User Trust Dynamics: An Investigation Driven by Differences in System Performance

Kun Yu Shlomo Berkovsky Ronnie Taib
Dan Conway Jianlong Zhou Fang Chen
Data61, CSIRO
13 Garden Street, Eveleigh, NSW 2015, Australia
{first.last}@data61.csiro.au

ABSTRACT

Trust is a key factor affecting the way people rely on automated systems. On the other hand, system performance has comprehensive implications on a user's trust variations. This paper examines systems of varied levels of accuracy, in order to reveal the relationship between system performance, a user's trust and reliance on the system. In particular, it is identified that system failures have a stronger effect on trust than system successes. We also describe how patterns of trust change according to a number of consecutive system failures or successes. Importantly, we show that increasing user familiarity with the system decreases the rate of trust change, which provides new insights on the development of user trust. Finally, our analysis established a correlation between a user's reliance on a system and their trust level. Combining all these findings can have important implications in general system design and implementation, by predicting how trust builds and when it stabilizes, as well as allowing for indirectly reading a user's trust in real time based on system reliance.

Author Keywords

Trust dynamics; system performance; acquisition and extinction; temporal examination; reliance

ACM Classification Keywords

H.1.2 [**Models and Principles**]: User/Machine Systems – *human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *user-centered design.*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

IUI 2017, March 13-16, 2017, Limassol, Cyprus
© 2017 ACM. ISBN 978-1-4503-4348-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3025171.3025219

INTRODUCTION

We live in an era booming with intelligent systems and automated devices. They were designed to help us whenever and wherever possible. Their capabilities and availabilities have been expanding relentlessly, with a speed that almost exceeds people's imagination and hence results in a trust crisis [2,17]: how can we know if a system can manage a job, and how much trust should we have in it? Will the system make a mistake in the future, and shall I still rely on it afterwards? Actually all these critical questions center around one important mental construct of humans, *trust*.

Trust roots in human mind and is believed to be related to a number of factors including human's disposition and experience [21], and hence is dynamic and difficult to capture. Human-machine (or human-system) trust plays a key role in affecting the way people work with intelligent systems: proper trust posited by a human is beneficial to the human-system collaboration, saving human effort and improving collaborative performance, while improper trust, e.g. a user trusts a system more than warranted or distrusts a reliable system, may lead to inappropriate system use or even task failure [13,20]. People adjust their trust in the system based on their interaction, during which system performance is posited as being among the most critical factors affecting a user's trust [13,15,16,19].

However, most existing work only reported the implication of varied system accuracy on the overall trust of users, while the examination of the temporal changes of trust is deficient. In order to shed more light on this area, our work explores how users' trust changes dynamically as time elapses when working with an automated system. Specifically, our efforts focus on how system failures and system successes occurring at different time affect a user's trust, or in other words, we study the *acquisition* and *extinction* of user trust due to varied system performances. We also propose that a reliance score can be used to unobtrusively quantify users' trust dynamically.

Via investigating the way user interact with four simulated Automatic Quality Monitor (AQM) systems, we have identified that a user's trust is affected to a higher level by system failures rather than system successes. Furthermore, at different time of interaction system failures and successes demonstrate different implications on trust change, which ultimately cause users' trust to converge. Finally, trust correlates with the reliance of users on the system, which implies that the level of user trust in a system can be inferred from their behaviors.

The remainder of this paper is organized as follows: First we discuss the existing work related to trust definition, composition, investigation context and method. This is followed by the methodology section which includes the description of our experimental design, procedure and introduction of the data we have collected. In the results section, our findings are illustrated, which reveal the patterns of users trust change over time. We explain our findings and discuss them from the user interaction design perspective in the discussion section, before concluding with some remarks about the implications of our work.

RELATED WORK

The research in human-system trust, or human-machine trust and its implications for system design has a rich history [8,23,31]. From Rouse's [25] and Glass's [7] ideas about adaptive aiding, to the later evolution into more refined HCI techniques, new methods to use trust for HCI system design have never ceased to be devised [10,28].

Various definitions have been proposed to represent user trust in human-machine interactions. One of the most widely deployed definitions is from Lee and See [14] where 'trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.' This definition succinctly encapsulates the primary sources of variance (the user, system, and context) and identifies a key aspect of this relationship – that of vulnerability. This definition is in line with our work, and thus is adopted in this paper. Other trust definitions show that trust is a hypothesised variable that has been demonstrated to be a key mitigating factor in system use or disuse (sometimes called reliance) [13], can be a means to characterize the interpersonal dependency [5] or the relationship within or between organizations [18].

Many researchers have also posited that trust is a multi-dimensional attribute [1,3,6]. Hoff and Bashir [9] proposed three layers of variability in human-automation trust: dispositional trust, situational trust and learned 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, a user's workload, perceived risks and benefits, and even mood. *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 affected by *dynamic learned trust* which evolves 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 behavioural sciecnces. However, it is not clear how trust changes specifically with time when a human is working with an automated system, and this work casts additional light on these issues.

Human-machine trust research has been refined through a myriad of different investigative lenses and contexts. Moray et al. [19] investigated adaptive industrial automation systems and found that reliability of automated fault diagnosis, mode of fault management (manual vs. automated), and fault dynamics strongly affect variables such as subjective trust in the system and operator self-confidence. Schaefer and Scribner [27] presented a theoretical review on the changing dynamic of the human-vehicle system. It was suggested that the construct of trust is a viable and important addition into performance models to better understand the complex dynamic of evolving automotive systems. Another work conducted on in-vehicle trust by Verberne etc. [29] has revealed that whether an intelligent system shares the same goal as the user, and the level of system transparency, i.e. the explanatory information provided by the system impacts the trust of the user, while the importance of explanations in trust building has also been confirmed in the work of Pu and Chen [24].

