Do I Trust My Machine Teammate? An Investigation from Perception to Decision

Kun Yu[*†], Shlomo Berkovsky[*], Ronnie Taib[*], Jianlong Zhou[*†], Fang Chen[*†]
*Data61, CSIRO Eveleigh, NSW, Australia
†School of Software, University of Technology, Sydney, Ultimo, NSW, Australia
Email: {kun.yu, jianlong.zhou, fang.chen}@uts.edu.au,
{shlomo,berkovsky, ronnie.taib}@csiro.au

ABSTRACT

In the human-machine collaboration context, understanding the reason behind each human decision is critical for interpreting the performance of the human-machine team. Via an experimental study of a system with varied levels of accuracy, we describe how human trust interplays with system performance, human perception and decisions. It is revealed that humans are able to perceive the performance of automatic systems and themselves, and adjust their trust levels according to the accuracy of systems. The 70% system accuracy suggests to be a threshold between increasing and decreasing human trust and system usage. We have also shown that trust can be derived from a series of users' decisions rather than from a single one, and relates to the perceptions of users. A general framework depicting how trust and perception affect human decision making is proposed, which can be used as future guidelines for human-machine collaboration design.

CCS CONCEPTS

• Human-centered computing → User studies • Human-centered computing → User interface design

KEYWORDS

Trust, perception, Decision making, Dynamic process, Machine performance

1 Introduction

Trust has been considered a critical factor affecting the decision, performance, experience and overall capability of humans when they interact with machines. According to Lee and Moray [1], the predictability of a system plays a fundamental role in a human's trust formation. However, due to the sophisticated technologies and increased levels of automation provided by machines today, humans are no longer able to know every technical detail or working mechanism of their machine teammate, and hence determining the system performance based on full system understanding becomes increasingly difficult. As a consequence, in many situations humans actually base their trust on limited perceptions of the machine partner, and make decisions accordingly [2].

Perception can be considered as the processed outcome of different sensory information, which is critical for human decision making. However, due to various reasons, the human mind is not always able to perceive the status and performance of a system accurately: a perception bias may occur which may ultimately compromise the quality of human decision making [3]. One of the most well-known forms of perception bias is the attribution bias as examined by Woods et al., in which people tend to neglect their own faults but attribute them to others, especially machines [4]. This has led to some typical collaboration issues in a human-machine team, such as algorithm aversion [5], when humans are much less tolerant to mistakes made by machines than by themselves. However, very little is known about the cause of the perception bias, or the methods to accurately quantify and mitigate it.

The limited, sometimes incorrect perception of the machine performance can lead to improper trust in the machine. The study of Lee and Moray [6] suggested that in many human-machine teams, for example, in the scenario of supervising an automatic system, human is the final decision maker, which grants them the right to reject suggestion of the system partner or totally abandon the automation. In Muir's works [7,8] it was explained that humans would override the machine if they had a higher confidence in themselves than their trust in the machine. However this is arguable as confidence is another subjective mental construct that can be even more difficult to measure, or to compare with trust. Actually so far there has been very limited knowledge of the quantitative relationship between perception, trust and decision.

The primary aim of this paper is to investigate the three key elements of human-machine teamwork: trust, perception and deception. Specifically, via manipulating the performance of a

simple decision support system, we seek answers to three questions with the findings as follows:

(i) When do people trust a machine teammate, and what is the dynamics of trust? We have found that users' trust, although initially different, approximates the system accuracy after a series of interactions. Furthermore, incremental trust is observed during the interactions with systems of over 70% accuracy, but decreased trust is observed for systems with lower accuracies.

(ii) How do users perceive the performance of the automatic systems and themselves in the human-machine collaboration context? Overall users are well able to perceive and estimate the system performance and discriminate their relative accuracies within limited trials. For the less accurate systems, users demonstrate a better estimation on their own performance than the system performance.

(iii) What is the implication of perception on user's trust, and further on decision making? Their mutual dependency is proposed as our understanding of decision making process, and we have also shown that trust can be inferred from a series of decisions rather than one or several single decisions.

The rest of the paper is organized as follows: existing literatures related to the relationship between perception and decisions are introduced in the next section, followed by the description of our experimental design, procedure and introduction of the data we have collected in the methodology section. In the result section, our findings are illustrated, showing the patterns of users trust, perception and performance over time and their mutual relations. We explained our findings and discussed their implications for future human-system interaction design in the discussion section before concluding the paper.

2 Related Work

The concept of trust roots back to the relationship between humans, and reflects the subjective willingness to collaborate with others. In the human-machine joint team scenario, trust has been considered as an attitude that an agent will help to achieve an individual's goal in a situation characterized by uncertainty and vulnerability as defined by Lee & Moray [6]. Existing research has revealed different findings regarding trust that is consistent with our intuitions: users tend to use machine that they trust but abandon those that they do not trust [9,10], different users have different trust propensity to the same machine [11,12] , system failures negatively affect trust but good performance of system helps to improve trust [13,14], and appropriate trust is beneficial to human-machine collaboration [8,9].

