AFRL-AFOSR-JP-TR-2018-0008

Trust Measurement using Multimodal Behavioral Analysis and Uncertainty-Aware Trust Calibration

Fang Chen
NATIONAL ICT AUSTRALIA LIMITED

01/05/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

FORM SF 298

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
16-01-2018

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
30 Sep 2014 to 29 Sep 2017

4. TITLE AND SUBTITLE
Trust Measurement using Multimodal Behavioral Analysis and Uncertainty-Aware Trust Calibration

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-14-1-0022

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Fang Chen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NATIONAL ICT AUSTRALIA LIMITED
L 5 13 GARDEN ST
EVELEIGH, 2015 AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2018-0008

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report summarizes our major research activities, study results and research accomplishments out of the trust measurement project in the past year. This is also the final report of the project. We have conducted different experiments on trust examination with varied system accuracy, and human trust in predictive decision making. From the study we have found that: 1) people can correctly perceive the accuracy of the system and adjust their trust accordingly; 2) there exists a strong link between human decisions, trust and perception, and trust can be inferred from a couple of decisions; 3) different uncertainty types (e.g. risk and ambiguity) affect human trust differently; 4) cognitive load levels also affect human trust differently because of cognitive resources available. These trust variations can be examined by physiological signals (e.g. GSR). Our future work will focus on investigating other physiological signals (e.g. BVP) as a means to quantify user trust, as well as identifying the trust patterns when humans play different roles in the human-machine collaboration.

15. SUBJECT TERMS
Trust

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			ROBERTSON, SCOTT
Unclassified	Unclassified	Unclassified	SAR	18	**19b. TELEPHONE NUMBER** *(Include area code)* +81-042-511-7008

Final Report for AOARD Grant FA2386-14-1-0022/AOARD134131

Trust Measurement in Human-Machine Interactions using Multimodal Behavioral Analysis and Uncertainty-Aware Trust Calibration

September 30, 2017

Name of Principal Investigators (PI and Co-PIs): Fang Chen

- e-mail address : fang.chen@data61.csiro.au
- Institution : DATA61, CSIRO, Australia
- Mailing Address : Level 5, 13 Garden Street, Eveleigh NSW 2015, Australia
- Phone : +61 2 9490 5601

Period of Performance: September/30/2016 – September/29/2017

Abstract: This report summarizes our major research activities, study results and research accomplishments out of the "trust measurement" project in the past year. This is also the final report of the project. We have conducted different experiments on trust examination with varied system accuracy, and human trust in predictive decision making. From the study we have found that: 1) people can correctly perceive the accuracy of the system and adjust their trust accordingly; 2) there exists a strong link between human decisions, trust and perception, and trust can be inferred from a couple of decisions; 3) different uncertainty types (e.g. risk and ambiguity) affect human trust differently; 4) cognitive load levels also affect human trust differently because of cognitive resources available. These trust variations can be examined by physiological signals (e.g. GSR). Our future work will focus on investigating other physiological signals (e.g. BVP) as a means to quantify user trust, as well as identifying the trust patterns when humans play different roles in the human-machine collaboration.

List of Publications

a) Award

J. Zhou, S. Z. Arshad, S. Luo and F. Chen received the "Reviewers' Choice Award" and the "Brian Shackel Award" for the paper "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making" at the 16th IFIP TC. 13 International Conference on Human-Computer Interaction (INTERACT 2017). The Brian Shackel Award is awarded to one selected paper in recognition of the most outstanding contribution with international impact in the field of human interaction with, and human use of, computers and information technology. (http://ifip-tc13.org/awards/)

b) Publications in peer-reviewed journals and books in the past project year period

[1] J. Zhou and F. Chen, "Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent", Springer, 2017. In press.

[2] J. Zhou, S. Z. Arshad, X. Wang, Z. Li, D. Feng, and F. Chen, "End-User Development for Interactive Data Analytics: Uncertainty, Correlation and User Confidence", IEEE Transactions on Affective Computing, 2017.

[3] J. Zhou and F. Chen, "DecisionMind: Revealing Human Cognition States in Data Analytics-Driven Decision Making with a Multimodal Interface", Journal on Multimodal User Interfaces, (Springer), 2017.