System failures have always been a key issue in the research of a user's dynamic trust changes, which may determine the way people behave and affect their reliance on automation [22]. Lee and Moray [13] used a simulated pasteurization system to induce consecutive system failures, and proposed that trust in a machine is related to overall human-machine joint performance, the system's fault and the user's prior trust. Johnson [11] examined how different types of automation errors affects user trust and reliance as well as the perceived reliability of automated decision aids. He found that perceived reliability is often lower than actual system reliability and that false alarms significantly reduced user trust in the automation. In comparison, Sauer et al. [26] investigated the effects of automation failures in training on trust. The results showed that if users are trained on miss-prone automation systems, automation bias (a tendency to follow the advice of the automation) was high when they encounter a failure in a different system, and that user errors resulting from automation bias were much higher when automation misdiagnosed a fault than when it missed one. It was suggested that trust remained stable over time in the absence of changes in reliability levels. However, when users were exposed to automation failures, their trust levels decreased rapidly [30].

The above research has shown that system failures cause significant declines of a user's trust. However, to the best of our knowledge, they have not been able to identify how a user's trust may be affected by system failures over time, and how consecutive occurrences of system failures or successes may change the user's trust. As a consequence, this paper addresses these very issues of how trust is affected over time by the observable system performance.

METHODOLOGY

Every human interaction with a system has unique characteristics and contexts making trust dependent on a broad range of human, technological and environmental factors. In order to provide results that can generalize to real-life system, we operationalize decision making as binary tasks, due to the fact that many complex decision processes can be decomposed into a series of atomic binary decisions. From the system design perspective, whether a complex system works can be verified via the examination of a simplified system [4]. The decision-trust relationship thus can be easily generalised to complicated decision-making problems. Furthermore, the simplified decision making protocol we implement is similar in effect to the microworlds introduced by Lee and See [14], which makes it convenient to map trust levels to decisions without the interference of other parameters.

Scenario

This experiment simulated a quality control task in a factory that manufactures drinking glasses [32]. 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 was designed to not necessarily be correct and occasionally exhibited false positives (suggesting failing 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*.

Trials

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 illuminated red for a faulty glass or turned off for a good glass (Figure 1). The user then needed to click a *Pass* button, if they thought the glass was good, or to click *Examine* if they thought the glass might be faulty. It is important to note that this decision is entirely up to the user who may opt either to comply with the AQM's recommendation or override it.

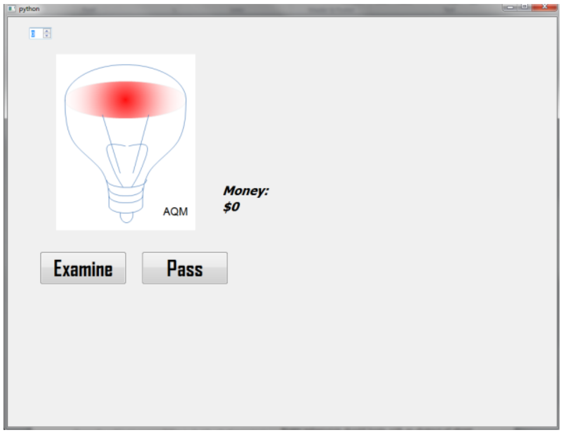
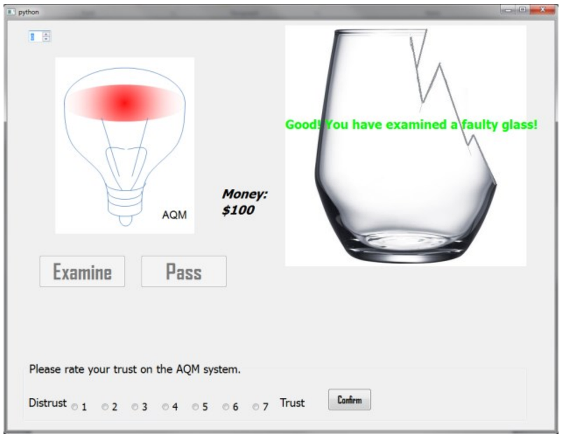


Figure 1: The trial starts with an AQM recommendation

After they made their choice, the users were shown the actual condition of the glass, providing them with direct feedback on their decision, as illustrated in Figure 2, 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 (examine faulty glass, or pass 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.



Good! You have examined a faulty glass!

Please rate your trust on the AQM system.

Distrust ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 Trust Confirm

Figure 2: Upon decision from the user, the actual glass is shown and fictitious earnings updated.

At the end of each trial the users were requested to explicitly 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.

AQM Accuracy and Blocks

The trials were presented sequentially, 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 experiment session was divided into four blocks of 30 trials each. The users were told that a different AQM would be used for each block; indeed, each AQM's accuracy was manipulated by varying the average rate of false positives and false negatives, as shown in Table 1.

