Second chance for a first impression? Trust development in intelligent system interaction

Tolmeijer, Suzanne; Gadiraju, Ujwal; Ghantasala, Ramya; Gupta, Akshit; Bernstein, Abraham

DOI
10.1145/3450613.3456817

Publication date
2021

Document Version
Final published version

Published in
UMAP 2021 - Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization

Second Chance for a First Impression? Trust Development in Intelligent System Interaction

Suzanne Tolmeijer
tolmeijer@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

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

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

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

Abraham Bernstein
bernstein@ifi.uzh.ch
University of Zürich
Zürich, Switzerland

ABSTRACT

There is a growing use of intelligent systems to support human decision-making across several domains. Trust in intelligent systems, however, is pivotal in shaping their widespread adoption. Little is currently understood about how trust in an intelligent system evolves over time and how it is mediated by the accuracy of the system. We aim to address this knowledge gap by exploring trust formation over time and its relation to system accuracy. To that end, we built an intelligent house recommendation system and carried out a longitudinal study consisting of 201 participants across 3 sessions in a week. In each session, participants were tasked with finding housing that fit a given set of constraints using a conventional web interface that reflected a typical housing search website. Participants could choose to use an intelligent decision support system to help them find the right house. Depending on the group, participants received a variation of accurate or inaccurate advice from the intelligent system throughout each session. We measured trust using a *trust in automation* scale at the end of each session.

We found evidence suggesting that trust development is a slow process that evolves over multiple sessions, and that first impressions of the intelligent system are highly influential. Our results echo earlier research on trust formation in single session interactions, corroborating that reliability, validity, predictability, and dependability all influence trust formation. We also found that the age of the participants and their affinity with technology had an effect on their trust in the intelligent system. Our findings highlight the importance of first impressions and improvement of system accuracy for trust development. Hence, our study is an important first step in understanding trust development, breakdown of trust, and trust repair over multiple system interactions, informing improved system design.

UMAP '21, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8366-0/21/06…$15.00
https://doi.org/10.1145/3450613.3456817

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**; **HCI theory, concepts and models**.

KEYWORDS

Human-AI Interaction, Trust development, Intelligent System, Decision Support, Trust Repair

ACM Reference Format:
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 (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3450613.3456817

1 INTRODUCTION

Technological advances in storage and computation have led to the unprecedented rise in the use of artificial intelligence (AI) and automation. This has resulted in the widespread adoption of intelligent systems across several domains including healthcare, transport, manufacturing, finance, and education [23]. Many everyday tasks are supported by AI systems today. From data-fueled cloud services on computers to smart apps on mobile phones, intelligent decision support is becoming increasingly ubiquitous. Although such support can make life easier for users, inappropriate reliance can also lead to failures [16]. Consider the example of a navigation support system. On the one hand, misuse or absolute reliance on the system can lead a user to follow an outdated speed limit. Disuse or lack of reliance on the system on the other hand, can lead to missed benefits, such as a user getting stuck in traffic due to the lack of trust in a suggested detour. Considering that AI systems are now penetrating critical domains [21], one can expect far graver consequences of user trust or the lack of it in such systems.

With AI playing a prominent role in our lives, important questions surrounding our trust in AI systems have emerged. How exactly does trust evolve in the interaction between humans and AI systems? To what extent is the trust that is established through interaction robust to system accuracy over time? What factors mediate trust formation? Since trust in intelligent systems is fundamental to their widespread adoption, these are pivotal knowledge gaps to address in the emerging field of *Human-AI Interaction.*

We adopt the following definition of trust as: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [16]. Trust is relevant when a situation contains a truster and trustee. The trustee has a task to perform with an incentive to perform it and the truster has the uncertainty and risk of failing the task [10]. Along with dispositional factors such as age and situational factors such as subject expertise, trust is learned over time [11].

To our knowledge, there has been little research on dynamically learned trust that evolves over different interactions with a system [12]. In particular, the influence of accuracy and reliability on trust formation over time have been insufficiently explored. However, learning about trust development is vital for successful system usage over time. Additionally, while dispositional factors such as age and affinity with technology have been shown to influence trust [11, 26], little is understood about their interaction with system accuracy. Thus, we pose the following research questions:

RQ1 *Does the accuracy of advice of an intelligent system over multiple sessions influence the reliance of users on that advice?*
RQ2 *Does inconsistency of accurate advice from an intelligent system over multiple sessions influence trust formation?*
RQ3 *Can inaccurate advice from an intelligent system harm trust formation and accurate advice recover trust formation over multiple sessions?*
RQ4 *Do dispositional factors such as age and propensity to trust influence trust formation in an intelligent system across multiple sessions?*

To investigate these questions, we considered a domain relevant to our everyday lives, and built an intelligent housing recommendation system to carry out a multi-session study consisting of 201 participants across 3 sessions in a week. In each session, participants were tasked with finding houses that fit a given set of constraints using a housing search website that we created (as shown in Figure 1). Participants could choose to use an intelligent decision support system to help them find the right house. The tasks were designed to make manual search relatively taxing, encouraging participants to use the intelligent system. We offered a return bonus to increase the chance of participants returning for all sessions as well as a task bonus to incentivize finding the correct task solution. Depending on the group, participants received a variation of accurate or inaccurate advice from the intelligent system throughout each of the sessions. We measured trust in the system at the end of each session using the established 'trust in automation scale' [15].