[4] J. Zhou, M. A. Khawaja, Z. Li, J. Sun, Y. Wang, and F. Chen, "Making Machine Learning Useable by Revealing Internal States Update – A Transparent Approach", International Journal of Computational Science and Engineering, Vol.13, No. 4, pp.378-389, 2016.

c) Papers published in peer-reviewed conference proceedings in the past project year period

[5] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen, "User Trust Dynamics: An Investigation Driven by Differences in System Performance", IUI2017.

[6] J. Zhou, S. Z. Arshad S. Luo and F. Chen, "Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making", the 16th IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT 2017), 2017. (Reviewer's Choice Award), ("The Brian Shackel Award" in recognition of the most outstanding contribution with international impact in the field of human interaction with, and human use of, computers and information technology)

[7] J. Zhou, S. Z. Arshad, S. Luo, K. Yu, S. Berkovsky, and F. Chen, "Indexing Cognitive Load Using Blood Volume Pulse Features", Proceedings of CHI 2017 Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1861-1868, Denver, USA, 2017.

[8] S. Luo, J. Zhou, H. Duh, and F. Chen, "BVP Feature Analysis for Intelligent User Interface", Proceedings of CHI 2017 Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2269-2275, Denver, USA, 2017.

d) Papers submitted

[9] J. Zhou and F. Chen, "Human-In-The-Loop Machine Learning with A 2D Transparency Space", PAKDD 2018.

Pending Patent

Generating a user-specific trustworthy interface, Kun Yu, Fang Chen and Shlomo Berkovsky

Contents

1. Introduction

In recent years, trust has been found to be a critical factor driving human behavior in human-machine interactions in many high-risk human-machine interaction domains such as aviation, the military command and control. However, due to the sophisticated technologies and increased levels of automation provided by machines today, humans are no longer able to know every technical detail or working mechanism of their machine teammate, and hence determining the system performance based on full system understanding becomes increasingly difficult. As a consequence, in many situations humans actually base their trust on limited perceptions of the machine partner, and make decisions accordingly. Trust is also influenced by the types and format of information accessible to humans, their individual approaches to develop and determine trust, and other aspects such as system capability and reliability.

During the past year, we have conducted three major studies involving:

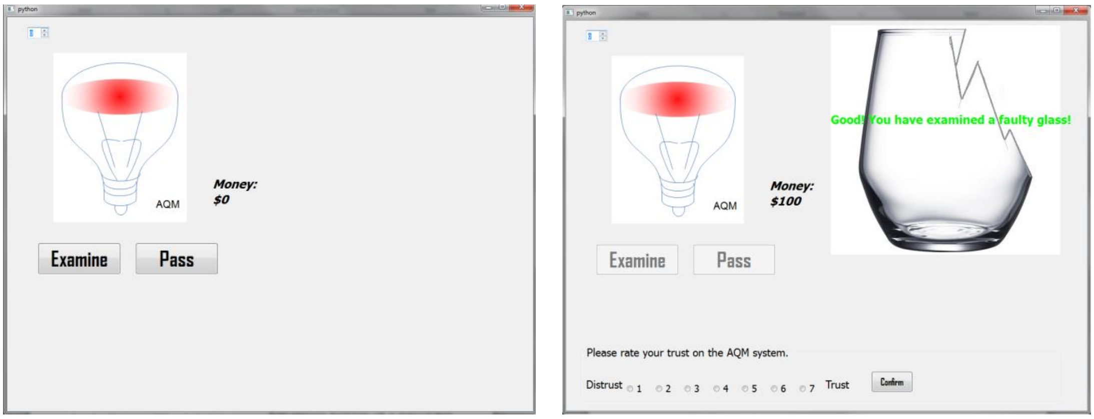
- Continued experiments to examine trust dynamics, its relationship with human decisions and how human perception may affect trust.
- Data analytics to identify the effects of uncertainty and cognitive load on trust as well as trust changes based on variations of system accuracy.
- Examination of physiological signals as a means to link human physiological responses to mental states in terms of cognitive load.
- A framework of informed decision making called *DecisionMind* is proposed to show how human's behaviour and physiological signals are used to reveal human cognition states (using user confidence as a case study) in predictive decision making.

In the following part of the report each study will be addressed in specific.

2. Trust Examination with Varied System Accuracy

2.1 Experiment

We operationalized a binary decision making task in our experiment to examine the decision-trust relationship, which is expected to be able to generalized to complicated decision-making problems. The scenario of the experiment was a typical product 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) (see Figure 1). 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).



Good! You have examined a faulty glass!