Table 1: AQM accuracies

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

In order to capture a trust baseline for each user, the experiment session systematically started with the 100% accuracy AQM, followed by the other three AQMs presented in a randomized order. Within a block, system failures occurred randomly over the 30 trials, so that the mean accuracy over the block matched the predetermined AQM accuracy as listed in Table 1. For instance, the 80% AQM made on average 6 errors over the 30 trials (on average: 3 false positives and 3 false negatives).

Participants

Twenty-one participant took part in this 45 minute experiment as users of the AQMs. Most participants were university students and the rest IT professionals. No specific background knowledge or inclusion criteria were required to participate in the experiment. Recruitment and participation were conducted in accordance with a University-approved ethics requirements for this study. Snacks were offered in return for taking part in the experiment.

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);
- Subjective trust rating.

We derived the following variables for each trial:

- Reliance: whether the user followed the recommendation from the AQM or not;
- Trust: normalized subjective trust rating. For each subject, all their inputs across all blocks are used to normalize their ratings in the [0, 1] range. More specifically, for all the trust ratings of a user, the normalized trust value T_i is calculated as

$$T_i = \frac{T_{io} - T_{\min}}{T_{\max} - T_{\min}} \tag{1}$$

Where T_{io} is the original trust rating of the user for the current i-th trial, and T_{max} and T_{min} are the maximum and minimum trust ratings respectively given by the same user across all four AQMs.

We also derive the following variables for sets of trials or blocks:

- AQM accuracy: as set for each block;
- Number of system failures: number of *consecutive* incorrect predictions made by the AQM;
- Number of system successes: number of *consecutive* correct predictions made by the AQM;
- Reliance score: the mean compliance over a set number of consecutive trials, in the [0, 1] range;
- Trust change: the variation in trust between the current trial and the previous trial or several trials before (may be positive or negative).

RESULTS

In the following sections, we examine the data collected during the study across all the AQMs.

Trust Correlation to System Accuracy

The acquisition and extinction of trust can be observed over the course of user interactions with the AQMs. Since the AQM errors were randomized over the 30 trials for each AQM, trust variations for each specific trial exhibited local variations when averaged between users. This issue is addressed by applying a 5-trial sliding window low-pass filter, i.e.

$$T_{mi} = \frac{1}{N} \sum_{i-N+1}^{i} T_i \tag{2}$$

Where T_{mi} is the filtered trust value, T_i is the trust value for the i-th trial, and N (equal to 5 in this case) is the size of the sliding window. Figure 3 shows the aggregated normalized trust for all 21 users, for all four AQMs. The horizontal axis represents the 30 trials in each block. Note that due to the 5-trial sliding window, normalized trust cannot be computed for the first 5 trials.

During early trials, trust in all AQMs seems to be close to uniform as would be expected since users know that each new AQM is different from the others they may have encountered, and the order is randomized. That said, trust in the 100% AQM appears to be above the other AQMs, but it is expected since it is consistently the first AQM that the users worked with.

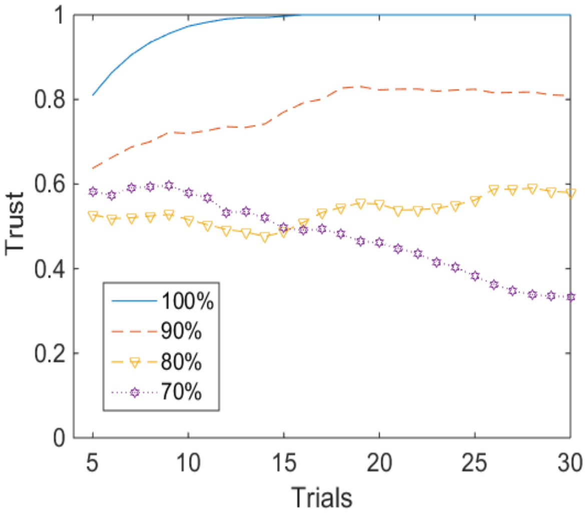


Figure 3. Mean trust for all users for each AQM.

For the 90%, 80% and 70% accurate AQMs, the users' initial trust is comparable, around 0.6, showing the initial disposition of users towards the AQMs. Investigating the temporal fluctuations of the trust values, we observe that they stabilize towards the end of the 30 trials, although the trust in the 100% AQM stabilizes at 1 after 13 trials only. It appears that the trust in the 90% and 80% AQMs stabilizes after 20 to 25 trials, while the trust in the 70% AQM stabilizes after trial 26.

An analysis of variance shows that the effect of AQM accuracy on the last trust point (trust mean over the last five trials) is significant for all the users ($F(3, 80)=27.03$, $p<0.05$). A further examination suggests that the average trust score on the last five trials and the accuracies of the AQMs have a strong correlation ($r=0.9996$, $p<0.05$). This finding implies that the users formed a stable mental model of the system trust levels towards the end of the 30 trial session and that this subjective perception correlates with the actual level of accuracy exhibited by the AQMs.

Effects of system performance on trust change

We observed that the trust levels exhibited some oscillations before stabilizing after about 25 trials. However, the cause for the trust change may still be undetermined. As discussed in Lee and Moray's work [13], system performance, especially system failures, can be one important factor that affects a user's trust. As a consequence, we examined the effects of system failures and successes on a user's trust change. In our study, a system failure refers to the case when the AQM recommends a wrong answer to the user, for example, the light is on but actually the glass is good, or vice versa. On the contrary, a system success means the AQM makes a correct recommendation consistent with the quality of the glass. The pattern by which the system fails or performs well is central to our investigation, so we coded different number of consecutive system successes and failures into our analysis.

