Do I trust a machine? Differences in user trust based on system performance

Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen

Abstract Trust plays an important role in various user-facing systems and applications. It is particularly important in the context of decision support systems, where the system's output serves as one of the inputs for the users' decision making processes. In this chapter, we study the dynamics of explicit and implicit user trust in a simulated automated quality monitoring system, as a function of the system accuracy. We establish that users correctly perceive the accuracy of the system and adjust their trust accordingly. The results also show notable differences between two groups of users and indicate a possible threshold in the acceptance of the system. This important learning can be leveraged by designers of practical systems for sustaining the desired level of user trust.

1 Introduction

Trust is a critical factor that impacts interpersonal relationship, and it used to be established via face-to-face communications between people until technologies made human-machine communications possible. The extensive usage of internet everywhere in the world has boosted the information revolution, under which circumstance the human alone is not capable of processing the vast amount of information which is booming exponentially over time, so people may resort to computers for help now and then. However, switching from a smiling colleague to a cold emotionless machine, the human does need time and experience to build up trust with the new partner, although it is capable of conducting many tasks that are beyond the capability of a human. In this context, it is particularly important for systems where users are required to make decisions based, at least partially, on machine recommen-

K. Yu · S. Berkovsky · D. Conway · R. Taib · J. Zhou · F. Chen
Data61, CSIRO
5/13 Garden St, Eveleigh NSW 2015, Australia
e-mail: {firstname.lastname}@data61.csiro.au

dations. For instance, consider a medical decision support system or an e-commerce recommender system. In both cases, a user decides on the course of actions be it medical treatment for a patient or product to purchase in uncertain conditions and based (in part) on the system's suggestions. Since in both cases there is something at stake, i.e., there are possible negative implications for incorrect decisions, the lack of user trust may deter the user from following these suggestions and be detrimental to the uptake of system.

Trust in automation, and, in particular, in decision support information technologies, has been the focus of many studies over the last decades [5, 7]. It has mainly been studied in the context of task automation and industrial machinery. In one of the seminal works in this field, Muir et al. [13] found a positive correlation between the level of user trust and the degree to which the user delegated control to the system. Furthermore, McGuirl and Sarter [11] found similar responses specifically within an automated decision support system. Note that both works highlighted the impact of establishing and maintaining trust on user reliance on system suggestions, and, indirectly, on the uptake of the system.

Although much work has been devoted to the impact of system performance [18] and transparency [21] on user trust, less attention has been paid to the temporal variations of trust, and to individual differences of such dynamic aspects. In this chapter, we discuss our investigations on the fine-grained dynamics of trust in an experiment that simulates an Automated Quality Monitoring (AQM) system that alerts users to of the existence of faulty items, in a fictional factory production line scenario. In the experiment, every one of the 22 participants interacted with four AQM systems each exhibiting a different level of accuracy. After each trial (30 per AQM system), the users reported their perceived level of trust in the system, which we refer to as explicit trust. In addition, we also measured implicit trust through reliance, quantified through the proportion of times the user followed the AQM's suggestion. It should be noted that for any decision made by the user, reliance for a single task is a binary feature, since it captures whether the user followed (or not) the system's advice.

Three hypotheses guided our examinations:

- H1: Learned trust, i.e. the trust gained after some experience and collaboration, would stabilise over time to a level correlated with the systems' accuracy;
- H2: Users would exhibit thresholds of acceptable accuracy for a system, under which reliance would drop;
- H3: Differences would exist for acceptable accuracy in terms of trust and stereotypical user profiles will still be able to be constructed.

This chapter will address our work which experimentally validates these hypotheses and draws practical conclusions that can help system designers maintain user trust in systems. In the following sections, we first present related work on user-system trust, followed by a detailed description of the experimental protocol. We then present and discuss the results, and finally conclude with a discussion on practical steps that might be taken to sustain user trust.

2 Background

Human-machine trust has generated an extensive body of literature since it was originally investigated within the context of industrial automation systems in the 1990s. Although multiple definitions, frameworks and decompositions of trust exist, there is convergent evidence about its central characteristics. We adopt the definition proposed by Lee and See [8] where trust can be defined as *the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability.* This succinctly encapsulates the primary sources of variance (the user, the system, the context) and identifies a key aspect of this relationship, that of vulnerability. Similar definitions exist by Rousseau et al. [15], Mayer et al. [10] and Hoff and Bashir [5]. Trust is a hypothesised variable that has been shown to be a key mitigating factor in system use/disuse (reliance) [7, 20]. It can be inferred from both self-report and behavioural measures [10], and importantly, is dynamic, with acquisition and extinction curves, subject to the users experience of system performance.

Trust has been proposed to be a multi-dimensional construct with a number of models existing in the current literature, each with slightly different proposed component subscales. We have adopted Hoff and Bashir's model [2], which he based on an empirical research overview of existing literature in the area. This model is also nicely applicable to our research focus in that it includes variables important to HCI contexts such as 'design features' as well as encompassing a number of important situational factors and individual differences such as culture, age, gender and personality. Hoff and Bashir also base their work on the Lee and See's definition of trust as mentioned above.

This model proposes that three conceptual types of factors influence user-system trust. Dispositional trust reflects the user's natural tendency to trust machines and encompasses cultural, demographic, and personality factors. Situational trust refers to more specific factors, such as the task to be performed, the complexity and type of system, user's workload, perceived risks and benefits, and even mood. Lastly, learned trust encapsulates the experiential aspects of the construct which are directly related to the system itself. This variable is further decomposed into two components. One is initial learned trust, which consists of any knowledge of the system acquired before interaction, such as reputation or brand awareness. This initial state of learnt trust is then also affected by dynamic learned trust which develops as the user interacts with the system and begins to develop experiential knowledge of its performance characteristics such as reliability, predictability, and usefulness. The relationships and interaction between these different factors influencing trust are complicated and subject to much discussion within the literature. In our work we focussed on how trust changes through human-machine interaction and therefore seek to manipulate experimental variables thought to influence dynamic learned trust, whilst keeping situational (and initial learned) variables static, and allowing for variation in individual differences via factors affecting dispositional trust.

Individual differences in trust response are a key focus of our research. In the original body of work on human-human trust, Rotter [14] established that trust

(human-human) was a stable character trait and developed an instrument that detected variations in propensity to trust between people. Extending this, Scott [16] demonstrated that trust was composed of at least two factors, one being situational, and the other being a stable, trait based factor (equivalent to Hoff's dispositional trust). When extending the original human constructs into the realm of humans and machines, Singh et al.[17] operationalised the construct of 'complacency' in automation, which included a subscale on 'trust' and found reliable, and stable variations between people. Lee and Moray [7] found differences between people's likelihood in using automation when error rates are held constant.