Basically, the work of Bernard [15] and Zuboff [16] provides theoretical foundations for the composition of trust, which proposes that human-machine trust is built on four dimensions, including natural laws, performance, transparency and design purpose. Natural laws provide the context under which the trusting relationship is possible, and regulates the basic behaviors of humans and machines. For example, fuel or electricity are necessary power for a machine to function properly. Performance indicates whether a machine will behave as expected, and how well it is capable of conducting a task. Transparency refers to human's understanding of the technical process that the machine partner is undergoing, or interpretations of the performance of the machine. The last dimension, design purpose, reflects the designer's intention for the function of a machine. Most research on trust have been conducted on the performance and transparency dimensions, as they directly relate to the overall human-machine team performance [17–19].

System performance is often manipulated via the occurrence of failures, which have always been key issues in the research of trust dynamics and affect the way people make decisions. Lee & Moray have used a simulated pasteurization system to induce consecutive system failures [1], and proposed that trust in a machine is associated with overall human-machine joint performance, system's fault and user's prior trust. Moray et al. further revealed that reliability of automated fault diagnosis, mode of fault management (manual vs. automated), and fault dynamics strongly affect subjective trust in the system, and operator self-confidence [10]. Sauer et al. investigated the effects of automation failures in training on trust and found that automation bias (a tendency to follow the recommendation of the automation) is high when users are trained on a miss-prone automation, which may ultimately lead to more user errors [20]. O'Donnovan et al. also proposed to elicit trust from system recommendation errors [21]. Many more work investigating the implications of system failures on trust can be found in the review of Muir [7,8], although very few of them provide quantitative interpretations on the relationship between trust and system performance. Some recent research has shown the implications of system failures on the dynamics of trust in a quantitative way [22,23], which paved the way towards further refined human-machine trust examination.

Along with the study of system failures and human trust, many attempts have been made in trust measurement, amongst which surveys and behavior-based methods are most popular [24]. The surveys are normally conducted before and after an experiment, asking the participants to rate their subjective trust in a given system [25,26]. They are helpful in determining the cause of trust and the overall subjective attitude towards the system. However, the survey-based methods often fail to capture the dynamics of trust, as people may not trust a machine exactly at the same level all through a thirty-minute experiment. In comparison, behavior-based trust measurement methods are usually based on the decisions of users in several final trials as conducted by Lee & Moray [8]. If a human makes decisions consistent with the system's suggestions, it is considered that the machine is trusted, otherwise it is not. Evaluating trust based on behaviors in this way may not be accurate, due to the fact that trust cannot be assumed to be binary, and there can be many intermediate levels between trust and distrust [27]. Furthermore, the mapping between decision consistent with a machine and trust in a machine is questionable: human may make decisions opposite to their actual

trust, in the case that the cost of incorrect decision is low as revealed in the study of Sutherland et al. [28].

Perception is another factor that relates closely to trust and decision, and a perceptual-motor system in human mind is suggested that affects the cognition and subsequent behaviors [29]. It is also demonstrated that perceptions contribute to the history-based trust, and the former play an important mediating role between human and machine [14]. Further evidence can be found on user trust and reliance and perception of automated decision aids, where perceived reliability is often lower than actual system reliability, and false alarms significantly reduce user trust in the automation [30]. In contrast, Cosmides & Tooby argue that humans can be good intuitive statisticians that are capable of making reliable judgements under uncertainty [31].

As a consequence, this work will revisit the question on how trust develops dynamically, and examine the capability of users to perceive differences in system performance. Due to the disadvantages of existing trust measurement methods, this examination also aims to identify new reliable means to measure trust. Furthermore, very few studies have disclosed the dynamics of trust, decision and perception, while in this study we aim to fill the gap.

3 Methodology

We consider the decision making process by human to be an essential part of human-machine interaction. To keep the potential of generalizing our investigation results to real-life systems, we adopted binary decision making tasks in our experiment, and postulate that any complex decision process can be decomposed into a series of atomic binary decisions. Furthermore, the simplified binary decision making protocol we implement is essentially similar to the micro-worlds discussed by Lee and See [32], which makes it convenient to map trust levels to decisions without the interference of other factors.

3.1 Scenario

This experiment simulated a quality control task in a drinking glass making factory. The users were asked to determine the condition of glasses, a binary choice between good or faulty. To make this decision, they only received the assessment from a simulated decision support system we call Automatic Quality Monitor (AQM), which alerted the user to potentially faulty glasses. However, the AQM did not always function properly and occasionally exhibited false positives (suggesting examining a good glass) and false negatives (suggesting passing a faulty glass). Hence, the trust the user placed into the AQM might fluctuate depending on the performance of the AQM, allowing us to explore the dynamics of trust.

3.2 Tasks

The experiment took place in a laboratory setting through a simple graphical user interface and was arranged in blocks of trials. Each individual trial started with the AQM providing its recommendation about a glass: a red warning light bulb was off for a good glass, or illuminated for a faulty glass (Figure 1), however the glass image on the top right of the interface was not shown. The user then needed to click a ***Pass*** button, if considered the glass was good, or to click ***Examine*** if considered the glass might be faulty. It is important to note that this decision is entirely up to the user who may comply with the AQM's recommendation or override it.