It should be noted that some collected data were intentionally discarded from the analysis to avoid the ceiling effect when using a Likert scale. For example, if the system succeeded three times in a row, resulting in a user's trust ratings of 6, 7, 7 respectively on a 7-point Likert scale, then the user may have arguably reached the scale's ceiling at the second success and been unable to rank the third trial as high as she/he would have wanted, i.e. above 7. In such instances, we excluded the third point from our analysis and considered only the first increase of trust. We applied a similar filtering to the lower end of the scale to avoid the floor effect, ignoring instances such as 2, 1, 1, for example. It should be noted that very few instances of the floor effect were observed in the data.

Figure 4 illustrates the average implication of consecutive system failures and system successes for trust change for all users. Here the trust change d_T between the k-th trial and $(k+n)$-th trial is calculated as:

$$d_T = T_{k+n} - T_k \tag{3}$$

Where T_{k+n} refers to the user reported subjective trust on trial $k+n$, and T_k is the user reported subjective trust on trial k. In the following part of the paper trust change is calculated in the same way. The number of consecutive system failures or successes n is plotted along the horizontal axis.

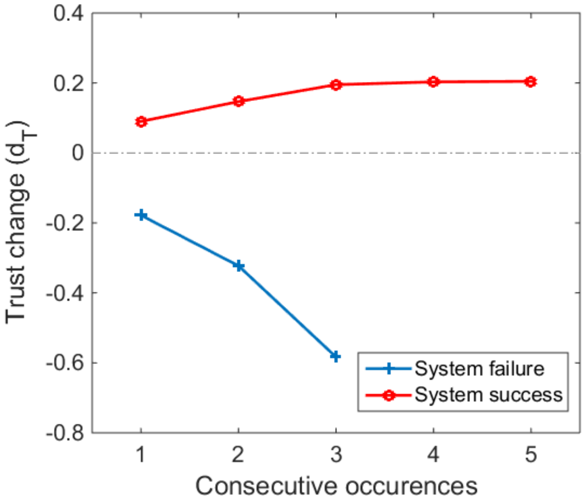


Figure 4. Implications of system failures and successes for user trust change.

Due to the fact that all AQMs used in the experiment perform better than the chance rate of 50% for the binary decision involved, more system successes than system failures are observed, and hence we are unable to extend the system failure graph beyond three consecutive failures.

The overall pattern is clearly visible in Figure 4: system failures cause trust to decrease, while the system's successes increase a user's trust. A higher number of consecutive failures and successes generally causes a greater change in trust values. Examining the trust change caused by system successes, there is a significant difference between one and two correct AQM recommendations (t=3.56, p<0.05), however no significant trust changes have been identified when more than two system successes occur. In fact, the mean trust increase for four consecutive system successes is 0.203, which is quite similar to that of five consecutive system successes at 0.205. This may imply that consecutive successes involving more than five trials do not result in any further significant trust increase. On the other hand, when examining the negative part referring to system failures, we observe a steady decrease in the trust changes for additional consecutive failures. Furthermore, a significant difference exists between a single failure and two consecutive failures (t=5.06, p<0.05), or between two and three consecutive failures (t=3.14, p<0.05). Thus it becomes clear that as more system successes occur in a row, trust change is affected less and less; however in contrast, system failures tend to cause an increasing trend to suppress trust. In this case, the mean values of trust change are 0.17, 0.32 and 0.58 respectively, which implies that when several system failures occur consecutively, one more system failure will cause more trust loss than the previous one.

When comparing trust changes caused by the same number of system successes and failures, significantly different trust changes occurred for each pair of comparisons. Specifically, between a single system failure and a single success (t=16.21, p<0.05), between two consecutive system failures and successes (t=17.93, p<0.05), and between three consecutive system failures and three consecutive successes, (t=10.87, p<0.05). This consistent trend implies that the system successes and failures affect trust change in different ways, and that the system failures have a stronger impact in terms of trust change amplitude.

Trust change on a temporal basis

We have identified the effect of system failures and successes on trust change. However, another factor, the user familiarity with the system, and therefore, the amount of evidence taken into consideration for forming the trust opinion, cannot be overlooked when examining factors affecting change in trust. We operationalize this variable as the time the user has already spent interacting with the system in our study, or basically, the number of past trials in each condition.

Our investigation starts with the impact of system failures on trust change, as observed at different interaction time points. Figure 5 illustrates the change of trust over time due to single and double system failures, which has been averaged across all the users and smoothed with a 10-trial sliding window as a means to reduce noise, using a method similar to that presented in formula (2), while keeping the filtered trust value at the center of the window. The 10-trial sliding window is the reason for having 21 trials in total, from trial 5 to 25 in this case.

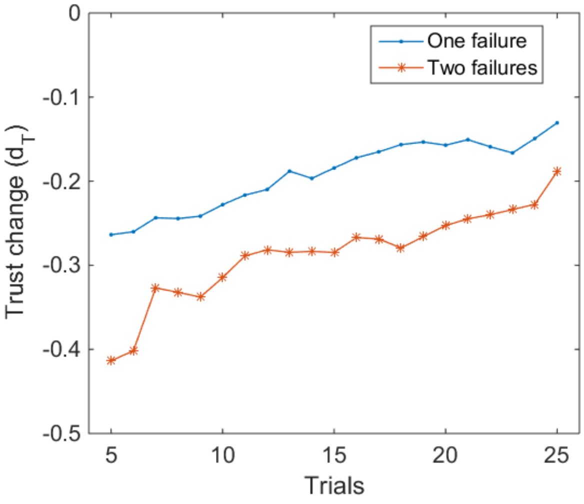


Figure 5. Trust change caused by system failures with time (averaged across all users, processed with a smoothing window of 10 points).