When comparing human-human to human-machine trust, Madhavan and Wiegmann's [9] review outlines a number of important differences. Jian et al.[6] found that people's ratings are less extreme towards other humans than towards machines. Earley [4] found that people evaluated system estimations as more trustworthy than human equivalents, but in contrast, Dietvorst et al. [2] found that people were more likely to under-rely on an automated aid in decision making even when shown that the machine performed more accurately than their own efforts and even when there was a financial stake involved. On the other hand, Dzindolet et al. [3] notes that human machine trust sometimes begins at a higher level than human-human trust and is characterised by more dramatic collapses when trust is proven to be misplaced. To explain this phenomena he suggested that some individuals harbor a 'Perfect automation schema' where expectations of system performance are unrealistically high. Such expectations result in differential reactions to system-failures, where those who possess this schema exhibit higher loss of trust on system failure than those who do not.

However, as Lee and See [8] and Hoff et al. [5] have claimed, individual differences are likely to be overcome by the experiential effects of steady state machine behaviour resulting in less variance between users after exposure to the machine. The experiment we outline below contradicts this finding to some extent. We have found that clustering users into two groups uncovers two patterns of trust behaviour where one group exhibits greater variance in trust ratings than the other. We use this finding to single out users, whose trust in system may be at risk and take proactive steps to sustain their trust.

3 Methodology

3.1 Context

The scenario of the experiment was a typical production factory quality control task. This simulated task consisted of checking the quality of drinking glasses on a production line, with the assistance of a decision support system called an Automatic Quality Monitor (AQM). However, the AQM was not always correct, i.e.,

it would occasionally exhibit false positives (suggesting failing a good glass) and misses (suggesting passing a faulty glass).

3.2 Trials

Each trial required the participant to make a decision about whether to pass or fail a glass, with no other information about the glass other than the AQM's suggestion. Trials were presented sequentially, providing a time-based history of interaction with a given AQM. At each trial, the participant could trust the AQM or override it and make their own decision. A simple graphical user interface coded in Python and running on a 64-bit Windows operating system was used, as shown in Fig. 1.

Each trial starts with the AQM providing a suggestion for a new glass, by illuminating a red warning light-bulb if it predicts the glass to be faulty. Otherwise the warning light remains off. It should be noted that the status of the AQM light and the possible quality of the glass are both binary features to help generalise results, as mentioned above.

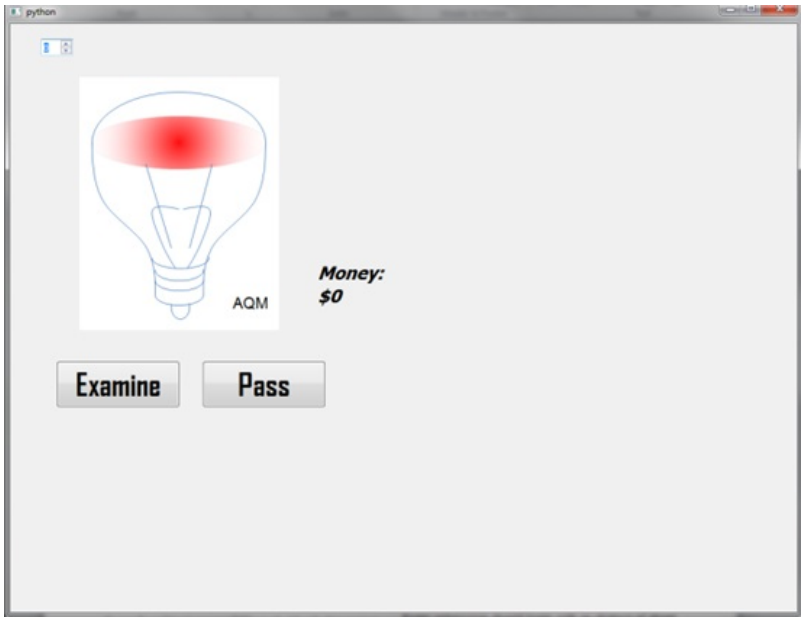

Fig. 1 The trial starts with an AQM recommendation, with two buttons (Examine/Pass) for users to make a decision

The participant must then decide whether to pass the glass by clicking the Pass button, or conversely to fail the glass by clicking the Examine button. The actual glass is then displayed, so the participant receives direct feedback on their decision,

as shown in Fig. 2 and Fig. 3. Furthermore, we gamified the experiment in an attempt to increase motivation and attention: each time the participant made a correct decision, i.e., examined a faulty glass or passed a good glass, they earned a fictional $100 reward. However, each incorrect decision cost them a fictional $100 loss. The total earnings were updated after each decision and displayed within the user interface. The rewards and the fines were used for gamification purposes only, and no actual remuneration was offered to the participants. Exemplary interfaces showing that the user has made correct and incorrect decisions are shown in Fig. 2 and Fig. 3 respectively.

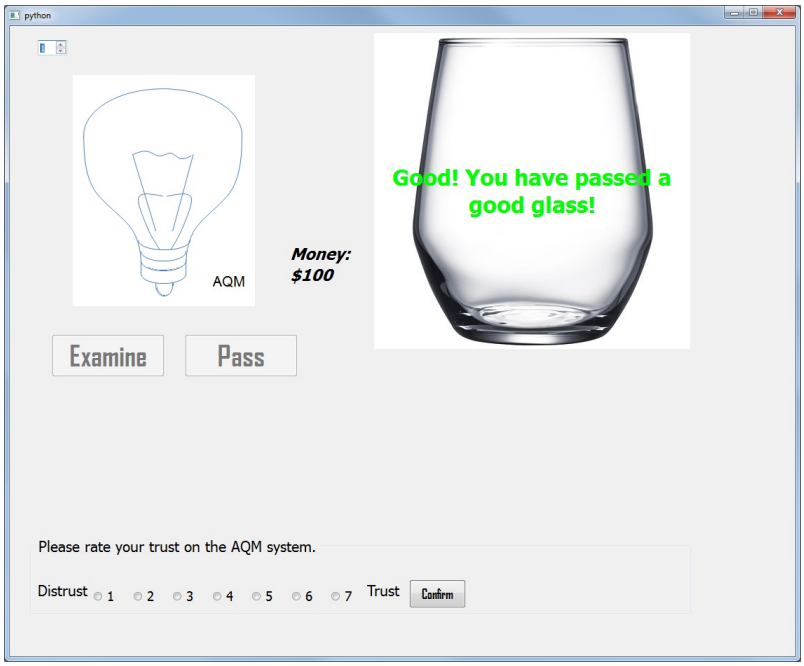

Fig. 2 Upon decision from the participant, the actual glass is shown and score updated.