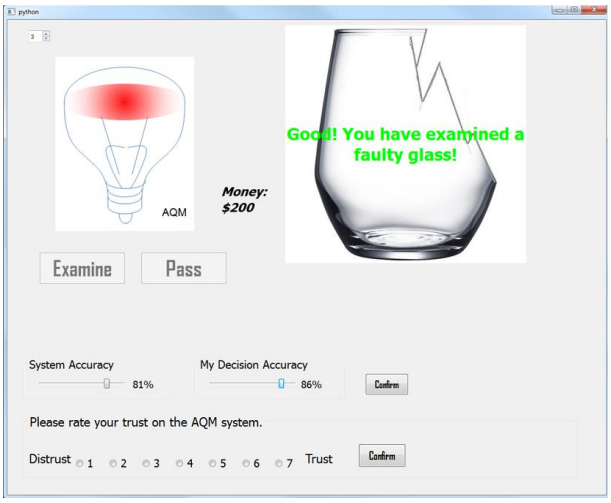

Figure 1: Interface for the experiment: the user is asked to make the decision between Examine and Pass, estimate the accuracy of AQM and their own, and rate the trust level in each trial. Note that the elements in the interface are shown stepwise to the users in the experiment.

After the decisions were made, the users were shown the actual condition of the glass, providing them with direct feedback on whether their decision was correct, as illustrated in Figure 1, where the user correctly decided to examine a glass that proved to be faulty.

In order to increase motivation and attention we gamified the interaction by introducing a fictitious $100 reward for each correct decision (examining faulty glass, or passing good glass) and $100 fine for each incorrect decision. The total earnings were updated and displayed after each decision. The users were aware that these rewards are only to help them track their score, without any actual remuneration offered.

After each trial the users were asked to input both the accuracy of the AQM and their own based on their subjective perceptions, using sliders ranging from 0% to 100% as shown in the bottom part of Figure 1. The users were informed that the accuracies refer to the ratio of correct decisions or AQM recommendations for all the prior trials within in a block. The users were also requested to indicate their level of trust in the AQM using a 7-point Likert scale ranging from 1: distrust, to 7: trust. In the instructions issued

at the outset of the experiment we explained that a rating of 4 meant neutral, or no disposition in either direction.

3.3 Block Assignment

The trials were randomly presented, providing a time-based history of interaction with a given AQM, and allowing us to explore how trust builds up or degrades over time based on the AQM's performance. The users interacted with a number of AQMs, for 30 trials with each AQM, and were told that a different AQM was used for each block; indeed, each AQM's accuracy was manipulated by varying the average rate of false positives and false negatives for every ten trials. For example, for the 80% AQM, two random machine errors occur between the trial 1 and trial 10, between trial 11 and trial 20, and between trial 21 and trial 30 respectively. This arrangement is made to serve two purposes: firstly, the occurrences of system failures do not cluster together; secondly, we can have three check points, i.e. trials 10, 20 and 30, where we can conduct quick checks on how much the perceived system accuracy deviates from the actual system performance. The experiment session involves seven randomized blocks of 30 trials each, and one 100% AQM block of ten trials prior to the seven randomized blocks to serve training purpose as shown in Table 1.

AQM Accuracy	False Neg. + False Pos.
100% (Training)	0%
90%	10%
80%	20%
70%	30%
60%	40%
50%	50%
40%	60%
30%	70%

Table 1. AQM accuracies in the experiment with respective false positives and false negatives.

We admit that in most realistic scenarios, people rarely interact with systems with accuracies as low as 30% or 40%. However, for those systems dealing with uncertainty, for example, some prototype systems or instable systems, their performance are hardly predictable and may be low. The low performance can also be encountered when a normal system malfunctions in a given period of time, and hence people may need to deal with such systems from time to time, and that is the reason we intentionally involve low accuracy systems in the research.

3.4 Participant

Thirty participants including four females took part in this 45 minute experiment as users of the AQMs. 23 of them were university students and the rest were IT professionals. No specific background or preparations were required to complete the experiment. Recruitment and participation were conducted in accordance with a University-approved ethics plan for this study. Snacks were offered for taking part in the experiment, and a gift voucher of $50 was offered in a draw after the experiment as a means of acknowledgement.

3.5 Data Collection and Processing

For each trial we collected:

• AQM's suggestion (light on or off);

• User's binary decision (pass or examine);

• Actual glass condition (good or faulty);

• Perceived system performance (0% to 100%);

• Estimated self-performance (0% to 100%);

• Subjective trust rating.

• We derive the following variables for each trial:

• Normalized subjective trust rating: For each user, all the inputs across all blocks are used to normalize the ratings in the [0, 1] range. More specifically, for all the trust ratings of a user, the normalized trust value T_i after trial i is calculated as

$$T_i = \frac{T_{io} - T_{\min}}{T_{\max} - T_{\min}} \quad (1)$$

where T_{io} is the originally provided trust rating of the user for a trial, T_{max} and T_{min} are the maximum and minimum trust ratings respectively given by the same user across all seven AQMs.