Original Contributions. In this work, we present experimental evidence which suggests that first impressions matter for trust formation in Human-AI interaction over multiple sessions. However, trust can be recovered and even improved significantly when a faulty first session is followed by consistent and accurate user support by an intelligent system. Trust formation shows slow upward and downward trends, confirming that trust develops over time and is influenced by system predictability and reliability. Finally, we find that the age of the user and their affinity with technology correlate with trust development. Our findings inform system designers of the importance of first impressions and (appearance of) system improvement over time during multiple interactions. Our contributions through this work inform future research directions pertaining to trust formation, loss of trust, and trust repair. We publicly share all our data, to promote open science.[1]

2 RELATED WORK

We discuss related literature in three realms: (i) how trust has been modeled and studied in HCI, (ii) trust formation in user interactions with intelligent systems, and (iii) the relationship between user trust and system accuracy.

2.1 Trust in Human-Computer Interaction

The interest of the HCI community in trust is apparent in recent literature. From trust in automation [16] and intelligent systems [7, 12] to trust in AI, machine learning, and robotics [27], prior works have explored trust in various systems over the years. Hoff and Bashir [11] have integrated research on trust factors into an overall model. According to them, trust in automation has three main components: dispositional trust, situational trust, and learned trust. Our focus lies on learned trust, which consists of initial learned trust (including expectations) and dynamic learned trust (which changes during the interactions with the system). Hoff and Bashir identify a research gap on how previous positive and negative experiences and resulting expectations influence trust in future interactions, which is the focus of our study.

Schaeffer et al. [26] also focus on factors impacting trust formation in automation in their meta-analysis. The four main influence categories of their model include 1) traits such as age, 2) emotive factors such as attitudes towards the system, 3) states including stress, and 4) cognitive factors such as expectancy. Among other research gaps, they list a lack of research on age impact, reliability and errors – all of which are discussed in this study.

2.2 Trust Formation

Trust develops over time and depends on many factors. Each interaction with a system alters the trust in that system. Holliday et al. [12] looked at trust formation within one user session. They found that the impression of a learning system, conveyed through explanations, led to higher levels of trust. In addition to a system learning over time, the impression of system reliability shapes trust. Case in point, consistent reliable support leads to steadily increasing trust, while consistent unreliable support led to constant decrease in trust [3]. First impressions are especially important: negative first impressions have a stronger negative influence on trust than negative impressions acquired later on [20].

Understanding trust formation does not only involve how trust is fostered, but also when it breaks and how it can be recovered. Trust break and recovery have been understudied [5, 29]. In this study, we thereby focus on the influence of accuracy on trust formation and whether improved accuracy is enough to regain trust after inaccurate advice.

[1]Open Science Foundation (OSF): https://osf.io/ndjfs/?view_only=502f2abc34714838918213a04d68dc58

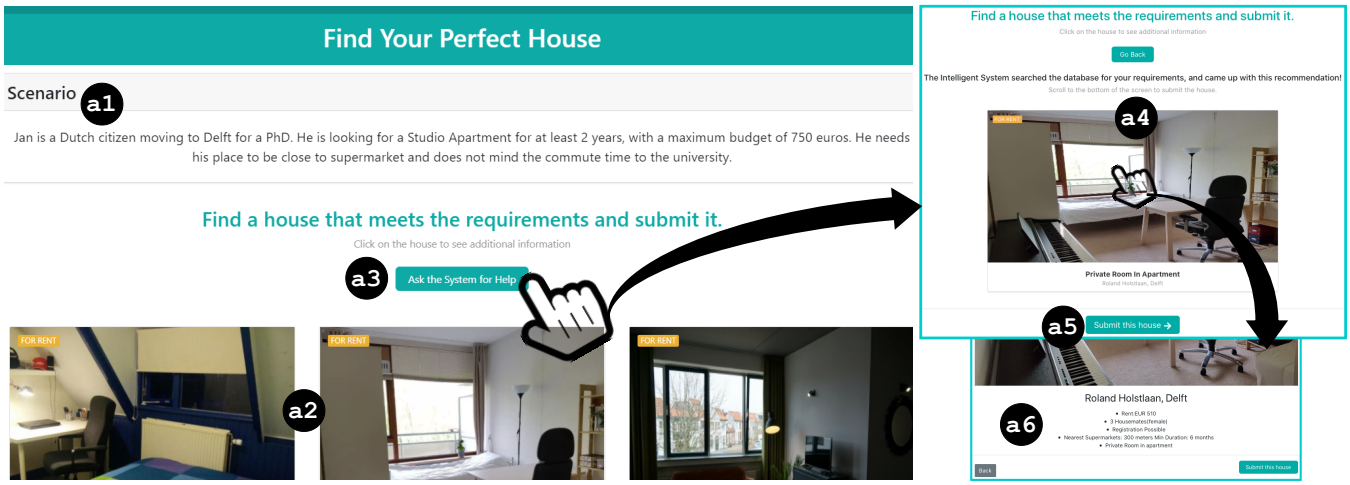


Figure 1: The housing search interface (left-hand side), and assistance from the intelligent system (right-hand side).

2.3 Trust and Accuracy

The influence of accuracy on trust has become more influential as artificial intelligent methods have become more opaque, e.g., when compared to earlier rule-based system. While results from AI have been very promising, users do not trust what they do no understand [19]. In fact, providing explanations for AI models that are less human-meaningful decreases perceived accuracy compared to actual accuracy [19]. The importance of the impression of the system is echoed in work by Yin et al [31]. Authors found a difference in trust formation between the effect of stated accuracy and observed accuracy: stated accuracy has a significant effect on trust independent of actual accuracy. Nevertheless, model accuracy is more important for trust formation than explanations [22].

If the system is indeed unreliable or inaccurate, the user takes longer to decide whether to follow the system's advice [28]. In robots, Desai et al. [6] found that early unreliability had a greater impact on trust formation than unreliability later on. Additionally, the error type also determines the impact on trust formation. For instance, in the autonomous cars domain, obstacles that were not detected but missed had a bigger impact than false alarms of obstacles [2].