We operationalised a binary decision making task in our experiment for two reasons. Firstly, any complex decision process can be arguably decomposed into a series of binary decisions. The decision-trust relationship thus can be easily generalised to complicated decision-making problems. Secondly, the simplified decision making protocol we implemented, similar in effect to the 'micro-worlds' discussed by Lee and See [8], makes it convenient to map trust levels to decisions without the interference of other parameters [19].

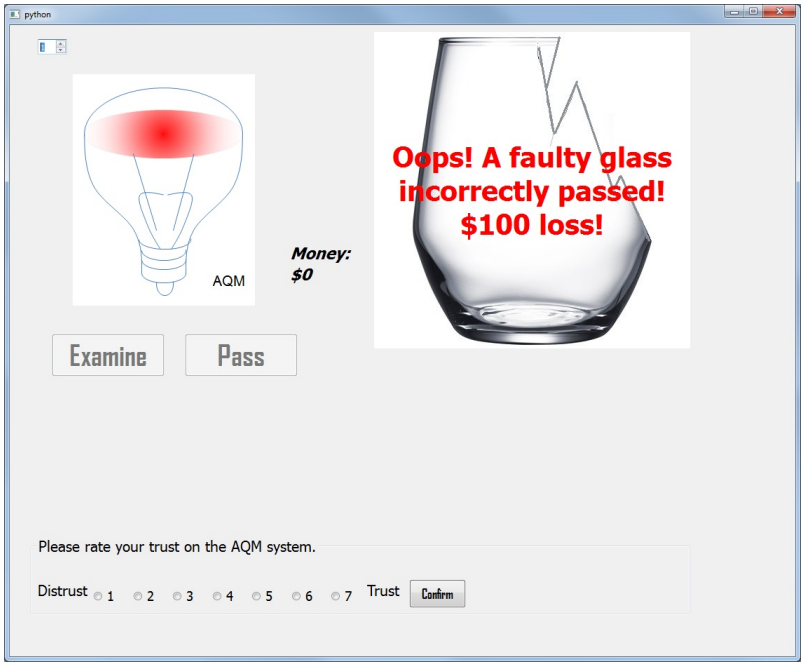

Fig. 3 A wrong decision leads to decreased score, and red text indicating the outcome of a glass.

3.3 AQM Accuracy and Blocks

The experiment session was separated into four blocks, and participants were instructed that a different AQM was used for each block. The accuracy of each of the four AQMs presented was manipulated by varying the average rate of false positives and false negatives exhibited by each system. These errors were presented in a randomised order within the 30 trials presented for each participants and each AQM.

Table 1 AQM accuracies

AQM Accuracy	False positives + negatives
100%	0%
90%	10%
80%	20%
70%	30%

We used four different AQM accuracies, as shown in Tab. 1. In order to capture a trust baseline for each participant, each experiment session systematically started with the 100% accuracy AQM, followed by the other three AQMs in random order. Each block consisted of one AQM that was used for 30 task trials. The AQM made

errors randomly over the trials, but in a way that the mean AQM accuracy over the block was as defined for that AQM. For instance, the 80% AQM would make, on average, 6 errors over the 30 trials (on average, 3 false positives and 3 false negatives).

3.4 Participants

Twenty-two participants took part in the 45 minute experiment. Twenty of the participants were university students and the rest two were IT professionals. No specific background or requirements were required to complete the task. Recruitment and participation were conducted in accordance to the University-approved ethics plan for this study. No reward or compensation was offered for taking part in the experiment.

3.5 Information logging

For each trial, we collected:

- The participant's binary decision (pass or examine);
- The AQM suggestion (light on or light off);
- The actual glass condition (good or faulty);
- The time required to make the decision, i.e., the time elapsed between the AQM light being presented to the participant and the Pass/Examine button being clicked;
- The subjective trust rating, collected after the actual state of the glass is revealed. This rating is collected 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.

One of the participants has consistently rated the trust at extreme levels (either 1 or 7) of the 7-point scale across the four sessions, and hence his data was excluded from the examination. Considering the individual differences, the trust data was normalised to the range of 0 to 1 on an individual basis, for all the trials conducted on the four AQMs. The binary decision of the participants was further quantified in terms of a reliance score R_s, i.e. the ratio between the number of decisions consistent with the AQM recommendation and the total decisions for a set number of consecutive trials, and thus the value of reliance score falls between 0 and 1.

$$R_s = \frac{N_r}{N_r + N_n} \tag{1}$$

where N_r and N_n refer to the number of decisions consistent and inconsistent with the AQM recommendation respectively for all the previous trials on it.

4 Results

In this section we present and discuss the results of our user study in the light of our hypotheses.

4.1 Trust Correlation to System Accuracy

We start with the investigation of acquisition and extinction of trust, as observed over the course of user interactions with the AQMs. The level of trust is measured subjectively after each trial, as described earlier. Since the AQM errors were randomised over the 30 trials for each AQM, and given the number of participants, trust variations for each trial exhibit a number of local variations. We address this issue by applying a simple low-pass filter, specifically a 5-trial sliding window, reducing our data to 25 points per AQM. That is, T_n, the level of trust after trial n, was computed as the average trust across the last 5 trials (n-*4* to n). Fig. 4 shows the aggregated normalised trust for all 21 participants, for all four AQMs.

$$T_n = \frac{\sum_{i=0}^{4} t_{n-i}}{5} \tag{2}$$

where t_n is the trust rating for trial n.

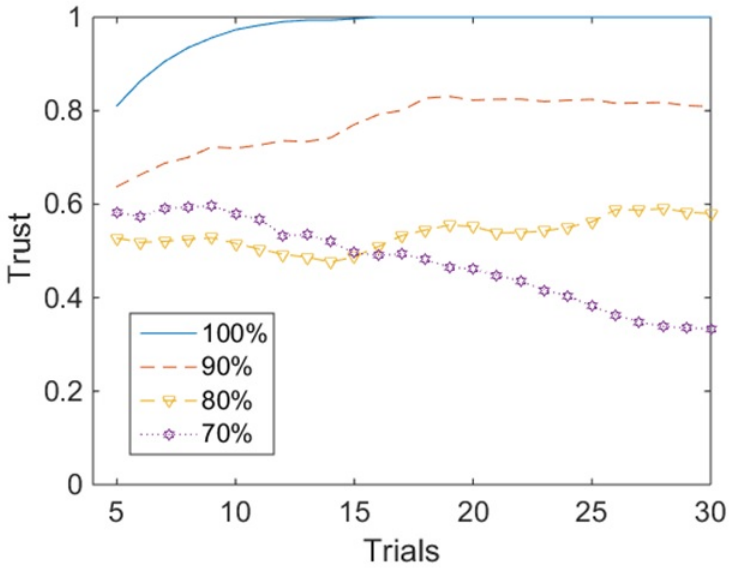

Fig. 4 Mean trust for all participants, all AQMs.