• Reliance rate: the proportion of decisions consistent with the system suggestions over a set number of consecutive trials, in the [0, 1] range.

4 Results

The results shown below comprise the decision behaviors and subjective ratings of all the users. To demonstrate the dynamic changes of trust, perception and decisions, the results will be presented along the 30 trial timeline wherever possible.

4.1 Trust Dynamics

The normalized trust of all the AQMs averaged across all the users is plotted in Figure 2. At the beginning, i.e. the trust rating after the first trial, the order of user's trust in the AQMs is randomized for all the AQMs according to an ANOVA examination ($F(6, 174)=1.28$, $p>0.05$), indicating that the users do not differentiate their trust significantly after a single trial, due to limited experience with the systems. Although visually, trust in the 90% AQM is higher than the rest, a comparison with the 80% AQM after the first trial does not show a significant difference ($t=1.54$, $p>0.05$).

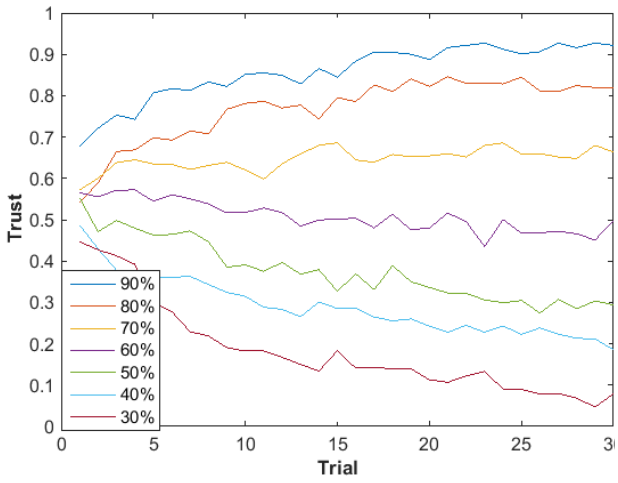

Figure 2: The mean trust of all users for all the AQMs.

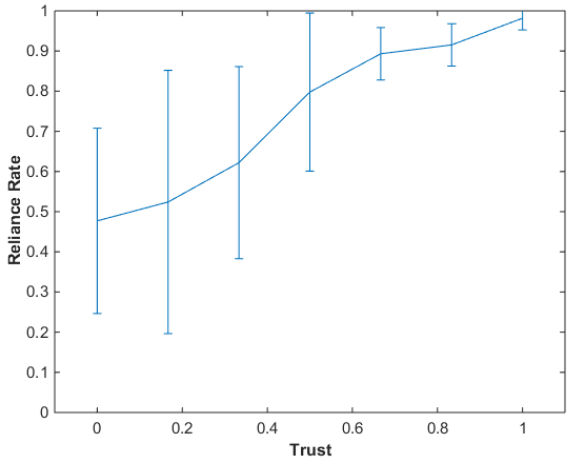

Figure 3: Trust affects the trend and variance of users' reliance rate (R_r). The error bars in the plot represent standard deviations.

As the users continue working with the AQMs, after trial 5 all the trust levels are well separated and align with the accuracies of the respective AQMs. Furthermore, it is found that from trial 5 onwards, the users have demonstrated different trust for the AQM (using ANOVA with repeated measures for the trust levels of individual users after trial 5, $F(6, 174)=20.88$, $p<0.05$). The trend of trust level separation continues towards the end of the trials, however examined with a t-test between trial 25 and trial 30, there are no more significant trust changes ($t= 0.18$, $p>0.05$), suggesting that trust levels have become stable.

4.2 User Decision Affected by Trust

The implications of users' trust on their decisions are investigated via examining the responses of all the users at different trust levels. We calculate the reliance rate R_r of users as the proportion of consistent decisions with the system over a set number of consecutive trials:

$$R_r = \frac{N_c}{N_c+N_d} \qquad (2)$$

where N_c is the number of user decisions consistent with what the AQM light indicates, and N_d is the number of decisions made different from the suggestion of the AQM. As shown in Table 2, based on the trust score for individual trials, when the users highly trust the AQM systems where the trust levels are 6 and 7, they rely on the system for decisions and there is no significant reliance difference between the early and late sections. In contrast, at trust levels 1 to 4, significant trust difference between the sections have been identified via repeated measures ANOVA examination ($F(5, 3)=19.9$, $p<0.05$), indicating that if users do not trust the system so much, they will decrease their usage of the systems.

Examining the individual columns of Table 2, a steady trend can be observed that the reliance rate decreases with trust ratings. An ANOVA test shows the significant difference between trust levels in terms of reliance rate ($F(6, 5)=53.8$, $p<0.05$), which suggests that when users rate low trust, they rely less on the suggestions of the system.