A study on accuracy over time with multiple sessions was done [14], but in relation to user feedback. They found that allowing users to provide feedback lowered trust in the system and lead to a lower experienced accuracy, independent of actual system accuracy. To our knowledge however, an in-depth understanding of the interaction between accuracy and trust formation over time is missing - especially whether (in)accuracy can lead to trust loss and trust recovery.

3 STUDY DESIGN

To address the aforementioned research questions, we conducted a crowdsourced multi-session study. In this section, we describe the measures, task design, and the procedure.

3.1 Measures

Measuring User Trust in the Intelligent System. We used a validated trust scale [15] to measure user trust in each case. The scale consists of 12 items pertaining to the intelligent system, and participants are asked to use a 7 point Likert-scale ranging from (*1: Not at all*) to (*7: Extremely*), to indicate their agreement with each item. While relatively recent scales for trust measurement such as the multi-dimensional measure of trust are available [17] or domain specific trust scales such as for online recommender agents have been proposed [4], we chose to use a more generic and validated scale of trust in automated systems [15]. To account for the dispositional component of user trust formation, we additionally used the validated and widely accepted 'propensity to trust scale' [9]. Each trust scale was aggregated into an average trust score per participant ranging from 1-7 and 1-5 respectively. In the case of the trust in automation scale, scores of negatively worded trust were reverse coded.

Measuring Affinity for Technology. Recent research has shown that affinity for technology interaction can be seen as a core personal resource for successful coping with technology and a facet of user personality [1]. We used the 9-item 'affinity for technology interaction' (ATI) to assess a user's tendency to actively engage in technology interaction [8].

3.2 Task Design and the Intelligent System

Trust requires three components: actors to form trust, an incentive to trust, and a risk to trust [10]. We modeled our task to integrate these three components. In the task, participants (*the trusters*) were presented with house searching scenarios with a given set of constraints. There was only one house per scenario that fit all constraints. Finding the right house that satisfies all requirements was rewarded with a monetary bonus of 0.25 GBP (*the incentive*). Participants were offered advice by an intelligent system (*the trustee*). If they did not consider the advice, they risked *losing valuable time* by having to manually click through each of the displayed houses to

find the right one that matched all constraints (*the risk*). We considered the housing domain since many people have experience with it and items naturally need to fit multiple requirements, making the search challenging enough to benefit from automated assistance.

Figure 1 presents an overview of the interface and the intelligent system that users were equipped with. On beginning the task, a house search scenario is presented to the user at the top of the interface (cf. a1). The scenario describes the constraints pertaining to the house search, in a situated search format. We manually crafted the tasks to be taxing, to create a realistic incentive for the users to engage with the intelligent system. We considered two levels of complexity within the house search task: in the relatively **easy** scenario, users were assigned a house search task with 3 constraints, while they had to deal with 5 constraints in the **complex** scenario (as shown in Table 1).

Table 1: Examples of easy and complex scenarios presented to users in each house search task. Each distinct constraint is colored for the benefit of the reader.

Complexity	Scenario
Easy	Peter is moving to Delft as a first year Bsc. student. He is a very easy going guy and is looking for a shared room which fits his rent budget of 300€. Further, he would require registration at the municipality.
Complex	Jan is a Dutch citizen moving to Delft for a PhD. He is looking for a studio apartment for at least 2 years, with a maximum budget of 750€. He needs his place to be close to a supermarket and does not mind the commute time to the university.

Note that there was a total of 12 houses displayed on the interface in a randomized order (cf. a2), and in each task only one house satisfied the given constraints. Participants could use assistance from the intelligent system by clicking the **Ask the System for Help** button, present below the scenario description (cf. a3). On clicking the button, the intelligent system presents the user with a house, that it claims matches all the required scenario constraints (cf. a4). Users can either submit the house directly by using the **Submit this House** button (cf. a5), or verify whether the constraints corresponding to the suggested house are indeed satisfied, by clicking on it and viewing the details (cf. a6). Based on the experimental condition (described in the following section), the intelligent system either provided an accurate or inaccurate suggestion. Users could freely switch between manually sifting through each house and using the intelligent system by using the **Back** button. By clicking on a house, users could view its details.

System Implementation. We created a web application using React.js for the front-end of the house search interface and Node.js as well as Express.js for the backend. MongoDB was used both for logging user interactions in the task and for storing data pertaining to the houses. The application was hosted on Heroku. In total, we created six distinct scenarios: three **easy** and three **complex**. Each participant was then randomly assigned one easy scenario and one complex scenario in each of the three sessions. The scenarios were also randomized across the sessions and between groups for participants, to prevent biases due to ordering effects. The total number of houses in the data set was 12 and for each scenario the position of the correct house in the displayed list was randomized to prevent biases due to ordering or learning effects. A fixed list of incorrect houses was created to support the sessions with incorrect advice from the intelligent system. In such sessions, a random house was selected and displayed from this list as a suggestion from the system. In case of session with correct advice from the intelligent system, the correct house was shown.

Although there are more elaborate systems for housing search (e.g., [25]), we opted for a simpler interface that allowed us to isolate the effect of the system advice on user trust.

3.3 Procedure and Experimental Setup

We recruited participants from crowdsourcing platform *Prolific*.[2] The platform has been shown to be an effective and reliable choice for running relatively complex and time-consuming interactive information retrieval studies [24, 30]. Crowdworkers on Prolific were invited to participate in a multi-session study titled, "*Finding the right house that meets your requirements*". To ensure reliable participation, we followed Prolific's guidelines and restricted eligibility to workers who had an acceptance rate of at least 80% and had at least 10 successful submissions on the platform. Participants were informed about the longitudinal nature of the task. Those who accepted our task received brief instructions about the task and were asked to check-off an informed consent before beginning their task session. As shown in Figure 2, participants were first asked to complete a pre-task questionnaire consisting of (i) demographic questions about their *age*, *gender*, and *country of origin*, as well as (ii) *the affinity to technology interaction* (ATI) scale. Next, participants were assigned two consecutive house searching tasks. They were incentivized with a monetary bonus of 0.25 GBP for finding the right houses. On completing the two house search tasks, participants were asked to complete an exit questionnaire consisting of (i) the 'trust scale', (ii) the 'propensity to trust' scale, and (iii) a text area to provide optional remarks or comments. On completing the exit questionnaire, participants received a completion code which they were asked to enter on Prolific to receive their base payment. We paid all participants Prolific's suggested fair wage of 7.50 GBP/h.