At first, trust in all AQMs seems uniform as would be expected since participants know that each new AQM is different from the others they may have encountered, and the order is randomised. Trust in the 100% AQM appears to be above the other AQMs.

An analysis of variance showed that the effect of the AQM accuracy on the first trust point (trust mean over the first five trials) was significant for all participants, $F(3,80) = 6.463, p < 0.001$. Post hoc Tuckey tests showed there was a significant difference between the 100% AQM and both the 80% and 70% AQMs. We think this may be linked to two possible factors. Firstly, since we use a sliding window to capture trust, the participants will have started to form a preliminary trust judgment on each AQM by the time of the first trust point (recall that the first point is actually after 5 trials). Secondly, it is possible that individual differences between participants combine in a way that creates such a wide initial variation in trust. We investigate this second possibility in later sections of this chapter, by grouping participants and then revisiting their initial trust assessment.

As a side note, the test of homogeneity (Levene's) for the first reliance point was significant, hence violating ANOVA's assumption of equal variances. However, the sample sizes being equal, this statistic should be robust. Hence, we accept the results.

Looking at the temporal fluctuations of the trust values, we observe that these stabilise with important differences between the AQMs. As expected, trust in the 100% AQM stabilises at 1 after 13 trials only. Also the 90% AQM converges to reasonably high levels of trust from trial 19. The 80% AQM is initially stable but exhibits a slight increase in trust starting from trial 15, while the trust in the 70% AQM steadily declines after less than 10 trials and eventually drops as low as 0.33.

An analysis of variance showed that the effect of the AQM accuracy on the last trust point (trust mean over the last five trials) was significant for all the participants, $F(3,80) = 27.03, p < 0.001$. Post hoc Tuckey tests show there is a significant difference between the 100% AQM and both the 80% and 70% AQMs, as well as between the 90% AQM and the 70% AQM, and again between the 80% AQM and the 70% AQM.

It should be noted that the final order of the trust ratings corresponds to that of the AQM accuracies. That is, the 100% AQM stabilises at the highest trust level, followed by the 90% AQM, 80% AQM, and 70% AQM, in this order. This finding supports our H1 hypothesis that learned trust would stabilise over time to a level correlated with the systems' accuracy. However, we will later examine what role individual differences may play in this process.

In addition, since we only selected a small set of discrete accuracies for our AQMs, it can be interesting to analyse our results from the perspective of a rank-ordering problem. Indeed, this would provide an indication of whether the reported trust ranking align to such discrete accuracy levels. A Friedman's test shows significant differences between the trust levels (Friedman's $\chi^2(20,3) = 45.31, p < 0.001$), with mean ranks of 3.8, 2.9, 2.0 and 1.3 for AQMs of accuracy 100%, 90%, 80% and 70% respectively. These statistics suggest that trust ratings correlate with increased

levels of AQM accuracy, when considered as discrete values (here 10% increments), again supporting our H1 hypothesis.

4.2 Acceptable Accuracy and Reliance

We now examine the dynamics of reliance, which we regard as an objective measure of trust. Recall that reliance is measured implicitly during each trial, as described earlier. Again, we apply a simple low-pass filter, but this time we use a 10-trial sliding window, reducing our data to 20 points per AQM. The reason for this larger window is mainly because reliance is a binary feature (at every trial the participant either did or did not follow the system suggestion). Hence, local variations tend to add weight to the reading for a small window size. Fig. 5 shows the aggregated reliance for all the 21 participants and all four AQMs.

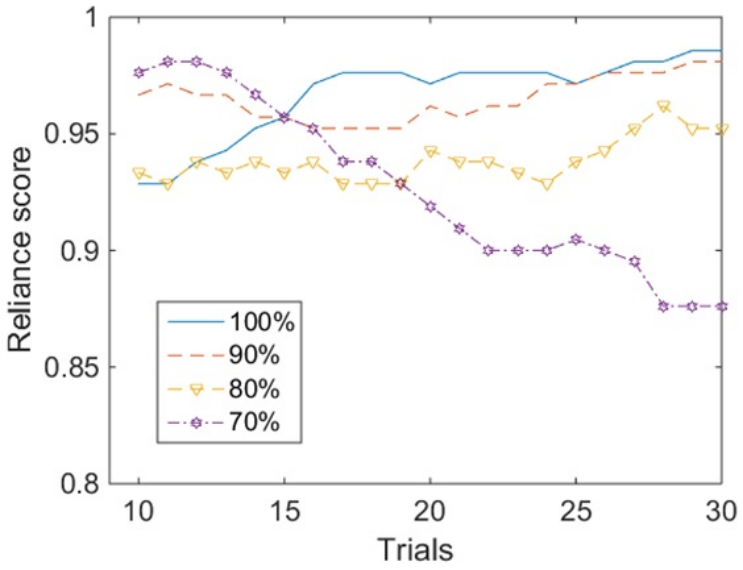

Fig. 5 Mean reliance for all participants, all AQMs.

We observe that despite the larger sliding window of 10 interactions, the reliance curves are less stable than the trust curves. We believe that the reason for this observation is two-fold. Firstly, the effect of a binary feature on smoothing is strong and could require a wider sliding window size, but this would mean losing temporal accuracy in our analysis of reliance dynamics. Secondly, we think that while participants exhibit relatively uniform trust trends, they have different strategies to

deal with it, as per our H3 hypothesis of individual differences. We will explore this aspect in the next sections.

Qualitatively, the AQMs exhibit different reliance patterns. These differences are not linked to the order in which AQMs are presented to the participant, because it is randomised. It is possible that, by way of randomisation, the AQMs in each subset behaved similarly over the first few trials, which would then be picked up as different levels by the sliding window. However, this explanation seems unlikely given the number of participants.

All curves, except for the 70% AQM, demonstrate slight (and often unstable) increases and their final levels are in the range of 0.95-0.98. The 100% and 90% AQMs seem to converge strongly, while the 70% AQM exhibits a steady decline in reliance. The 80% AQM seems close to the 100% AQM baseline. This could indicate that the acceptable level of accuracy for a system is around 80%, possibly a bit above since the AQM 80% is slightly lower. An analysis of variance showed that the effect of the AQM accuracy on the first reliance point was not significant for all participants, $F(3, 80)=1.597$, $p=0.197$ n.s. That is, the apparent reliance pairs observed are not significant in view of the variance, further demonstrated by Fig. 6. This means that the participants interacted with all four AQMs with a comparable level of dispositional trust, as comes through the implicit reliance measure.