The relation between trust and reliance rate is further depicted in Figure 3. The error bars indicates the variance at each trust level, and the trust of all users is normalized to the [0,1] range. It should be noted that the data from all the users are plotted in this figure, however for individual users a similar trend is observed as

Trust level	*Trial [1,5]*	*Trial [6,10]*	*Trial [11,15]*	*Trial [16,20]*	*Trial [21,25]*	*Trial [26,30]*
7	0.973	0.951	0.972	0.99	0.985	0.983
6	0.957	0.913	0.932	0.927	0.91	0.927
5	0.912	0.896	0.866	0.824	0.819	0.827
4	0.905	0.765	0.745	0.742	0.679	0.717
3	0.817	0.682	0.669	0.613	0.713	0.523
2	0.766	0.624	0.53	0.597	0.503	0.509
1	0.797	0.6	0.527	0.535	0.508	0.465

Table 2: Reliance rate (R_r) at different trust levels. The trust levels are the original ratings of the users. The 30-trial block is segmented into 6 sections, each composing 5 trials in each column.

well. The reliance rate demonstrates a clear rising trend with trust, suggesting that users rely more on systems when they trust them which is consistent with existing understanding. On the other hand, the decreasing variance of reliance rates reveals another interesting finding: at low trust levels, although the overall reliance rate are low, users demonstrate high variance in reliance rates. This suggests that users rely on the system in different ways, sometimes even if they do not trust the system, they may try decisions consistent with its recommendation. In comparison, as trust level increase, the rate of reliance also converge, implying that users tend to follow the system suggestions when they believe the system to be highly reliable.

4.3 User Performance and Perception

Performance refers to the proportion of correct decisions amongst all the decisions made on one AQM. We have asked users to estimate their performance based on their estimation on all prior trials. In the meanwhile via comparing the decisions of users with the outcome of glasses, we are able to calculate their actual performance.

Figure 4 shows both the actual performance of the users and the perceived performance of their own. Interestingly, in the initial several trials users are not able to precisely estimate their performance, although it is easier compared with situations when more trials have been done. It should be noted that if a user is good at memorizing the previous trials, he/she should be able to increase the accuracy of performance estimation as she/he approaches the end of the 30 trials. An interesting finding from Figure 4 is that at the end of the trials, for the more accurate AQMs (90%, 80% and 70%), users' estimated accuracies are significantly higher than their actual performance; however they are still capable of discriminating the order of these AQMs. Table 3 shows the difference between the perceived and realistic performance of users and whether it is statistically significant

AQM	90%	80%	70%	60%	50%	40%	30%
F	13.86	21.22	7.99	0.13	4.42	0.003	0.06
p	<0.01	<0.01	<0.01	0.73	0.04	0.95	0.80

Table 3: Differences between actual and perceived user accuracies at trial 30 with repeated measures ANOVA: users are better capable of estimating their own performance when working with less accurate AQMs.

(using repeated measures ANOVA), from which we can see that for the less accurate AQMs, users estimated their performance better than when they were working with the more accurate AQMs.

4.4 Perception of System Performance

If the users estimate their own performance differently from their real performance, how about their perceptions on the AQMs? Figure 5 provides the answer and depicts the dynamics of AQM perceptions. The results suggest that the users are capable of perceiving the system performance with high accuracy. At the fifth trial, the perceived system accuracies for different AQMs already differ significantly based on repeated measures ANOVA ($F(6, 174)=27.69$, $p<0.05$). A paired t-test between trial 25 and trial 30 ($t=0.46$, $p>0.05$) indicates that towards the end of the 30 trials, there are no more significant perception changes for all AQMs, implying that the perceived system accuracies have stabilized. These findings imply that the users are able to adjust their perceptions and reach accurate estimations towards the end of the trials, especially for the most accurate AQMs (90%, 80% and 70%). For the other less accurate AQMs especially the 50%, 40% and 30% ones, perception bias of over 10% can be observed towards the end of the trials, but the order of accuracy is still correctly perceived.

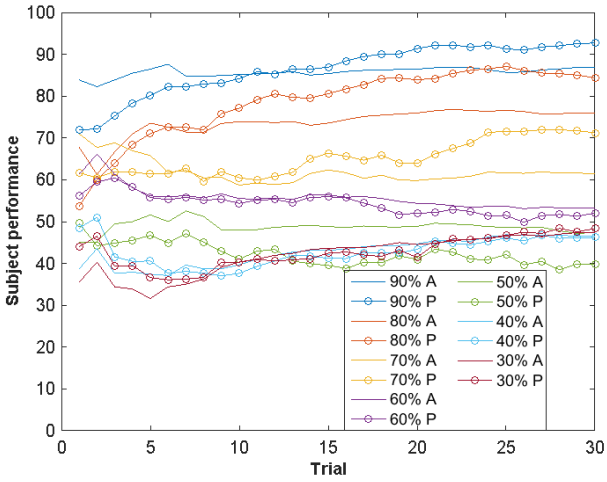

Figure 4: Perceived vs actual subjective performance, where 'A' denotes actual performance and 'P' denotes perceived performance of the user.