After a successful session, participants were invited to join a second session two days later and a third session two days after the second session. To maximize the return rate of participants, we rewarded participants with a return bonus of 1 GBP in addition to their base pay for completing Session 2 and 3. Since we logged participant data using their Prolific IDs, we ensured that in each session participants received two distinct house search scenarios (one **easy**, one **complex**), which they did not encounter previously. While the study flow was identical in the three sessions, participants were only asked to respond to the demographic questions, and fill out the 'propensity to trust' scale in the first session.

Participants were randomly assigned to one of eight different experimental conditions (referred to as 'groups' hereafter),

[2] https://www.prolific.co

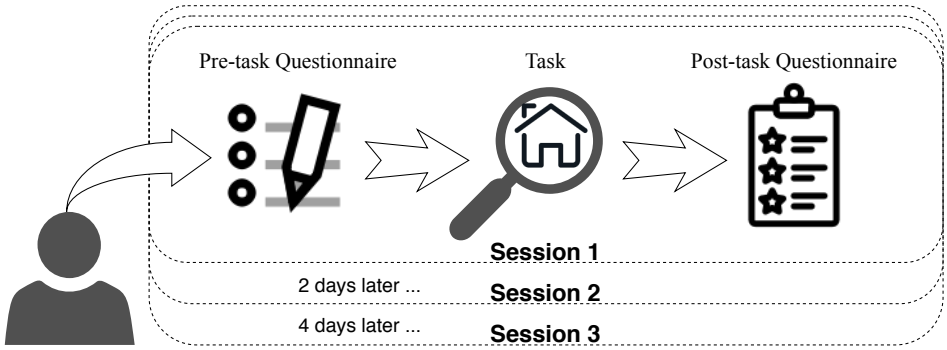


Figure 2: Overview of the study workflow.

that differed in the sequence of accuracy of the intelligent system across the three sessions. Assuming that **1** represents accurate advice and **0** represents inaccurate advice given by the system in a given session, the experimental conditions were as follows: '**1 1 1**', '**1 1 0**', '**1 0 1**','**0 1 1**','**0 0 1**', '**0 1 0**','**1 0 0**', and '**0 0 0**'. For instance, this means that participants assigned to group '**0 1 0**' received incorrect advice in session one, correct advice in session two, and incorrect advice in session three.

3.4 Hypotheses

The aforementioned experimental conditions (or groups) allow us to test different hypotheses by comparing sessions and groups. Specifically, the hypotheses we test to answer our research questions can be found in Table 2.

4 RESULTS

In our first session, 255 subjects participated. Of those participants, 83% returned for the second session two days later. 96% of these participants returned to complete the third session two additional days later. This resulted in a total of 203 participants, who completed three sessions. Two participants were excluded based on clearly evident unreliable participation. Thus, the results and analysis presented hereafter pertain to these 201 participants unless specified otherwise (see Table 3). A compromise power analysis of the mixed ANOVA revealed that with over 24 participants per group, we have a power of 0.9 (considering a medium effect size of $f = 0.25$, $\alpha = 0.05$).

We found that 26 participants did not use the intelligent system in any of the three sessions. Trust scores for these participants were therefore excluded in our analyses pertaining to user trust. Since questions in the trust scale refer to system performance, the responses from users who never utilized the system are meaningless.

Each session consisted of two scenarios. We will refer to sessions using S1 to S3, scenarios will be denoted as S1.1 and S1.2 for each session. To control for Type-I error inflation in our multiple comparisons, we use the Holm-Bonferroni correction for family-wise error rate (FWER) [13], at the significance level of $\alpha < .05$. Significance levels are marked as follows: $p < 0.05^{*}$, $p < 0.01^{**}$, and $p < 0.001^{***}$.

4.1 Participant Demographics

One hundred thirty-five participants reported to be female (66 male). The age of the participants ranged from 18 to 62 years old (*M=27.5, SD=9.2*). Education ranged from high school or less (29%) and college without degree (25%) to some form of degree obtained throughout their life (46%). 40% of the participants reported to have studied computer science or some related field. Participants originated from 30 different countries, with most participants reportedly born in the United Kingdom (41), Poland (36), Portugal (26), and Italy (22).

4.2 Success of Participants Across Sessions

Independent of the experimental group assignment, on average participants were able to successfully find the right house in 78% of the scenarios in the first session, 66% in the second session, and 92% in the third. Part of the difference in user accuracy in finding the right house can be explained by the difficulty of the scenario: in four out of six scenarios, there is a significant difference, using Fisher's exact test, between **`easy`** and **`complex`** scenarios and correct/incorrect answers given by the participants. Complexity does not explain user accuracy in the second scenario of session 1 and 3. Another explaining variable to user accuracy is correctness of the system's advice. Except for the second scenario of session 1, there is a significant difference between user and system (in)accuracy: users made less mistakes when the system gave correct advice. The summarized results of the two Fisher's exact tests can be found in Table 4.

4.3 System Accuracy influences Reliance

We analyzed the reliance of users on the intelligent system. Indicators of user reliance on the system can be distinguished at two levels: users clicking the **Ask the System for Help** button to open the system's suggestion (*open*) or users submitting the system's suggestion by clicking the **Submit this House** button as their answer (*submit*).