It is interesting that both for single system failure and for double system failures, a consistent trend of decreased impact on trust change can be observed as users conduct more trials. A comparison between the two curves shows that double failures cause a greater decrease in trust than single failures, as suggested above. More importantly though, the observed decrease in the trust change over time implies that as time elapses and the user's experience with the system accumulates, the rate of trust loss caused by system failures steadily decreases. This implies that as time passes, the users seem to form their subjective trust and modify it less and less. At the initial stages of interaction trust changes are higher than at the later stages and we refer to this as the *inertia* of trust, similar to the effect observed by Lee and Moray [12], where the trust level between human and the system is kept stable once it has been formed.

However, examining the system successes' impact on trust change over time reveals different results, as shown in Figure 6. Although the overall trust change trend is close to decreasing, it should be noted that the trust change, during the 30 trials, consists of three distinct phases (separated by the dotted lines). Phase I corresponds to approximately the first nine trials, during which system successes result in increased trust changes. This can be posited as the stage when users form their perception of the system trust: they are learning to adjust their trust, starting from a small change and then increasing the change gradually. Phase II starts from around trial 10 and lasts until trial 18, which comprises of the peak trust change part. In this second phase, the system successes have the highest impact on trust change, and it can be interpreted as a mental process where users are more confident in their understanding of the system and they are making rapid adjustments of their trust level to reflect the system performance. Finally, in Phase III starting from trial 19 approximately, users consistently decrease their trust adjustments, as their trust level already approaches the stable levels and only small fine-tuning is needed. This phase essentially matches the inertia of the trust decrease observed in Figure 5. Combining the right parts of Figure 5 and Figure 6, it can be inferred that user's trust in a system tends to converge at Phase III.

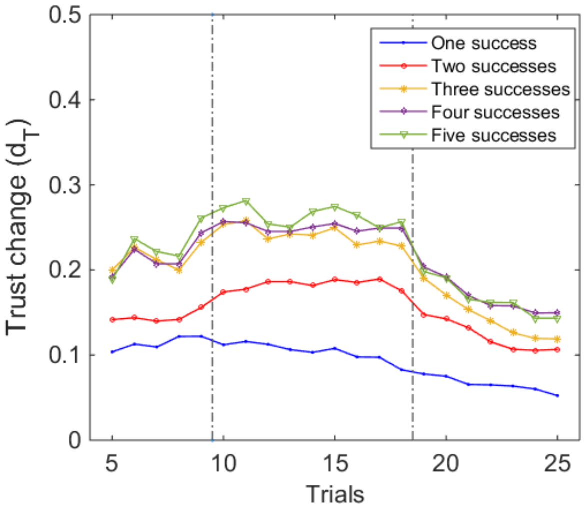


Figure 6. Trust change caused by system successes with time (averaged across all users, processed with a smoothing window of 10 points).

It should be noted that the three phases are only evident for system successes but not for system failures, and significant difference in trust changes can be identified between Phase II and Phase III for a single system success (t=2.51, p<0.05) and double system successes (t=2.71, p<0.05). This finding implies that a user's trust adjustment goes from coarse to fine, and in a simple system like ours this change occurs after around 18 trials. On the other hand, there is insignificant difference in trust changes between Phase I and Phase II, and one possible explanation is that users are making continuous and rapid adjustments on trust in both Phase I and Phase II, between which only minor differences in trust levels exist.

The results shown in Figure 5 and Figure 6 suggest that trust formation processes mainly take place during the initial stages of user interaction with the system as expected. Higher fluctuations, both in the positive and negative direction are observed at the beginning, while the magnitude of changes generally decays towards the later interaction stages. Presumably, this happens due to the fact that users initially form their subjective trust perceptions and only modify them slightly later on due to the observed system performance. In simple words, we found that the users *form their judgements at the beginning* of interaction and they only *adjust it later on* depending on the system performance.

Relation between trust and reliance

In real-life applications, knowing the way that users' trust changes is not enough: it is not practical to keep asking people to report their trust levels during their interaction with a system. As a consequence, methods that can uncover the level of a user's trust in a non-intrusive way can be promising for practical use.

As mentioned earlier, one purpose of our designed experiment is to identify the relationship between a user's trust and their decisions, and their decisions are observable in real systems. Due to the simple design of our experiment, we are able to determine on a per trial basis whether a user is following the recommendation of the AQM, or intentionally making decisions opposite to its recommendation, and thus a reliance score R_s is defined as

$$R_s = \frac{|N_r - N_u|}{(N_r + N_u)} \tag{4}$$

Where N_r is the number of trials that a user follows the recommendation of a system, and N_u is the number of trials that the user makes decisions opposite to the system's recommendation. This computation of reliance reflects the extent to which the user bases their decisions on system recommendations. For example, for ten consecutive trials, if the user either always follows the system's recommendation, or always decides against the system (i.e., consistently flips the recommendations), R_s=1. And this makes sense, since in both cases user decisions – be it to follow the system or to flip the suggestion – are consistently based on the suggestions. However, if the user follows the system five times but not for the other five, R_s=0.