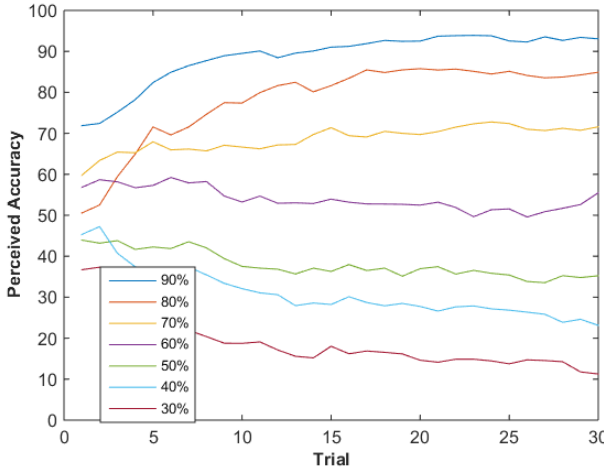

Figure 5: User perceptions of the AQM accuracies.

4.5 System Perception and User Decisions

Due to the similarity between perceived system accuracy and users' trust in the AQMs, we would like to see how the system perceptions affect user decisions. Table 4 illustrates all the user's decisions at different levels of perceived AQM accuracy. It suggests that the higher a system's performance is perceived, the more decisions consistent with the system are made. However, noting the reliance rate at the top of the table, even if the perceived system accuracy is extremely low, the user may still take a chance to follow the system's suggestions now and then, although overall a decreasing trend is suggested when the perceived system accuracy is below 70%.

For all the users, the relationship between their perceived AQM accuracies and the rate of reliance is illustrated in Figure 6. A linear regression is calculated to predict the reliance rate based on the perceived accuracy. A significant regression equation is found ($F(1,99)=187.42$, $p<0.05$) with an r^2 of 0.654. The predicted trend of reliance rate R_r with perceived accuracy is

$$R_r = 0.47 \times P_a + 0.521 \quad (3)$$

where P_a is the perceived accuracy range from 0% to 100%. This finding implies that as the perceived accuracy increases, users rely more on the recommendations of the AQMs. It should be noted that for the majority of the cases the reliance rate is above the chance level of 0.5, even when the perceived accuracy of the systems is very low, which is consistent with our finding shown in Figure 3. Intuitively, the regression coefficient 0.47 indicates that the reliance increase is about two times slower than the system perception increase.

5 Discussion

The results of this study provide evidence on several important findings regarding perception, trust and human decision, and reveal their mutual relationship when users interact with machines as a collaborative team member. We have shown that users are

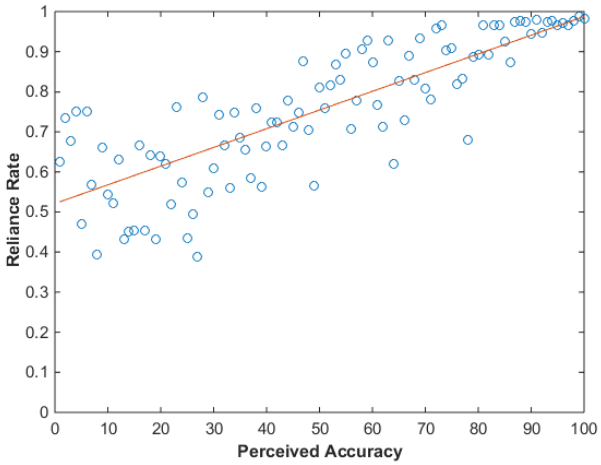

Figure 6: Reliance rate (R_r) increases with perceived system accuracy for all users. The linear regression result is shown in red.

capable of estimating the system accuracies reasonably well and gradually adapting their trust levels to the system performance within 30 trials. The positive relationship between trust and user perception suggests the tight link between the two mental constructs. This finding implies that if a user perceives the performance of a system, their trust in the system will be affected accordingly; furthermore, the increased trust may result in more decisions consistent with the recommendation of a decision support system.

Examining the way trust and system perception evolve, it is found that after five trials, both trust and user perception are well separated, indicating that the users are capable of discriminating the performances of the systems very quickly and trust them accordingly, although the accuracy of perception can be incrementally improved later as more interaction occur and more experience gained. After 25 trials, both the trust level and the perceived system performance reached a stable level, and we can infer that no significant change of them will happen if the user

System Accuracy Perception (%)	*Trial [1,5]*	*Trial [6,10]*	*Trial [11,15]*	*Trial [16,20]*	*Trial [21,25]*	*Trial [26,30]*
0-10	0.729	0.658	0.536	0.603	0.524	0.453
11-20	0.712	0.604	0.608	0.623	0.512	0.477
21-30	0.815	0.591	0.483	0.603	0.521	0.544
31-40	0.744	0.723	0.621	0.531	0.677	0.522
41-50	0.944	0.734	0.673	0.701	0.586	0.560
51-60	0.895	0.890	0.785	0.754	0.729	0.753
61-70	0.861	0.768	0.867	0.771	0.750	0.773
71-80	0.926	0.925	0.933	0.788	0.860	0.788
81-90	0.959	0.938	0.911	0.953	0.898	0.940
91-100	0.975	0.925	0.955	0.959	0.990	0.984

Table 4: Reliance rate (R_r) at different system perception levels. The perception levels are segmented into 10 intervals respectively, e.g. 11-20 means the perception rating interval between 11 and 20. The 30-trial block is segmented into 6 sections, each composing 5 trials in each column.

continue interacting with the systems.