Please rate your trust on the AQM system.

Figure 1. The interface of the experiment, which involves the AQM system (the light at top left of the interface), the human decisions (middle left of the interface), the revealed glass quality (top right of the interface), and the participant ratings in terms of trust and perceptions (bottom of the interface, only trust rating is shown).

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. In each trial, the participant could trust the AQM or override it and make his/her own decision. Twenty-two participants took part in the experiment.

We collected the following information for each trial:

- AQM’s suggestion (light on or off);
- Participant’s binary decision (pass or examine);
- Actual glass condition (good or faulty);
- Perceived system performance (0% to 100%);
- Estimated self-performance (0% to 100%);
- Subjective trust rating.

We also calculated the normalized trust (to scale the trusting rate of a single participant to the [0, 1] range) and reliance rate (the proportion of human decisions consistent with the system suggestion).

2.2 Results and Discussions

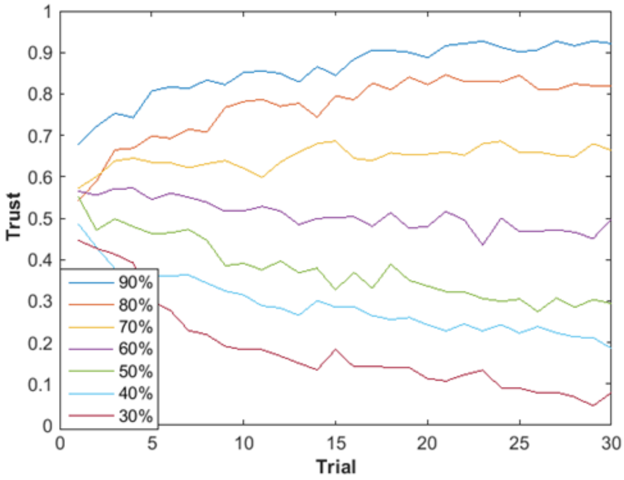


Figure 2. The dynamics of trust during the 30-trial period.

The normalized trust averaged across all the participants and AQMs is plotted in Figure 2. At the beginning, i.e. the trust rating after the first trial, the order of participants' trust in the AQMs is somehow randomized for all the AQMs, indicating that the participants do not differentiate their trust significantly after a single trial, due to limited experience with the systems. As the participants continue working with the AQMs, after trial 5 all the trust levels are well separated and align with the accuracies of the respective AQMs. Furthermore, it is found that from trial 5 onwards, the participants have demonstrated different trust for the AQM. The trend of trust level separation continues towards the end of the trials until the trust levels are stable.

The implications of participants' trust on their decisions are investigated via examining the responses of all the participants at different trust levels. We calculate the reliance rate of participants as the proportion of consistent decisions with the system over a set number of consecutive trials, and their relationship between reliance rate and trust is depicted in Figure 3. The error bars indicates the variance at each trust level, and the trust of all participants is normalized to the range between 0 and 1. The reliance rate demonstrates a clear rising trend with trust, suggesting that participants rely more on systems when they trust them which is consistent with existing understanding. On the other hand, the decreasing variance of reliance rates reveals another interesting finding: at low trust levels, although the overall reliance rate are low, participants demonstrate high variance in reliance rates. This suggests that participants rely on the system in different ways, sometimes even if they do not trust the system, they may try decisions consistent with its recommendation. In comparison, as trust level increase, the rate of reliance also converge, implying that participants tend to follow the system suggestions when they believe the system to be highly reliable. This finding is interesting and reveals that participant's trust level can be inferred from a couple of decisions rather than a single decision, however the latter has been used as an indicator of trust in many other investigations.

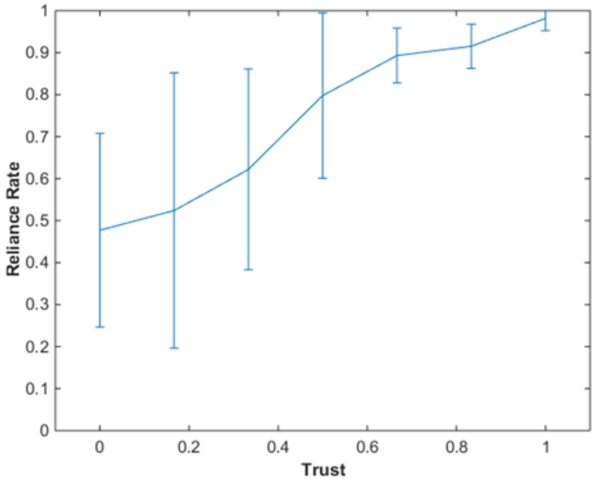


Figure 3. Trust affects the trend and variance of participants' reliance rate. The error bars in the plot represent standard deviations.