Figure 7 illustrates the changes of trust and reliance score with time. Figure 7 (a) shows the normalized subjective trust levels similar to Figure 1 (the 100% AQM curve is omitted since the trust there peaked at 1 for the majority of time), while Figure 7 (b) shows the computed reliance scores. The reliance score was calculated on a 10-point sliding window, i.e. N_r+N_u=10. To keep the number of sample points consistent, a 10-trial filter presented in equation (2) was applied to the trust scores, resulting in no trust values for the first 4 trials and the last 5 ones.

A comparison between Figure 7 (a) and Figure 7 (b) reveals that for different AQMs, the users' reliance curve follows a trend similar to their trust. The correlation analytics show that for the 70% AQM, a very strong correlation exists between the trust and reliance scores (r=0.977, p<0.05). In this case, both trust and reliance follow a uniformly decreasing trend, indicating that users do not trust this AQM and thus sometimes make decisions different from the recommendation of it. For the 80%

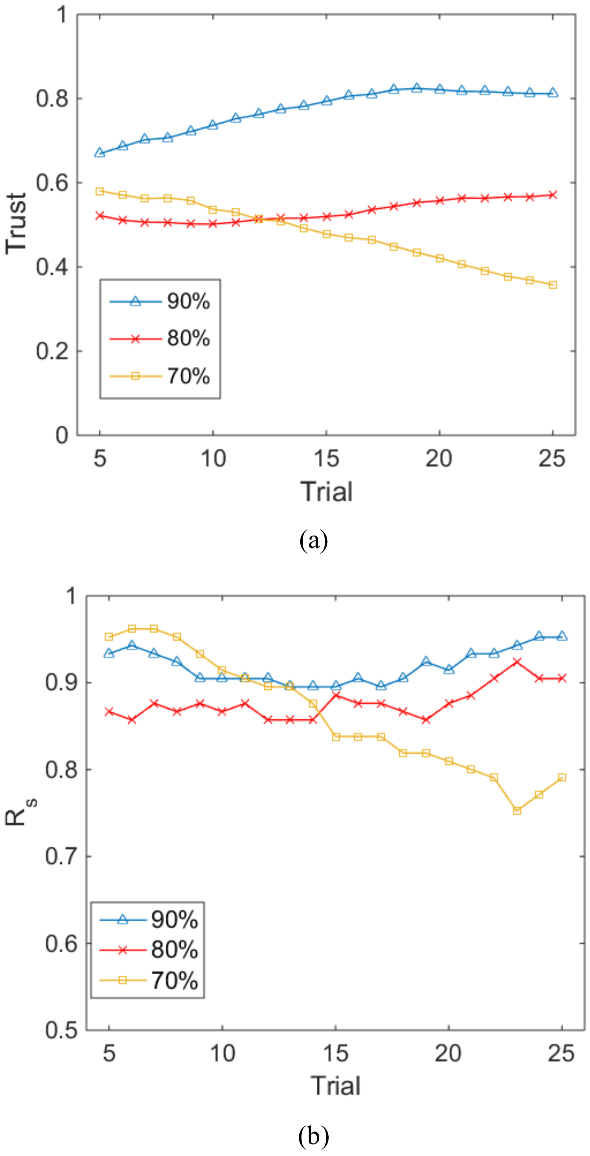


Figure 7. Trust and reliance score as a function of time for three AQMs. (a) Trust score filtered with a ten-point window[1]; (b) R_s calculated with a ten-point window.

AQM, the trust still considerably correlates to the reliance (r=0.673, p<0.05), where an increasing trend can be identified for both curves. For the 90% AQM, there is practically no correlation between the trust and reliance scores (r=0.06, p>0.05), while towards the end of the 30 trials, both trust and reliance stabilize at high levels, indicating that for a system with high accuracy, e.g. 90%, users tend to trust it, and make decisions in accordance with the system recommendations.

Furthermore, according to our earlier examination of trust change as a function of time shown in Figure 5 and Figure 6, we have identified that during Phase III, i.e. approximately from trial 19 onwards, the trust level stabilizes. Examining the correlation between trust and reliance for trials within this period, strong correlations are found for the 90% AQM (r=0.92, p<0.05), 80% AQM (r=0.86, p<0.05) and 70% AQM (r=0.76, p<0.05) respectively.

The latter results suggest that upon forming their trust perceptions, the users exhibit correlation between the explicitly reported trust and the implicitly computed reliance, as observed across all three AQMs. This finding implies that when the trust has stabilized the reported trust levels and the observed user behavior do correspond. This means that the users correctly perceive the performance exhibited by the system and not only report generally accurate trust perceptions, as already reported in the analysis of Figure 3, but also adjust their decisions accordingly. Reversing the process, monitoring user reliance in a system can provide a direct insight into the current level of trust the user is placing in the system. In simple words, this means that *trust may not need to be reported but can be learned* from the observed user behavior.

DISCUSSION

In this work we conducted a detailed analysis of the acquisition and extinction of trust. This uncovered several novel findings related to system performance and the dynamics of user trust.