Results pertaining to the reliance of users on the intelligent system can be found in Table 5. The group average for *opening* the system was above 50% for all groups except group '**000**'. In this case, usage drops gradually to below 33% in the last session — the only session across all groups where this is observed. Interestingly, system usage within a session sometimes dropped despite the system providing correct suggestions or increased despite inaccurate

Table 2: Hypotheses and their required comparisons. Comparisons are made either between sessions within a single group, or between different groups. Investigation of dispositional factors is not related to specific sessions or groups.

RQ	Hypothesis	Comparison	Groups
1	H1. Increased amounts of accurate advice leads to more user reliance, while inaccurate advice will lower intelligent system dependence.	Groups	All groups
2	H2. Consistent inaccurate advice over multiple sessions leads to significantly lower trust than inconsistent accuracy over time.	Groups	'000' vs. '100', '010', and '001'
2	H3. Trust is significantly higher for users that receive consistent accurate advice.	Groups	'011', '101', and '110' vs. '111'
2	H4. Inaccurate advice is more harmful in earlier sessions rather than later sessions.	Sessions	'001' vs. '010' vs. '100' '011' vs. '101' vs. '110'
3	H5. Trust is lost significantly if an inaccurate session follows an accurate session.	Sessions	'110', '010', '101', and '100'
3	H6. Trust does not recover significantly when consistent accurate advice follows an inaccurate first impression.	Sessions	'011'
4	H7. The dispositional factors of gender, age, culture, experience with computer science, propensity to trust, and affinity with technology all influence trust formation across multiple sessions.	Sessions and Groups	All groups

Table 3: Number of participants per experimental condition

Group	000	001	010	011	100	101	110	111
Participants	24	24	26	26	26	24	25	26

advice. However, the highest average usage of the system was observed for group '**111**' and lowest for group '**000**'. The order of average usage ratios suggest that first impressions matter, which is further discussed in the next subsection. Nearly all first usage in a session stayed equal or went up if the previous session had correct advice, or stayed equal/went down if the previous session contained incorrect advice from the system. Exceptions to this trend are the first session transition within group '**101**' and the two session transitions within group '**111**', possibly because system usage was already relatively high to begin with. This partially supports H1: while in general, more accurate advice leads to consulting the intelligent system more often, there are exceptions (such as group **'011'** having less system usage than group **'100'**).

The fraction of users who submitted the system's suggested house after seeing it in each session, and across all groups, is reported in Table 6. Of the users who opened the system during a session, the fraction that *submitted* the suggested house ranged from 12% in group '**000**' to 66% in group '**111**' on average. The average submission reliance ratio for all incorrect sessions was 0.15, while it was 0.62 over all correct sessions. Again, we see that first impressions, i.e., correct advise in session 1, influenced reliance on the system, with respect to following the suggestion. This supports H1: once the system is consulted, advice usage increases with system accuracy.

4.4 First Impressions Matter

Participants were assigned to different orders of session support accuracy. To answer RQ2, trust scores were compared using a mixed ANOVA between groups within sessions.

We found a statistically significant effect of experimental group assignment on trust scores ($F(7, 146) = 18.456, p < .0001, \eta_p^2 = 0.470$). While session alone did not influence user trust ($F(2, 292) = 0.127, p = 0.881, \eta_p^2 = 0.001$), there was a strong interaction between the accuracy of a group and session in explaining the reported average trust scores ($F(14, 292) = 19.910, p < .0001, \eta_p^2 = 0.488$). Pairwise comparisons between groups are illustrated in Table 7, while the pairwise comparisons of sessions within groups are presented in Table 8. Tukey's HSD test was used for post-hoc analysis (reported in Table 7 and 8).

Given that Table 7 shows comparisons for aggregated trust scores over all three sessions, we expect groups with equal number of correct suggestions to receive equal average trust scores. If not, order and consistency of accurate suggestions would appear to matter. In some cases, we found that order does not matter. For example, there is no significant difference between groups '**110**', '**101**' and '**011**' ($p = 0.797$, $p = 0.155$, and $p = 0.114$ respectively). However, group **'111'** scored significantly higher than any of the groups with two accurate sessions, supporting H3.

In other cases, the importance of reliability and validity does influence trust averages, leading to significant differences between groups with equal correct suggestions. This is especially the case for groups that received accurate support from the intelligent system in only one of the three sessions. Shifting accurate system behavior by one session did not lead to a significant difference, i.e., neither group '**001**' and '**010**' ($p = 0.196$) nor group '**010**' and '**100**' ($p = 0.134$) differ from each other. However, group '**100**' was found to have a significantly higher average trust score than group '**001**' ($p = 0.013$). This suggests that a first good impression is significantly better for trust development than a repair through correct advice at a later point in time, supporting H4.

Additionally, average trust scores in group '**000**' did not differ from the groups '**001**' or '**010**' ($p = 0.960$ and $p = 0.269$ respectively), but were found to significantly differ from group '**100**' ($p = 0.027$). This partially supports H2. Additionally, this once again corroborates that the first good impression can make all the difference. Missing this opportunity for trust development in a first session causes later possible trust recovery to be futile. In fact, group '**100**' and '**011**' do not differ significantly in average trust

Table 4: P-value results of two Fisher's exact tests on user accuracy. Difficulty (`easy/complex`) and system accuracy (correct/incorrect) were compared against user accuracy (correct/incorrect).

Sessions	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2
Difficulty	0.006**	0.854	1.076e-5***	0.011*	0.018*	0.814
System Accuracy	0.001e-1***	0.104	5.634e-6***	1.422e-7***	0.018*	0.003**

Table 5: Ratio of participants who used the system per group by clicking the system suggestion at least once. Average usage ratio per group is shown in the last column.

