participants of the study were provided

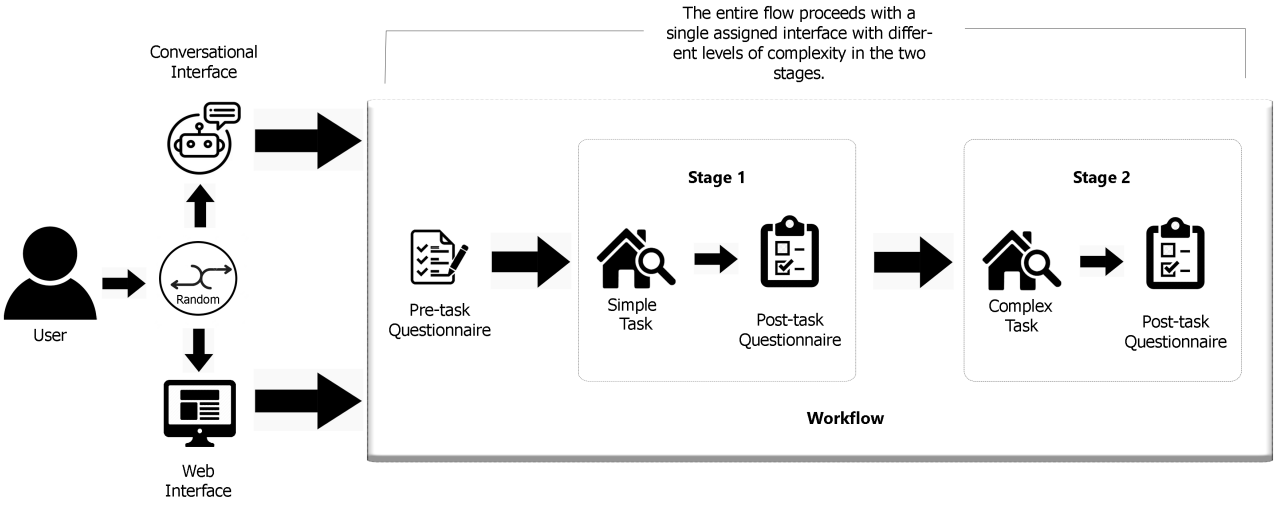

Figure 4: An overview of the study workflow.

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 questionnaire 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 house search scenario using either the chat interface or the web-based graphical user interface. They are then directed to the post-task questionnaire based on ResQue, regarding the recommendation provided by the system. On completing the questionnaire, the participants are then redirected to a transition page, from where they could continue to the second stage of the study.

The second stage consists of a second house search task with a different scenario 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. 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. On completion, the participants are provided a completion code which they were asked to enter on Prolific to get paid.

4.3 Workers and Rewards

We recruited 60 online workers for each condition (30 workers first complete easy task, followed by hard task and 30 workers first complete hard task, followed by easy task). Thus, $60 \times 4 = 240$ unique crowd workers from Prolific participated in our experiment. To further ensure the output quality, we make our studies available to crowd workers whose approval rates are higher than 90%. Participants in our study receive 1.25 GBP upon the acceptance of their submissions. According to the report from Prolific, the actual average hourly reward of our study was 7.0 GPB (9.6 USD/hour).

5 RESULTS AND ANALYSIS

Some participants had completed only one scenario and some had submitted the same task multiple times by interacting with the conversational agent again after submitting the house. Filtering was carried out to remove participants with incomplete submissions. 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). The behavioral trends for each interface and the responses for the post-task questionnaire are analysed below.

5.1 Conversational Interface Behaviour Analysis

The analysis conducted from these submissions is shown in Table 2. From the submission 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 is also seen that user performance was mostly similar for simple and hard 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 hard scenarios and the condition 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 Web Based Graphical User Interface Behaviour Analysis

The analysis conducted from these submissions is shown in Table 3. From the submisssions, it is seen 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 hard conditions, we observed that the split between correct and incorrect submissions for the hard scenarios is almost half, while the percentages are in favour of correct submissions for the easy condition (64.957%). Counter-intuitively, the time taken for hard 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 hard scenarios than in the case of easy scenarios.

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%

5.3 Analysis of Trust across Interfaces

The trust scores of the interfaces were obtained by computing the mean scores of the post-task questionnaire provided by the participants. In Table 4, we see the descriptive statistics for the Trust scores for the respective interfaces, moderated by the accuracy of the scenarios. 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 ($M \pm SD$)
Conversational 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
Graphical User 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 was performed to analyse the effect of interface type and accuracy of scenarios on the trust score. The results (Table 5) show 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 than the web interface.