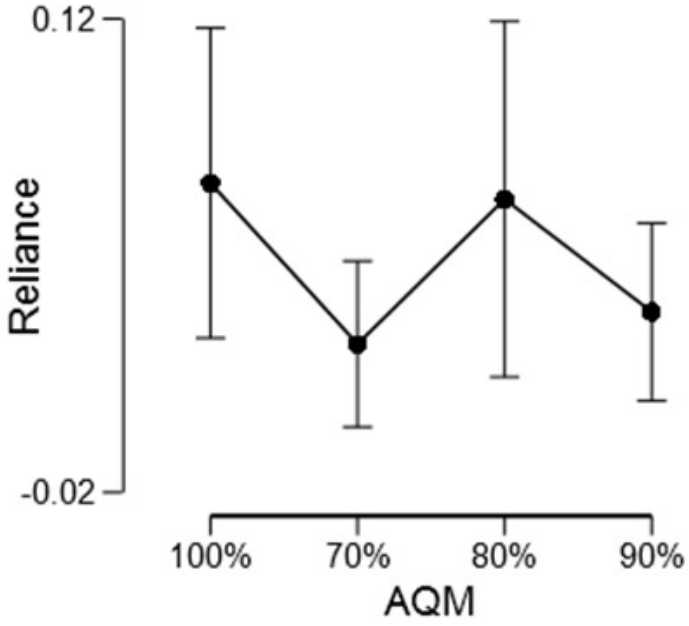

Fig. 6 First reliance point variance for all participants.

Focussing on the last reliance observed after 30 trials, an analysis of variance showed that the effect of the AQM accuracy on the last reliance point was significant for all participants, $F(3, 80)=4.182$, $p=0.008$. The test of homogeneity (Levene's) was significant, but again the sample sizes are equal. Due to the binary notion of reliance, we can test our hypothesis of acceptable level of accuracy by comparing all the AQMs to the 100% AQM baseline, in order to determine where the threshold for accuracy may lay. To do so, we applied a simple contrast in the ANOVA for the last reliance point, and obtained significance only for the pair 100% AQM versus 70% AQM. This means that the 80% AQM, while being visually apart from the 100% and 90% AQMs, is actually not significantly different. However, the AQM 70% is significantly different from the other three AQMs. These results support our hypothesis H2 that users have thresholds of acceptable accuracy for a system, under which the reliance drops. Since there is no significant difference between the AQMs in terms of the initial reliance levels, participants start interacting with the AQMs free of pre-disposition. But later on we observe a specific behaviour only for the 70% AQM, whereby the reliance of the participants on that AQM declines significantly compared to other AQMs. This indicates that a threshold of acceptable accuracy in our AQM for the cohort of our participants lies somewhere between 70 and 80%. Having said that, the high values and narrow range of reliance values should be highlighted. Over the course of the whole experiment, reliance curves of all the four AQMs remain fairly compact and above the 0.9 mark. This behaviour is not surprising, however, and can be explained by the relatively high accuracies chosen for all the AQMs. Even the poorest AQM operating at 70% accuracy can correctly classify a glass 7 times out of 10, which is well above chance. We believe that the participants rightfully perceived this benefit of the AQM over pure random choice. Hence they decided to follow the AQM's suggestions, leading to very high levels of reliance. However, examining individual differences and grouping users can help understand the substantial reliance drop observed for the 70% AQM.

4.3 Clustering of Participants

While individual user profiles can be appealing for high-precision applications, it may not be justifiable in the context of trust, which as a construct has broadly defined metrics. In addition, the number of participants in our experiment would not allow to generate fine grain profiles, if they were to exist. So, we endeavour to partition the participants into two groups using clustering.

The participants were clustered using the reported trust for the last five trials of each AQM, with a K-means method. The trust ratings for the last five trials of each AQM were used because most participants approached stable trust during these trials. We clustered the participants into two groups due to the limited number of participants involved in this study, and end up with Group 1 including 13 participants and Group 2 with the remaining 8 participants.

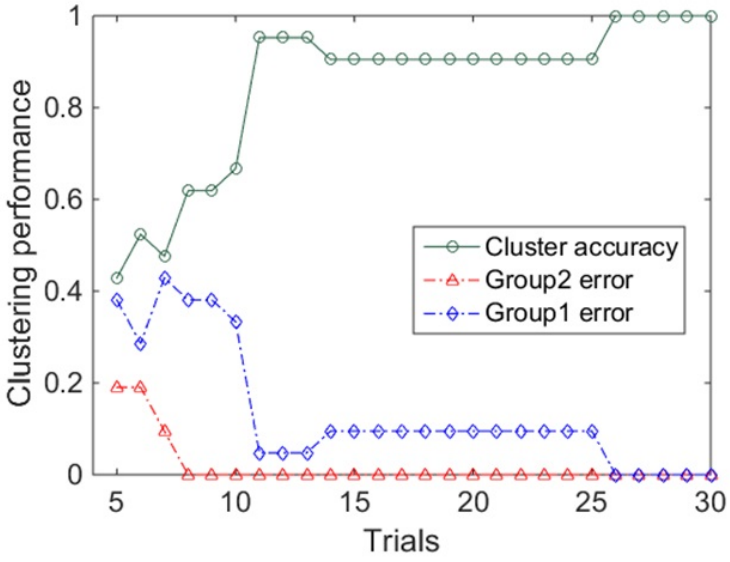

Fig. 7 Clustering performance. The Group2 error indicates the participants that belong to Group 1 but are incorrectly clustered to Group 2, and the Group1 error indicates the participants that belong to Group 2 but are incorrectly clustered to Group 1. The clustering accuracy indicates the overall rate of participants correctly assigned to the final two groups.

Initially, we set out to examine the stability of the clusters. For this, we consider the final clustering produced after the 30 trials as the ground truth and evaluate the relative accuracy of the clusters as they could have been generated at earlier stages of interaction. That is, we execute the above clustering method after a smaller number of trials, say 20, and measure the proportion of users that are correctly mapped to their ground truth cluster. The results of this analysis are shown in Fig. 7. Since the clustering is based on the trust levels calculated with a 5-trial sliding window, no clustering can be done for the first 5 trials. The relative accuracy of the clustering increases between trials 5 and 11, as more user information becomes available, and stabilises thereafter above the 0.9 mark. That is, the clusters become stable after 11 trials, after which less than 10% of users are incorrectly mapped to the other cluster.