Performance refers to the proportion of correct decisions amongst all the decisions made on one AQM. We have asked participants to estimate their performance based on their estimation on all prior trials. In the meanwhile via comparing the decisions of participants with the outcome of glasses, we are able to calculate their actual performance. Figure 4 shows both the actual performance of the participants and the perceived performance of their own. Interestingly, in the initial several trials participants are not able to precisely estimate their performance, although it is easier compared with situations when more trials have been done. It should be noted that if a participant is good at memorizing the previous trials, he/she should be able to increase the accuracy of performance estimation as she/he approaches the end of the 30 trials. An interesting finding from Figure 4 is that at the end of the trials, for the more accurate AQMs (90%, 80% and 70%), participants' estimated accuracies are significantly higher than their actual performance; however they are still capable of discriminating the order of these AQMs.

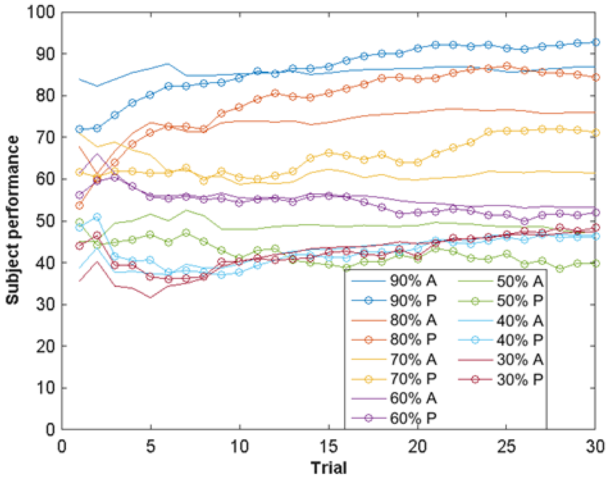


Figure 4. Perceived vs actual subjective performance, where 'A' denotes actual performance and 'P' denotes perceived performance of the participant.

If the participants estimate their own performance differently from their real performance, how about their perceptions on the AQMs? Figure 5 provides the answer and depicts the dynamics of AQM perceptions. The results suggest that the participants are capable of perceiving the system performance with a higher accuracy. At the fifth trial, the perceived system accuracies for different AQMs already differ significantly. However towards the end of the 30 trials, there are no more significant perception changes for all AQMs, implying that the perceived system accuracies have stabilized. These findings imply that the participants are able to adjust their perceptions and reach accurate estimations towards the end of the trials, especially for the most accurate AQMs (90%, 80% and 70%). For the other less accurate AQMs especially the 50%, 40% and 30% ones, perception bias of over 10% can be observed towards the end of the trials, but the order of accuracy is still correctly perceived.

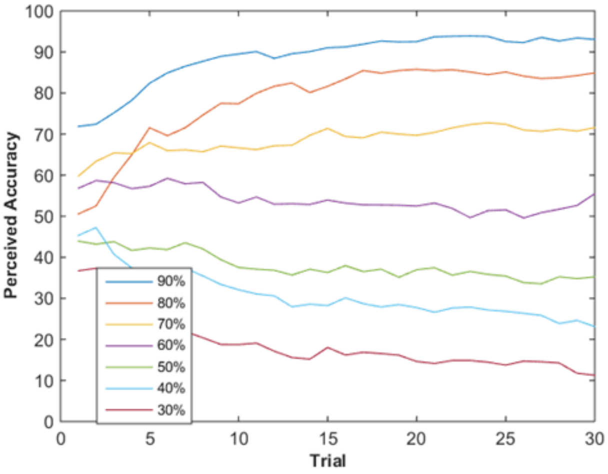


Figure 5. Subjective perceptions of the AQM accuracies.

Due to the similarity between perceived system accuracy and participants' trust in the AQMs, we would like to see how the system perceptions affect participant decisions. For all the participants, the relationship between their perceived AQM accuracies and the rate of reliance is illustrated in Figure 6. A linear regression is calculated to predict the reliance rate based on the perceived accuracy, with

$$R_r = 0.0047 \times P_a + 0.521$$

Where R_r refers to the reliance rate and P_a refers to the perceived system accuracy. However, even if the perceived system accuracy is extremely low, the participant may still take a chance to follow the system's suggestions now and then.