Group	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	*avg.*
000	0.54	0.42	0.54	0.42	0.33	0.29	*0.42*
001	0.71	0.50	0.63	0.58	0.63	0.63	*0.61*
010	0.85	0.58	0.65	0.58	0.65	0.62	*0.65*
011	0.54	0.50	0.54	0.58	0.65	0.69	*0.58*
100	0.62	0.69	0.77	0.85	0.58	0.54	*0.67*
101	0.79	0.54	0.71	0.67	0.58	0.58	*0.65*
110	0.68	0.60	0.76	0.84	0.76	0.72	*0.73*
111	0.77	0.73	0.69	0.77	0.73	0.69	*0.73*

Table 6: Ratio of participants who submitted the system's suggestion after opening it.

Group	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	*avg.*
000	0	0.20	0.08	0.20	0.13	0.14	0.12
001	0.12	0	0	0.07	0.67	0.73	0.26
010	0.09	0.07	0.47	0.67	0.06	0.06	0.24
011	0.07	0.31	0.50	0.67	0.53	0.72	0.47
100	0.44	0.72	0.30	0.32	0.07	0.21	0.34
101	0.42	0.46	0.41	0.44	0.71	0.93	0.56
110	0.41	0.67	0.42	0.71	0.11	0.11	0.41
111	0.35	0.63	0.61	0.65	0.84	0.89	0.67

scores ($p = 0.325$), even though the latter group corresponds to more correct suggestions than the former, underlining this finding further.

4.5 Trust Recovery is Possible

A further understanding of group differences can be derived from session differences within groups. The results of session comparisons per group can be found in Table 8.

Following H5, we expect trust to be significantly lower for an inaccurate session after it follows an accurate session. This is supported: we find this trust loss for groups '010', '100', '101', and '110'. The one exception to 'first impressions matter' and the one comparison that had an unexpected significant results, was within group '**011**'. While trust increase between session one and two was expected, trust increased further between session two and three, therefore, not supporting H6. A potential explanation can be that the impression of an improving system positively influenced perceived reliability of the system, leading to increased trust in the system. One possible explanation is that the impression of a learning system leads users to accept an initial fault when the system improves [12].

4.6 Dispositional Factors have Little Influence

Although system interactions influence trust development greatly, certain dispositional factors also shape trust evolution. These factors include for example age, gender, and country of origin [11]. To investigate **RQ4**, we gathered participant information for these factors, as well as the following: level of education, whether they studied computer science, their affinity with technology [8], and their propensity to trust [9].

We used linear mixed effects models to compare the influence of different dispositional factors. The fixed effects were set to "group * session", since the mixed ANOVA results from the analysis displayed in Table 7 and Table 8 showed a very strong interaction effect between experimental groups and sessions. Models with different added random effect variables were compared using ANOVA.

We found that out of all measured dispositional elements, two factors have a significant influence on trust evolution: age of the participant ($p = 0.006$) and their affinity with technology ($p = 0.012$). However, these traits only show a small effect ($sd = 0.24$ and $sd = 0.21$ respectively). Therefore, H7 was only partially supported.

The summarized results of our tested hypotheses can be found in Table 9.

4.7 Trust Evolves Slowly

The most detailed overview of trust scores can be found in Figure 3. In addition to the results of our hypothesis testing, we want to highlight interesting trends in the observed pace of trust formation. Many session comparisons with the same provided accuracy did not show a significant difference, but did show a trend in the expected direction. An example of this is group '**000**', where there is no significant difference, but trust drops slightly over the sessions. Every group that has two inaccurate session suggestions shows a downward trend for the second incorrect session, no matter the order of accuracy. For positive trends, this is only the case when two accurate sessions are presented sequentially. Results from group '**101**' even show that trust between session one and three shows a downward trend. While these results are not significant differences, all found differences are in the expected direction.

5 DISCUSSION AND FUTURE WORK

Our study revealed important interactions between trust formation and accuracy during intelligent system usage. Our work consolidates and complements previous studies of trust in HCI, and

Table 7: Results of mixed ANOVA for average trust scores between groups. Green cells imply a significant difference between groups. The group mentioned in a green cell had a higher average trust based on Tukey's HSD test.

Group	000	001	010	011	100	101	110	111
000								
001								
010								
011	011**	011***	011**					
100	100*	100*						
101	101***	101***	101***		101*			
110	110***	110***	110***		110*			
111	111***	111***	111***	111***	111***	111**	111***	

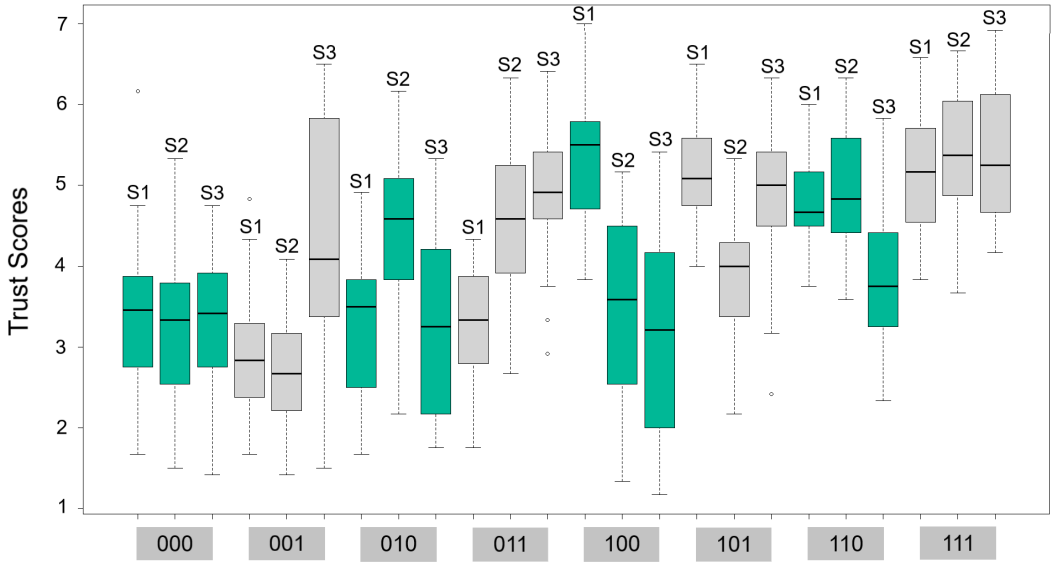


Figure 3: Boxplots representing trust scores (*x-axis*) per session, across each experimental group (y-axis).

provides further insights on trust formation and evolution over time. In this section, we discuss our key findings and present further research directions that we believe are necessary to further understand user trust formation in intelligent systems.