These findings reveal two important aspects of interactive system design, especially for the decision support systems like the AQM used in our experiment. Firstly, the users are capable of comparing system performances after a limited number of trials. In that sense, we can hypothesize that for a system to function properly, special attention should be paid when the user just starts to use it, as the outcome of these trials will significantly affect the future trend of user trust change. For a system as simple as the AQM, the first five trials are of prior importance to shape user's trust. Secondly, as it takes longer for users to perceive the actual performance of the system, sufficient interaction should be allowed if the designer wants to know how people usually use the system. Approximately 25 iterations of interaction have occurred before the user's trust and perception become stable, however we can imagine that if working with a more complicated system, more interaction time with repeated interactions will be required before reaching a reasonable understanding of users' trust feeling about it.

It can be observed in Table 2 and Figure 3 that even if at the same trust level, the users may not always make the same decisions. This finding has shed light to the way trust is measured using behaviors, while it can be misleading if the outcome of a single decision or several limited decisions are considered as indicators of trust even after a long period of interaction. In our view, behavior-based methods can be improved via using the reliance rate as shown in Figure 3, which increases with the increment of trust. Another optional choice for trust measurement is the variance of decisions, however this measurement may not be reliable enough alone especially when trust levels are low, and it is possible to combine the reliance rate and decision variance for better measurement of trust.

Another interesting finding is that the users perceive the performances of the interaction system and themselves differently. Comparing Figure 4 with Figure 5, apparently users have better overall estimations on the system performance than themselves. Furthermore, when working with the three relatively high performance AQMs, the users have significantly overestimated their own performance but their perceptions on the system performance are reasonably good. When working with the low accuracy systems, users' self-estimations do not differ much from their actual performance, but their estimations on the system performance are less accurate. Revisiting the question of whether humans are good intuitive statisticians, our result is consistent with Cosmides & Tooby [31] that humans are good at perceiving uncertainties and make judgement accordingly, however to be more accurate, it should be further addressed that the capability of human to perceive uncertainties is related to the object being estimated.

Based on the dynamics of trust and user perception of the system, it can be observed that the 70% accuracy is the threshold between the increase and decrease of trust and system perception. We can also see that based on the self-estimation of performances, the users overestimated their performance when the system accuracy is no lower than 70%, suggesting that users' self-confidence is higher when working with such systems. Existing research has shown that a user's self-confidence generally enhances motivation [33] and relates to the tendency to make improvements when interaction with systems [34]. As a consequence, it can be inferred that 70% accuracy is a threshold that automatic system designers should consider, above which users are able to grow trust and achieve good system perceptions with better self-confidence.

Although we endeavor to provide quantitative examinations for all the findings, there are three limitations that should be highlighted and discussed. Firstly, the AQM system we designed is a typical form of the simplest decision support systems, in which the recommendation accuracy is the only factor considered. The users demonstrated an overall reliance rate over 50% for all the AQMs, which implies that for such systems with binary decisions, overall more than half of the decisions are made consistent with the system's recommendation, although for the least trustworthy systems the final reliance rate dropped below 50% as shown in the last row of Table 2. However, many realistic systems are much more complicated, and the trust and perception of them can be much difficult to characterize in a quantitative way. As a consequence, it will be necessary to examine every single factor involved in other systems, e.g. system transparency, complexity and modality of interaction, before generalize the current findings to them. Secondly, the findings in this study is mainly correlational, which may limit the causal conclusions that can be drawn from this study. Finally, in the examinations we do not consider the implications of prior trials or the effect of consecutive positive or negative system performance, although this has been addressed in another study [23]. Combining the findings from both investigations will produce a full picture of how users perceive and trust a decision-support system.

The present study, being quantitative, revealed a number of findings that should be considered in interaction system design and analytics. Furthermore, there are a few directions of interest to be examined in the coming research. Currently all the AQMs are featured with a fixed accuracy, however for many realistic systems their performance may not be stable. We are interested in how users perceive, trust and interact with a system of dynamic performance, and in which way the dynamics of the system is able to affect the users' attention and perception. Generalization is another issue to examine – whether our findings can be used to interpret the interaction patterns with other types of design support systems will be examined.

6 Conclusion

In this study, we investigated user trust, perception and decisions in the human-machine interaction context and revealed how they interplay with each other. Overall the results indicate that users are capable of perceiving the performance of themselves and systems, adjusting their trust and decision schemes accordingly.

We also propose that trust can be measured via repeated user decisions instead of isolated ones, and can be inferred from the subjective perceptions of the machine performance. Finally, our examinations uncover that 70% is the system accuracy threshold that determines whether users will trust and use the system with high self-confidence. So, back to the key question: "Do I trust my machine teammate?" The answer lies in how the machine is designed, perceived and interacted, and can be detected via the user decisions and perceptions as revealed in this study.

ACKNOWLEDGMENTS

This research is supported by the AOARD grant FA2386-18-1-4091. The authors would also like to acknowledge all the participants for their time and helpful comments.