The curves marked Group1 or Group2 error provide details about users mapped to the incorrect cluster. We observe that the majority of these come from Group 2 users mistakenly mapped to Group 1. Beyond several initial incorrect mappings, Group 1 users were reliably identified and mapped to the right cluster.

Having clustered the participants, we repeat the above analyses of trust and reliance dynamics, but this time for each group separately. The trust curves for the four AQMs observed for Group 1 and Group 2 are shown in Fig. 8 and Fig. 9 respectively. Since clustering was based on trust levels, we expect to find differences in trust between the two groups. Notably, the curves for the 100% AQM are similar in both groups, which can be expected based on H2, since a 100% accuracy AQM

is very likely to be acceptable to all users, regardless of their sensitivity. Therefore, we focus the rest of the analysis on differences between groups with regards to the other three AQMs.

Qualitatively, the Group 1 curves are much more spread out than those of Group 2. The initial trust levels of the 90% AQM, 80% AQM, and 70% AQM in Group 1 are in the range of 0.5 to 0.63, whereas in Group 2 they are in the range of 0.55 to 0.77. Despite this, the range of final trust is fairly different: it ranges 0.08 to 0.73 for Group 1 versus 0.70 to 0.94 for Group 2. It should also be highlighted that the three trust curves are clearly separable for Group 1, while the differences are less pronounced for Group 2. Also note that the order of the curves for Group 2 does not correspond the accuracy levels of the AQMs.

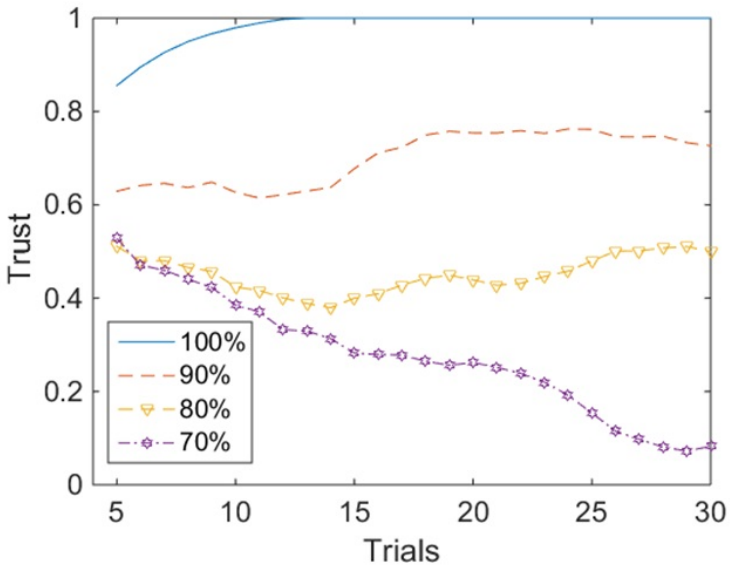

Fig. 8 Mean trust for Group 1, all AQMs.

An analysis of variance showed that the effect of the AQM accuracy on the first trust point was significant for Group 1, $F(3,48) = 7.267, p < 0.001$. Post hoc Tukey tests have identified the significant difference between Group 1's trust on the 100% AQM and all the remaining AQMs. For Group 2, no significant difference has been found for the first trust point $F(3,28)$=0.820, p=0.494. The test of homogeneity (Levene's) for the two groups' first reliance point was not significant.

Examining the last trust point now, an analysis of variance showed that the effect of the AQM accuracy on the last trust point was significant for Group 1, $F(3,48) = 48.51, p < 0.001$ and also for Group 2, $F(3,28) = 7.510, p < 0.001$. In terms of pairwise post hoc comparison, Group 1 showed significantly different trust for any pair of AQMs, while for Group 2, significant difference was observed between three

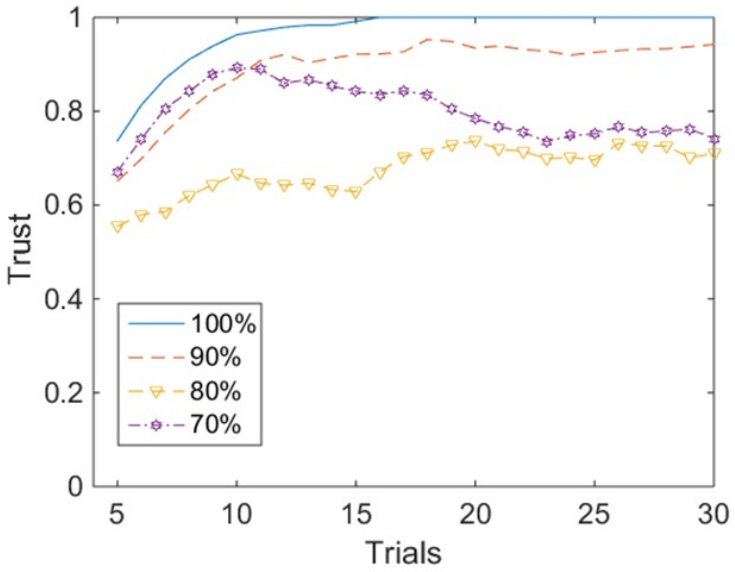

Fig. 9 Mean trust for Group 2, all AQMs.

pairs of AQMs, i.e. 100% and 80%, 100% and 70%, 90% and 80%. The test of homogeneity (Levene's) for both groups last reliance point was significant, but can be ignored because of the equal sample sizes.

In order to address hypotheses H2 and H3 of acceptable levels of accuracy, it is needed to compare all the AQMs to the baseline 100% AQM. To do so, we applied a simple contrast in the ANOVA for the last trust point for Group 1, and obtained significance only for the pair 100% AQM versus 70% AQM. This finding means that the 80% AQM, while being slightly apart is not significantly different. However, the 70% AQM was indeed found to be significantly different from the other AQMs, and we argue that this AQM falls below the threshold of acceptable accuracy postulated in our hypothesis.

These results support the hypothesis H3 that individual differences exist for acceptable accuracy, but typical user groups may be constructed, where Group 1 demonstrates significant difference in terms of trust on different AQMs, however Group 2 doesn't show significant trust difference.

As observed earlier, the differences between the AQMs in terms of reliance level are less clear than the differences in trust. The reliance curves of the four AQMs observed for Group 1 and Group 2 are shown in Fig. 10 and Fig. 11 respectively.