The first finding refers to a better contextualization of trust changes to previously exhibited system performance and user trust levels. Figure 4 shows that system failures have much higher negative impacts on the user's subjective trust than the positive impacts of system successes. That is, the trust extinction due to wrong system recommendations is much faster to occur than trust acquisition due to system successes. In brief, it is *easier to destroy trust than to build it*. Although this is observed from AQMs with accuracies no lower than 70%, we posit that there exists a threshold for the accuracy of an automatic system, like our AQM, below which trust cannot be maintained. For example, for an AQM with 50% accuracy, it is very likely that the trust loss caused by system failures cannot be regained by the same number of system successes. As a consequence, to maintain a user's trust in a system, its accuracy should be well above 50%. Re-examining Figure 3, as the trust for the 70% AQM demonstrates a declining trend but the 80% AQM doesn't, the threshold of system performance is possibly between 70% and 80%, or in other words, a

[1] This figure is generated similar to Figure 1 with two exceptions: i) a wider (10-trial) window is applied due to the 10-trial window used for Figure 7 (b), which is used to remove the noise involved for the calculation of the reliance score. ii) the labelled trial is at the center of the sliding window, instead of at the end in Figure 1.

system with a performance of 80% is capable of maintaining a user's trust, but a 70% accurate system is not.

This finding reveals an important considerations for system designers: if an automatic system may fail the user with a chance higher than 20%, the designer should give a second thought whether such a system will be able to help the user given the possibility of a user's trust loss in the system. This leads to another question which deserves further attention, especially in the context of practical decision support systems. Very low levels of trust may lead to a situation, in which the user still relies on the system, but flips its recommendations, i.e. consistently follows the opposite of the option recommended by the system, in the case of binary decisions. However, this is likely to lead to user attrition, which may turn out to be more destructive. For example, if important financial investment decisions of customers are based on recommendations of a poorly-performing system, low trust may quickly lead to customer loss.

Furthermore, users are able to correctly perceive the accuracy of an automated systems, and adjust their trust accordingly to match the performance of the system. Although the adjustments may take up to twenty or even more trials, users have the capability to learn and adjust their decision pattern gradually. As shown clearly in Figure 5 and Figure 6, at the late stages of user interaction trust changes are close to zero, although some reasonably minor trust changes are still observed, e.g., lower than 0.25 for trust extinction and lower than 0.20 for trust acquisition. This demonstrates that our experiments conducted along 30 user-system trials and trust level reports capture the majority of the trust dynamics, but may still not fully cover the entire trust formation processes. Although the trust changes stabilize, some changes may yet occur, hence some experiments employing longer sequences of trials may be needed. This, however, may introduce carry-over bias as some users reported that 30 interactions were sometimes too long already. Hence, different experimental designs may be needed, e.g., where users report their trust every third interaction or where the implicit reliance score is considered as an indicator of the subjective trust.

An important finding is the three phases that depict the way users adjust their trust dynamically, following a learning, adjustment, and fine-tuning procedure. In the learning phase, users are cautious with their judgments on the system, and may have a high expectation on its performance. A system failure may cause huge trust loss, while a system success may not improve trust much as the users consider it normal for the system. However, with more trials conducted, users enter the adjustment phase: they have more experience and are able to make substantial adjustments to their trust, until its level approaches their overall feeling about the system. Then comes the fine-tuning phase, when their adjustments to trust are minor, which may not result in severe fluctuations in the trust level. Identifying these three phases of trust dynamics is an important step towards a refined and accurate user system design. If a user is asked how much a product is trusted during the learning or adjusting phase without sufficient experience with that product, the user is unlikely to provide a reliable trust score for the product which may mislead the product design. On the other hand, for the AQMs used in this study, Phase I and Phase II involve approximately 18 trials in total. It will be very interesting to see how many trials are required for more complex systems in these two phases, or if it is possible to quantify the number of trials required for each phase based on different system characteristics. This knowledge could be extremely helpful for realistic system design, e.g. for some critical systems such as those used in clinical settings for patient life support. We do not want doctors or nurses to undergo a long cold start period but we are keen to know how much they trust their systems in the long run, and this technology, if put into practice, may contribute to the resolution of many trust-related issues in system design and implementation.

Finally, our analysis of the correlations between trust and reliance shows high correlation scores, especially focusing on the reliance behaviors observed in Phase III. This finding implies that trust as an intrinsic mental construct may be possibly tracked via some external means, i.e. examination of the behavior or decision of the users. More precisely, our work suggests that trust can be monitored in real-time through reliance behaviors of users. Detecting trust variations can help tune system responses. For example, a consistent trust drop can be a signal to adapt the way a system interacts with the user, e.g., switching from automatic mode to manual mode, or changing the method of information presentation. Conversely, when a system detects over-trust, it could display warnings or mitigation messages accordingly.

However, it should be noted that the high correlations were observed for the 90%, 80%, and 70% AQMs only, and that the values of the correlations decrease with the accuracy level of the AQM. This highlights one limitation of this work, and raises a question about the correlations that would have been observed if lower-accuracy AQMs were deployed in the experiment. We posit that users would have identified the poor performance of AQMs at the 10%-20% accuracy levels and then consistently flipped the system recommendations. This could lead to negative correlation between trust and reliance, as the trust level would be low, while the reliance would still remain high. However, the correlation between the two at the moderate levels of system performance, e.g., 40%-60% AQM would be hard to predict. That said, from a practical perspective, we do not see any valid reasons leading to the development and adoption of such low performance systems.