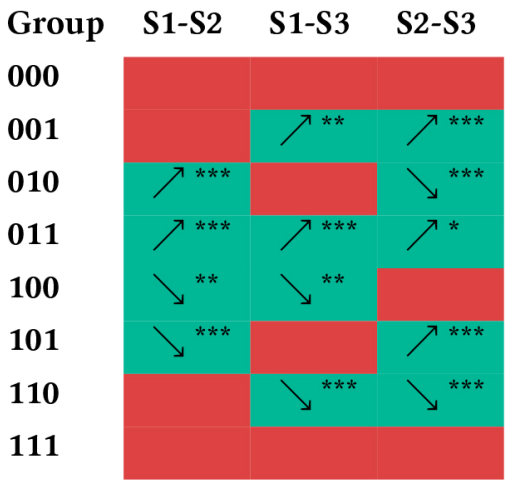
5.1 Result Discussion

User Success: The fraction of correct houses found by the participants depended on two factors: system accuracy and difficulty of the scenarios. Particularly, difficult scenarios were more likely to be answered wrong, as were scenarios where the system gave the wrong advice. The fact that scenario complexity does influence user accuracy in session 1 and 3 seems to be caused by a learning effect: once users are used to the task in the first scenario, the difference between three and five constraints had less of an influence.

System reliance: Participants seemed to especially rely less on the system when they were in group **'000'**. Longer negative experience over time influenced their usage to decrease, especially in the last session. This indicated that even opening the system was not worth their effort. Submission reliance on the other hand had a closer correlation with system accuracy. Intuitively, groups with more accurate suggestions were more likely to submit those suggestions.

First impressions: The importance of first impressions in intelligent system interactions has been reflected by recent work [6, 20]. Our findings corroborate this understanding of Human-AI interaction. However, in contrast to related work that has primarily looked at trust formation within a single session, we measured trust development across multiple system interactions. We found that first impressions are not only important within a session, but also between sessions and over time. Interestingly, this is mostly the case when the system only had one accurate session. When the system provided two accurate sessions, there was no difference in trust values in all possible session orders (**110**, **101**, and **011**), even

Table 8: Results of mixed ANOVA for average trust scores within groups between sessions. Green cells imply a significant difference between sessions. ↗ implies trust increased between the compared sessions, ↘ indicates trust decreased.



when the first session was inaccurate. More research is needed further investigate if increased system reliability, i.e., being accurate *most* of the time, indeed trumps the importance of first impressions in trust formation.

Trust recovery: We found that trust recovery is possible when a first system mistake is consistently corrected in later sessions. One possible explanation lies in the learning capabilities of the system, Earlier studies found that the impression of a learning system could lead to higher levels of trust [12], in line with our findings. As such, an interesting follow-up study can focus on features influencing perceived intelligence and how it influences trust formation.

Slow trust changes: The reported trust scores showed upward and downward trends when the system showed consistently accurate and inaccurate support respectively. These slow but steady trends are reflected in research in the domain of autonomous cars [3]. We advanced the current understanding of trust evolution by introducing three consecutive sessions for each user. By expanding the number of sessions in future work, it can become apparent when and whether these trends become significant and/or plateau to a steady trust score in the longer-term.

Influence of dispositional factors: We found that age and affinity with technology influenced user's reported trust scores to an extent. While this is in line with earlier work [11, 26], other factors did not have a significant impact, including level of education, country of origin, gender, and propensity to trust. There are various possible explanations for these results. Firstly, as intelligent systems are becoming more pervasive, people from all levels of education come in touch with intelligent systems. The lack of significant effect of country of origin can be due to our sample: most of the participants were from Europe. It is possible that inter-continent comparison results in less effect than a comparison between continents. Finally, gender and propensity to trust did not have a significant effect. One possible explanation can be that participants did (not) experience the system to be intelligent. As the system starts to show more human-like traits, mental models related to trust in humans are more likely to be activated. We measured propensity to trust *in humans*, which does not correlate with the trust in our system if it is not perceived as intelligent enough. Future research could include different levels of anthropomorphism and system intelligence, to investigate its influence on trust.

It is striking that most dispositional traits had little to no effect on reported trust scores. Potentially, dispositional traits become less important as system experience increases. Alternatively, dispositional traits could influence trusting behaviors more than trusting beliefs. More research is needed on the effect of dispositional trust factors over time.

5.2 Caveats and Limitations

We make important contributions by advancing the current understanding of trust formation in Human-AI interaction. To position our findings within the scope of our study, we discuss the caveats and limitations of this work.

Firstly, we did not distinguish how wrong the intelligent system was. Incorrect advice consisted of a randomly assigned house that did not satisfy one or more requirements. The degree of incorrectness of an intelligent system can potentially influence trust formation. For example, a system that is very clearly wrong in its advice might lose user trust earlier than a system that is just slightly off. We aim to explore this in our imminent future work. The perceived utility of the system can also vary; adding more items in the search space could relate to more time saved by using system advice, while a larger bonus may also increase system usage.