Qualitatively, the two groups exhibit distinct patterns. Group 1 starts with a fairly uniform level of reliance across the AQMs, which can be expected, since the accuracy of each AQM is not known to the participants at the start. Conversely, the reliance in Group 2 is initially split into two ranges: the 90% and 70% AQMs hover above the 0.95 mark, while the 100% and 80% are much closer to 0.9. This result is in

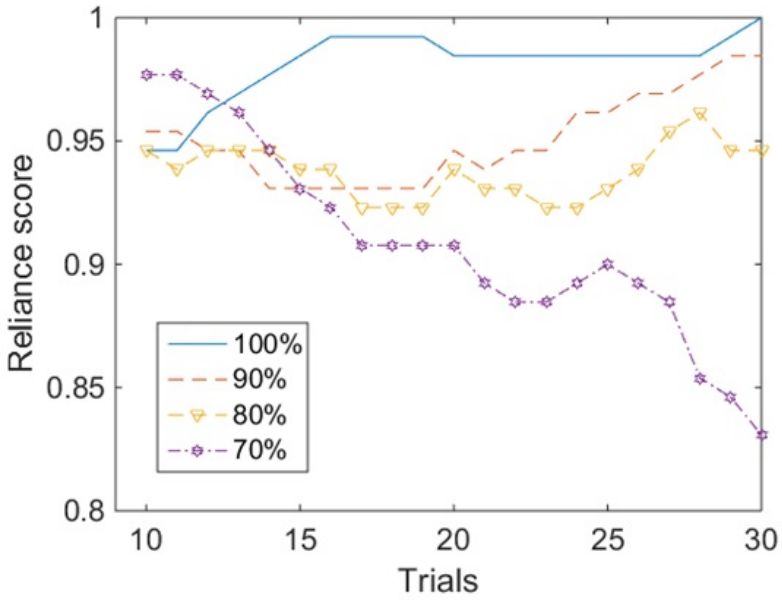

Fig. 10 Mean reliance for Group 1, all AQMs.

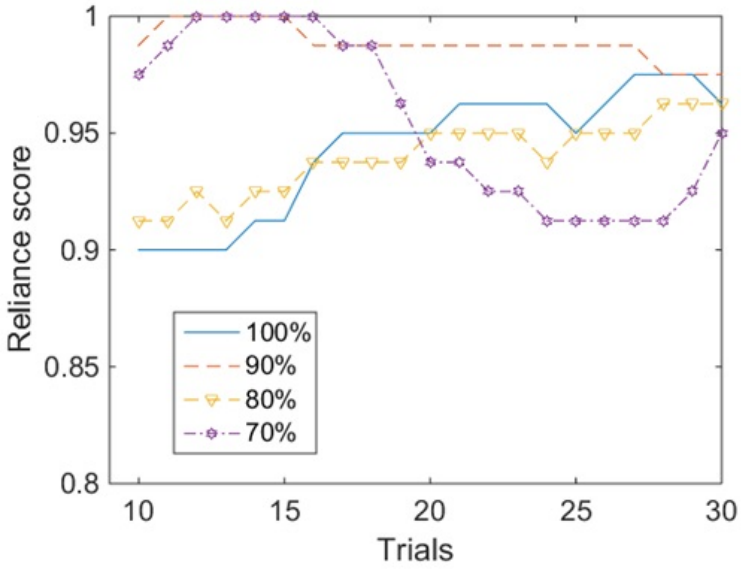

Fig. 11 Mean reliance for Group 2, all AQMs.

line with our earlier observations of reliance using all the participants. An analysis of variance showed that the effect of the AQM accuracy on the first reliance point was not significant for Group 1, $F(3, 48)=0.441$, $p=0.725$ n.s. and also not significant for Group 2, $F(3, 28)=1.584$, $p=0.215$ n.s. This means that the apparent subsets observed for Group 2 are not significant in view of the variance, as demonstrated by Fig. 12.

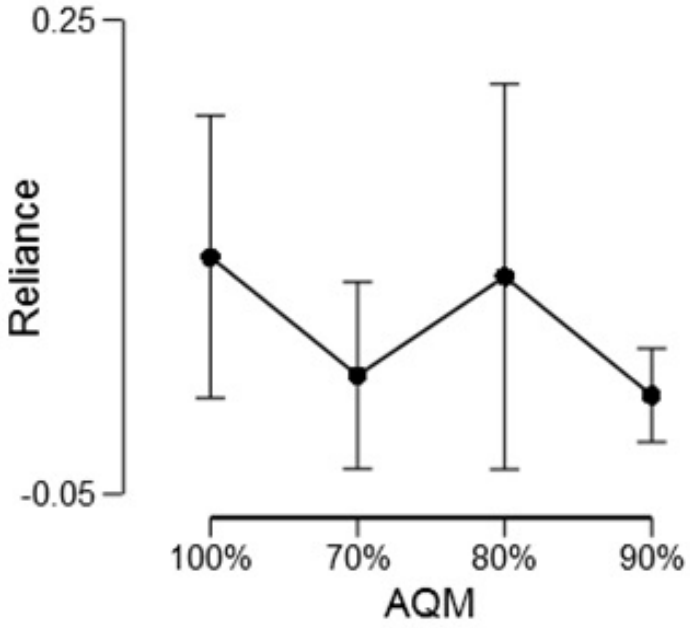

Fig. 12 First reliance point variance for Group 2.

Examining the last reliance point, we observe that the reliance levels in Group 1 correctly reflect the order of the AQM accuracies. Also note that the curves of the 100% AQM, 90% AQM, and 80% AQM obtain high reliance scores of almost 0.95 or greater than this, while the 70% AQM is clearly placed below the others. This indicates that in Group 1 the acceptable level of accuracy for a system is around 80% (possibly a bit above that), since the reliance on the 80% AQM is slightly lower than on the 100% AQM and 90% AQM. For Group 2, all the curves converge around the 0.95 mark, although the 70% AQM is slightly lower than the others. An analysis of variance showed that the effect of the AQM accuracy on the last reliance point was significant for Group 1, $F(3, 48)=4.532$, $p=0.007$ and not significant for Group 2, $F(3, 28)=0.153$, $p=0.927$, n.s. The test of homogeneity (Levene's) for the last reliance point in Group 1 was significant, but the variances were equal for all the other analyses above.

Again, we applied a simple contrast in the ANOVA for the last reliance point for Group 1, and obtained significance only for the pair 100% AQM versus 70% AQM. Just like for trust in Group 1, this means that the 80% AQM while being slightly apart is not significantly different from the 100% and 90% AQMs. However, the 70% AQM is significantly different from the other levels, arguably because it is under the threshold of acceptable accuracy postulated in our hypothesis. Similarly to subjective trust, these results support our hypothesis H3 that Individual differences exist for acceptable accuracy, but typical user profiles may be constructed. Both groups start interacting with the AQMs free of pre-disposition, but Group 1 later on exhibits a threshold of acceptable accuracy in the range of 70% to 80%. Group 2 again seems to be more resilient or have a lower threshold of acceptable accuracy.