REFERENCES

[1] J. D. Lee & N. Moray (1994). Trust, self-confidence, and operators' adaptation to automation. International journal of human-computer studies, 40(1), 153-184.

[2] M. T. Ribeiro, S. Singh & C. Guestrin (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

[3] E. Pronin (2007). Perception and misperception of bias in human judgment. Trends in cognitive sciences, 11(1), 37-43.

[4] D. D. Woods, L. J. Johannesen, R. I. Cook, & N. B. Sarter (1994). Behind human error: Cognitive systems, computers and hindsight (No. CSERIAC-SOAR-94-01). DAYTON UNIV RESEARCH INST (URDI) OH.

[5] B. J. Dietvorst, J. P. Simmons, & C. Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

[6] J. D. Lee & N. Moray (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270.

[7] B. M. Muir (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics, 37(11), 1905-1922.

[8] B. M. Muir & N. Moray (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics, 39(3), 429-460.

[9] J. D. Lee & K. A. See (2002). Trust in computer technology and the implications for design and evaluation. Etiquette for Human-Computer Work: Technical Report FS-02-02, 20-25.

[10] N. Moray, T. Inagaki & M. Itoh (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. Journal of experimental psychology: Applied, 6(1), 44.

[11] P. Madhavan & D. A. Wiegmann (2007). Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277-301.

[12] S. Sarkar, D. Araiza-Illan & K. Eder (2017). Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker. arXiv preprint arXiv:1703.02335.

[13] M. T. Khasawneh, S. R. Bowling, X. Jiang, A. K. Gramopadhye & B. J. Melloy (2003). A model for predicting human trust in automated systems. Origins, 5.

[14] S. M. Merritt & D. R. Ilgen (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. Human Factors, 50(2), 194-210.

[15] B. Bernard (1984). The Logic and Limits of Trust. (New Brunswick, NJ: Rutgers University Press, 1983. Pp. 190. $27.50, cloth; $9.95, paper.). American Political Science Review, 78(1), 209-210.

[16] S. Zuboff (1988). In the age of the smart machine: the future of power and work. New York: Basic.FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[17] S. Rice & K. Geels (2010). Using system-wide trust theory to make predictions about dependence on four diagnostic aids. The Journal of general psychology, 137(4), 362-375.

[18] E. Onal, J. Schaffer, J. O'Donovan, L. Marusich, S. Y. Michael, C. Gonzalez, & T. Hollerer. (2014, March). Decision-making in abstract trust games: A user interface perspective. In Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on (pp. 21-27). IEEE.

[19] D. Holliday, S. Wilson, & S. Stumpf (2016, March). User trust in intelligent systems: A journey over time. In Proceedings of the 21st International Conference on Intelligent User Interfaces (pp. 164-168). ACM.

[20] J. Sauer, A. Chavaillaz, & D. Wastell. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. Ergonomics, 59(6), 767-780.

[21] J. O'DONOVAN & B. Smyth (2006). Mining trust values from recommendation errors. International Journal on Artificial Intelligence Tools, 15(06), 945-962.

[22] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, & F. Chen (2016). Trust and reliance based on system accuracy. In Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (pp. 223-227). ACM.

[23] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, & F. Chen (2017). User trust dynamics: An investigation driven by differences in system performance. In Proceedings of the 22nd International Conference on Intelligent User Interfaces (pp. 307-317). ACM.

[24] E. L. Glaeser, D. L. Laibson, J. A. Scheinkman & C. L. Soutter (2000). Measuring trust. The quarterly journal of economics, 115(3), 811-846.

[25] J. M. McGuirl & N. B. Sarter (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors, 48(4), 656-665.

[26] Z. Yan, V. Niemi, Y. Dong & G. Yu (2008, June). A user behavior based trust model for mobile applications. In International Conference on Autonomic and Trusted Computing (pp. 455-469). Springer, Berlin, Heidelberg.

[27] R. J. Lewicki, D. J. McAllister & R. J. Bies (1998). Trust and distrust: New relationships and realities. Academy of management Review, 23(3), 438-458.

[28] S. C. Sutherland, C. Harteveld & M. E. Young (2016). Effects of the Advisor and Environment on Requesting and Complying With Automated Advice. ACM Transactions on Interactive Intelligent Systems (TiiS), 6(4), 27.

[29] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, & Y. Qin (2004). An integrated theory of the mind. Psychological review, 111(4), 1036.

[30] J. D. Johnson, Sanchez, A. D. Fisk, & W. A. Rogers (2004, September). Type of automation failure: The effects on trust and reliance in automation. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 48, No. 18, pp. 2163-2167). Sage CA: Los Angeles, CA: SAGE Publications.

[31] L. Cosmides & J. Tooby (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition, 58(1), 1-73.

[32] J. D. Lee & K. A. See (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

[33] R. Bénabou & J. Tirole (2002). Self-confidence and personal motivation. The Quarterly Journal of Economics, 117(3), 871-915.

[34] P. Briggs, B. Burford, & C. Dracup (1998). Modelling self-confidence in users of a computer-based system showing unrepresentative design. International Journal of Human-Computer Studies, 49(5), 717-742.