Our focus in this work was on self-reported trusting beliefs. This is a direct measure, but can be subject to a reporting bias. Behavioral analysis, for example exploring whether the participant heeded the advice, was used to corroborate our findings. However, further analysis can explore trusting behavior of users in comparison to trusting beliefs.

It is important to note that participants in our study were primarily European and fairly educated. The sample size of around 25 participants per experimental condition can limit the generalisability of our findings to other populations. Finally, as with much trust research, it can be questioned whether findings achieved in online studies can be replicated in real life. Experiments with intelligent systems being used in real life can both provide longer research windows to see if trends in trust formation over time become significant, as well as check the validity of online studies.

5.3 Implications and Future Work

Complex machine learning models and intelligent systems are currently being deployed in several critical domains, albeit as functional black-boxes. When human interaction with such systems, particularly in the first iteration, results in a sub-par experience, system adoption can be gravely affected. Impressions of a learning system can increase trust in the system, but only when the system actually appears to learn the correct behavior. Given that trust evolves slowly, system designers should focus on consistent behavior over time. Subsequently, system designers could benefit

Table 9: Results of tested hypotheses.

Hypothesis	Result
H1. Increased amounts of accurate advice leads to more user reliance, while inaccurate advice will lower dependence on the intelligent system	Partially supported
H2. Consistent inaccurate advice leads to significantly lower trust than inaccurate consistent advice.	Partially supported
H3. Trust is significantly higher for users that receive consistent accurate advice.	Supported
H4. Inaccurate advice is more harmful in earlier sessions rather than later sessions.	Supported
H5. Trust is lost significantly if an inaccurate session follows an accurate session.	Supported
H6. Trust does not recover significantly when consistent accurate advice follows an inaccurate first impression.	Not supported
H7. The dispositional factors of gender, age, culture, experience with computer science, propensity to trust, and affinity with technology all influence trust formation across multiple sessions.	Partially supported

from trying other trust recovery mechanisms, especially when the user group is younger or has less affinity with technology.

Consistent system behavior over time can be investigated over a longer period of time with more sessions, to see if our results hold when users become used to the system and have calibrated their trust according to their experience.

In our work, we focused on the self-reported trust scores of users, or trusting beliefs. This in fact is only one aspect of trust: trust can for example be formalized as a disposition, attitude, belief, intention, or as behavior [18]. For example, while trusting belief usually has more emphasis on integrity of the trustee, trusting behavior focuses more on integrity and benevolence of the trustee [18]. There have been early results that suggest a mismatch between trust beliefs and trust behavior [28], which needs further investigation. We mainly focused on lack of accuracy as a cause for trust breakdown and improved accuracy as a form of trust recovery. To prevent trust loss in case of inaccurate AI support, different strategies for trust recovery can be deployed besides improving system accuracy. More research is needed into the effectiveness of such strategies related to different kinds of errors [5, 29].

6 CONCLUSION

Appropriate trust in intelligent systems is vital for successful and correct usage. Trust is not a static concept, but evolves during interactions over time. We presented a crowdsourcing study on the influence of system accuracy on trust formation over time. Answering **RQ1**, we find that accuracy explains opening of and using intelligent system advice. Following **RQ2**, inconsistent accuracy of advice influences trust formation. Specifically, inaccurate advice leads to trust loss, earlier inaccuracy is more harmful than inaccurate advise in later sessions, and trust is significantly higher for users that receive consistent accurate advice. Session-wise comparison resulted in the answer for **RQ3**: inaccurate advice harms trust formation when it follows an accurate session and trust can be recovered after an initial inaccurate advice if the system provide accurate advice afterwards. With regards to influence of dispositional factors researched in **RQ4**, we discover that participant's age and affinity with technology have a small influence on trust formation. We identified the influence and importance of accuracy for trust formation and point to further research avenues on trust formation, trust break, and trust repair. In sum, this study provides first insights into trust development in response to system performance over multiple system interactions. Hence, it provides a first building block to understand this important and timely topic.

REFERENCES

[1] 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. In *International Conference on Human-Computer Interaction*. Springer, 19–29.

[2] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T Esterwood, X Jessie Yang, Lionel P Robert Jr, and Dawn M Tilbury. 2020. Comparing the Effects of False Alarms and Misses on Humans' Trust in (Semi) Autonomous Vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 113–115.

[3] Matthias Beggiato and Josef F Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour* 18 (2013), 47–57.

[4] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[5] Ewart J de Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.

[6] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–258.

[7] 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.

[8] 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.

[9] M Lance Frazier, Paul D Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[10] Russell Hardin. 2006. Trust. (2006).

[11] 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.

[12] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 164–168.

[13] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[14] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.

[15] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[16] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[17] Bertram F Malle and Daniel Ullman. 2020. A Multi-Dimensional Conception and Measure of Human-Robot Trust.
[18] D Harrison McKnight and Norman L Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*. Springer, 27–54.
[19] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
[20] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[21] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.
[22] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*.
[23] 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.
[24] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 229–238.
[25] Shashank Mouli Satapathy, Rutanshu Jhaveri, Ujjwal Khanna, and Ashish Kumar Dwivedi. 2020. Smart Rent Portal using Recommendation System Visualized by Augmented Reality. *Procedia Computer Science* 171 (2020), 197–206.
[26] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.
[27] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.
[28] Hiroyuki Tokushige, Takuji Narumi, Sayaka Ono, Yoshitaka Fuwamoto, Tomohiro Tanikawa, and Michitaka Hirose. 2017. Trust lengthens decision time on unexpected recommendations in human-agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 245–252.
[29] 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*. 3–12.
[30] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2020. How Does Team Composition Affect Knowledge Gain of Users in Collaborative Web Search?. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 91–100.
[31] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.