CONCLUSION

This paper examines the relationship between system performance, a user's trust and reliance on the system. We observe significant differences in trust change after a single, double or triple consecutive failure or success, showing that trust acquisition becomes slower and slower with the number of successes, but trust extinction accelerates with the number of errors. In addition, we identify significant differences in the absolute value of trust between consecutive failures on the one hand and consecutive successes on the other, showing that trust acquisition is harder to obtain than extinction. Importantly, we show that increasing system familiarity decreases the rate of trust change, which we call the *inertia of trust*, which has important implications for the development and assessment of systems in general: practitioners should wait to the end of the learning and adjustment phases to measure trust when the fluctuation of subjective trust is only subject to a fine-tuning process. Finally, our analysis establishes the correlation between a user's reliance on a system and their trust level. These findings taken together, have important implications for general system design and implementation, by predicting how trust builds and when it stabilizes, as well as allowing for indirectly reading users' trust in real time based on their system reliance behaviours.

ACKNOWLEDGEMENT

This work is supported by the AOARD grant FA2386-14-1-0022/AOARD134131. The authors would also like to thank all the participants for contributing their time and handwriting to this investigation.

REFERENCES

1. Cynthia L. Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58, 6: 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7
2. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1: 114–126. https://doi.org/10.1037/xge0000033
3. Jinjuan Feng, Jonathan Lazar, and Jenny Preece. 2004. Empathy and online interpersonal trust: A fragile relationship. *Behaviour & Information Technology* 23, 2: 97–106. https://doi.org/10.1080/01449290310001659240
4. John Gall. 2002. *The Systems Bible: The Beginner's Guide to Systems Large and Small*. General Systemantics Press.
5. Shankar Ganesan and Ron Hess. Dimensions and Levels of Trust: Implications for Commitment to a Relationship. *Marketing Letters* 8, 4: 439–448. https://doi.org/10.1023/A:1007955514781
6. Yolanda Gil and Donovan Artz. 2007. Towards Content Trust of Web Resources. *Web Semant.* 5, 4: 227–239. https://doi.org/10.1016/j.websem.2007.09.005
7. Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward Establishing Trust in Adaptive Agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* (IUI '08), 227–236. https://doi.org/10.1145/1378773.1378804
8. Stephan Hammer, Michael Wißner, and Elisabeth André. 2015. Trust-based decision-making for smart and adaptive environments. *User Modeling and User-Adapted Interaction* 25, 3: 267–293. https://doi.org/10.1007/s11257-015-9160-8
9. Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3: 407–434. https://doi.org/10.1177/0018720814547570
10. Holger Hoffmann and Matthias Söllner. 2012. Incorporating behavioral trust theory into system development for ubiquitous applications. *Personal and Ubiquitous Computing* 18, 1: 117–128. https://doi.org/10.1007/s00779-012-0631-1
11. Jason D. Johnson. 2004. *Type of automation failure: the effects on trust and reliance in automation*. Georgia Institute of Technology.
12. J. Lee and N. Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10: 1243–1270. https://doi.org/10.1080/00140139208967392
13. John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1: 153–184. https://doi.org/10.1006/ijhc.1994.1007
14. John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1: 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
15. J McGuirl and N Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors* 48, 4: 656–665.
16. Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57, 1: 34–47.
17. Erik Millstone and Patrick van Zwanenberg. 2000. A crisis of trust: for science, scientists or for institutions? *Nature Medicine* 6, 12: 1307–1308. https://doi.org/10.1038/82102
18. Christine Moorman, Gerald Zaltman, and Rohit Deshpande. 1992. Relationships between Providers and Users of Market Research: The Dynamics of Trust within and between Organizations. *Journal of*

Marketing Research 29, 3: 314–328. https://doi.org/10.2307/3172742

19. Neville Moray, Toshiyuki Inagaki, and Makoto Itoh. 2000. Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks. *Journal of Experimental Psychology: Applied* 6, 1: 44–58.
20. Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5–6: 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5
21. Bonnie 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. https://doi.org/10.1080/00140139408964957
22. Susanne van Mulken, Elisabeth André, and Jochen Müller. 1999. An empirical study on the trustworthiness of life-like interface agents. In *Human-Computer Interaction (proceedings of Hci-International 1999), 152–156. Mahwah*, 152–156.
23. R. Parasuraman, T. Sheridan, B., and D. Wickens C. 2008. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making* 2, 2: 140–160.
24. Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *IUI*, 93–100. https://doi.org/10.1145/1111449.1111475
25. William B. Rouse. 1988. Adaptive Aiding for Human/Computer Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 30, 4: 431–443. https://doi.org/10.1177/001872088803000405
26. Juergen Sauer, Alain Chavaillaz, and David Wastell. 2016. Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics* 59, 6: 767–780. https://doi.org/10.1080/00140139.2015.1094577
27. Kristin E. Schaefer and David R. Scribner. 2015. Individual Differences, Trust, and Vehicle Autonomy: A pilot study. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*, 786–790.
28. Matthias Söllner, Axel Hoffmann, Holger Hoffmann, and Jan Marco Leimeister. 2012. How to use behavioral research insights on trust for HCI system design. 1703. https://doi.org/10.1145/2212776.2223696
29. Frank M. F. Verberne, Jaap Ham, and Cees J. H. Midden. 2012. Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors* 54, 5: 799–810.
30. Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6: 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9
31. Zheng Yan and Silke Holtmanns. 2008. Trust Modeling and Management: From Social Trust to Digital Trust. *http://www.igi-global.com/chapter/trust-modeling-management/6870*: 290–323. https://doi.org/10.4018/978-1-59904-804-8.ch013
32. Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (UMAP '16), 223–227. https://doi.org/10.1145/2930238.2930290