Seen from a rational behaviour perspective, the behaviour of Group 1 is not optimal, since adhering to the AQM's recommendations would provide a 20% better than chance outcome. We conjecture that one possible explanation for this may be that the AQM accuracy perceived by the participants is lower than the actual AQM accuracy, i.e., participants in Group 1 may perceive the 70% AQM as being worse than chance. It should be noted, however, that our experiment does not allow us establish the exact cause of the observed behaviour, as this would require a substantially different experimental set-up.

5 Discussion

In this chapter we discussed the fine-gained dynamics of user-system trust, an important construct of human interaction with a decision support system. We specifically focused on an automated quality monitoring (AQM) simulation, which provided indication of faulty glasses being produced. In our study, each user interacted with four AQMs and out of these interactions we populated the explicit trust and implicit reliance scores.

We analysed the temporal dynamics of both trust and reliance, as well as their dependence on the accuracy exhibited by the AQM. It was found that the reported trust levels aggregated across the entire cohort of users, stabilised over time and, at large, corresponded to the accuracy of the AQMs. Somewhat surprisingly, we discovered that the implicit reliance levels were very high and comparable across the four AQMs. We attribute this finding to the relatively high accuracy of the AQMs in our experiment. Following this, we conducted an additional analysis of individual user differences in trust and reliance. For this, we split the users into two clusters and compared the trust and reliance scores obtained in these clusters. This analysis discovered differences in the dynamics of user trust in the two clusters and also some differences in user reliance on the low-accuracy AQMs. Hence, the obtained experimental results support the hypotheses raised at the beginning of this chapter. Firstly, we observe that the learned user-system trust stabilised over time and generally correlated with the level of accuracy exhibited by the system. Secondly, our findings indicate that at reasonably high levels of system accuracy, user reliance is

high, whereas once the system accuracy falls below an acceptance threshold, the reliance may deteriorate as well. Thirdly, we show that these acceptance thresholds are dynamic and user-dependent, and we successfully manage to separate users into two groups with different trust profiles and reliance patterns.

These observations surfaced an important practical question referring to the implications of our work on the sustainability of user-system trust. Due to the low transparency but complex structure of most machine learning systems, users are mostly unable to understand their internal working mechanism or parameters, however there is a possibility to improve the users' performance, based on their interaction history. Once the system recognises that certain users are in the 'risk group' and the performance exhibited by the system is not up to their expectations, additional steps may need to be taken in order to sustain the trust of these users. For instance, the system may show its historical performance to these users, thus increasing the experience of these users, or revealing some details of the internal machine algorithm which allows for further understanding of the users. Alternatively, system designers may want to enrich the interactions of these users, e.g., through additional explanations of the suggested actions or through implanting persuasive messages strengthening user trust [1].

Another intriguing question refers to identifying the users at risk. In our study, we conduct a posterior clustering of the participants and split them into two groups. However, a more relevant task would be to identify the type of a user and their system acceptance at the beginning of interaction or, even, before the interaction. The analysis of our clustering shows that the clusters were stable and little changes of cluster were observed at late stages of interaction, possibly indicating stable acceptance preferences. One possible predictor of this preference could potentially be the user's personality or behavioural traits, which can be derived, for example, from the user's past interactions with other systems. Prior research shows that trust correlates to some personality characteristics [12], and this information can be extracted and leveraged in order to sustain user trust. We leave this research beyond the scope of our work.

In addition, we should put upfront that our findings are based on a fairly limited cohort of participants, all of which had reasonably short interactions with the system. Validating our findings with a larger set of users (and, possibly, with a different target system) is a natural future extension of our work. Also, we would like to increase the length of interactions on the account of reducing the frequency of users reporting their explicit trust. For example, we could collect the explicit trust level every second interaction, allowing us to double the length of interactions without over-burdening the users. This would allow us to collect a more solid empirical evidence and better support our hypotheses.

Finally, more work is needed to address the fine-grained dynamics of trust acquisition and extinction. In our work, we assumed a stable level of accuracy of every system. This, however, may vary over the course of user interaction. Hence, it is important to validate the evolution of user trust as a function of the user's initial trust disposition, observed system performance, and temporal aspects of this performance

(e.g., initial failures vs. failures when the trust was already formed). We highlight the importance of these research questions, but leave this work for the future.

6 Conclusion

This chapter examines the relationship between system performance, a users trust and reliance on the system. We observe that users correctly perceive the accuracy of the system and adjust their trust and reliance accordingly. We have successfully segmented the users into two groups who showed different patterns in trust dynamics and reliance with different AQM systems. This important learning can be leveraged by designers of practical systems for group-focused interaction systems. Furthermore, we have established a possible threshold in the acceptance of the system. These findings taken together, have dramatic implications for general system design and implementation, by predicting how trust and reliance change as human-machine interaction occurs, as well as providing new knowledge regarding system performance that is necessary for maintaining a user's trust.

References

1. Shlomo Berkovsky, Jill Freyne, and Harri Oinas-Kukkonen. Influencing individually: fusing personalization and persuasion. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):9, 2012.
2. Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
3. Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94, 2002.
4. P Christopher Earley. Computer-generated performance feedback in the magazine-subscription industry. *Organizational Behavior and Human Decision Processes*, 41(1):50–64, 1988.
5. Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.
6. Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.
7. John D Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40(1):153–184, 1994.
8. John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.
9. Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.
10. Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

11. John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4):656–665, 2006.
12. Stephanie M Merritt and Daniel R Ilgen. Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, 50(2):194–210, 2008.
13. Bonnie M Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.
14. Julian B Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 35(4):651–665, 1967.
15. Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.
16. Cuthbert L Scott III. Interpersonal trust: A comparison of attitudinal and situational factors. *Human Relations*, 33(11):805–812, 1980.
17. Indramani L Singh, Robert Molloy, and Raja Parasuraman. Automation-induced" complacency": Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2):111–122, 1993.
18. Weiquan Wang and Izak Benbasat. Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, 24(4):249–273, 2008.
19. Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 223–227. ACM, 2016.
20. Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317. ACM, 2017.
21. Jianlong Zhou, Zhidong Li, Yang Wang, and Fang Chen. Transparent machine learning—revealing internal states of machine learning. In *Proceedings of IUI2013 Workshop on Interactive Machine Learning*, pages 1–3, 2013.