Table 5: Results of a two-way ANOVA on the 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

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

5.4 Analysis of 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 by 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 respective interfaces, moderated by the accuracy of the scenarios. 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.

A two-way ANOVA was performed to analyse the effect of interface type and accuracy of scenarios 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) show 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 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 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 recommendation.

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

6 DISCUSSION

It is clear that users tended to trust the decision support system based on conversational interfaces more than web based graphical user interfaces. This result was found to be independent of the accuracy of the interfaces under consideration. Further, for the conversational interfaces, a sizeable difference between trust scores was seen between the system configured with low-accuracy condition and system with high-accuracy condition. Whereas, for the web-based graphical user interface, the trust scores did not show any significant difference.

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, while the web based graphical user interface did not show any significant differences between low accuracy and high accuracy conditions.

6.1 Trust vs Performance - Effect of Time

Interestingly, it is also seen from Table 2 and Table 3, the overall time of completion for conversational interfaces is significantly lower than web based graphical user interfaces. 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 suggestion is given. Whereas for the the web interfaces, due to distrust in the system, the users tended to be more careful in rechecking 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. It is seen 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 [11, 25], most previous studies mainly focused on the output quality and the time [4, 6, 28, 29]. Our work suggests a three-way trade-off between trust in interface, active completion time, and user performance.

6.2 Implications for Designing DSS

The results suggests that for decision support systems of the future, the choice of interface can play a major impact in the development of both trust. We found that users working with conversational interfaces in general trusted the decision support system more, compared with users working with typical web-based graphical user interfaces.

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 graphical user interface. The results on trust and satisfaction across both the decision support systems with low accuracy and high accuracy conditions conveys that, mistakes such as those in configuration and user experience in case of conversational interfaces are more detrimental in developing trust and user satisfaction, than a web-based graphical user interfaces.

6.3 Limitations and Future Work

While our results are evaluated in the domain of housing suggestions, it could be beneficial in any domain where the system designers have the choice between graphical user interface and conversational interfaces and the development of trust is one of the goals of the system. We believe that domains where the input parameters have a fixed set of options to choose from will be amenable to our results. This may range from the control interfaces in cars to robotics as well as in e-commerce domain. Further, it would be interesting to see the how the trust in conversational interfaces evolves in long term if the interface is configured with low accuracy earlier and high accuracy later. With a similar approach, in the future, we can

also compare the trust formation for purely voice user interfaces and purely chat user interfaces.

7 CONCLUSION

In this work, we investigated the effect of conversational interfaces in building trust for the decision support systems. We designed novel conversational interfaces and used typical web based graphical user interfaces for a decision support system (of house recommendation). We recruited 240 online participants and performed crowdsourcing experiments on Prolific. It was found that the mean trust scores are significantly higher for conversational interface tasks compared to graphical user interface tasks. By comparing a conversational interface with a graphical user interface for building trust in the context of decision support systems, our study highlights the impact of conversational interfaces in human computer interaction for trust development. This study has valuable insights for system designers to build resilient and trust worthy decision support systems of the future.

ACKNOWLEDGMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (no. e-infra190082).

REFERENCES

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

[2] 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.

[3] 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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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

[10] 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/.

[11] 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).

[12] 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.

[13] 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

[14] 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.

[15] 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

[16] 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

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

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

[19] 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.

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

[21] 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.

[22] 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

[23] 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

[24] 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

[25] 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.

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

[27] 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

[28] 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.

[29] 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

[30] 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.

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

[32] 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

[33] 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

[34] 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

[35] 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.

[36] 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

[37] 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

[38] 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.

[39] 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

A ANALYSIS OF AFFINITY TOWARDS THE INTERFACES

From the 222 valid submissions for the conversational interface, we found that the mean ATI score was 3.857 ± 0.75, which represents a 'Medium' affinity towards interacting with the conversational interface. Among the workers, 27.43% and 9.73% showed a 'High' and 'Very High' affinity respectively, and 12.39% showed a 'Low' affinity, while the remaining showed a 'Medium' affinity.

From the 234 valid submissions for the web-based graphical user interface, we found that the mean ATI score was 3.908 ± 0.698, which again shows a 'Medium' affinity towards interacting with a web-based graphical user interface. Among the workers, 44.35% and 3.48% showed a 'High' and 'Very High' affinity respectively, and 8.7% showed a 'Low' affinity, while the remaining showed a 'Medium' affinity.