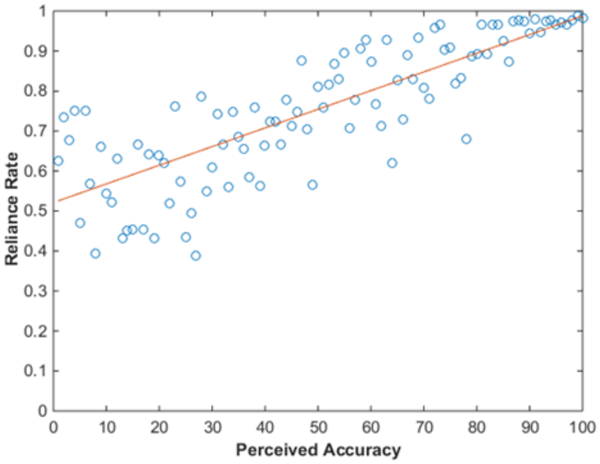


Figure 6. Reliance rate increases with perceived system accuracy for all participants. The linear regression result is shown in red.

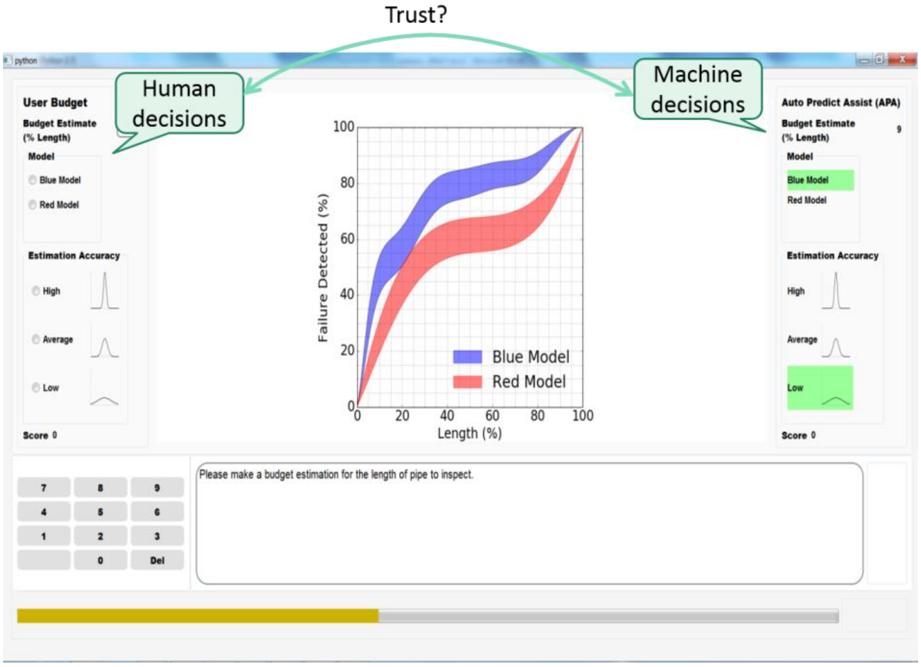
To sum up, we have shown that participants are capable of estimating the system accuracies reasonably well and gradually adapting their trust levels to the system performance within 30 trials. The positive relationship between trust and system accuracy perception, as implied by the high correlation, suggests the tight link between the two mental constructs. This finding implies that if a participant perceives the performance of a system, his/her trust in the system will be affected accordingly; furthermore, the increased trust may result in more decisions consistent with the recommendation of a decision support system.

3. Trust in Predictive Decision Making

3.1 Experiment

For this study, 42 participants were recruited with three different backgrounds with the ages ranging from between 20 to 57. The three groups are divided into people with machine learning researchers, non-machine learning researchers and administrative staff. The machine learning researcher group contains 14 participants (11 male, 3 female). The participants have a background in machine learning or data mining. The second group contains 19 participants (18 male, 1 female) with non-machine learning research background which do not use a large amount of mathematics as a part of their work. The majority of these participants are software engineers or researchers in fields such as cloud computing and mobile systems. Lastly the administrative staff consistent of 9 participants (3 male, 6 female) with non-technical background. These include staff from a combination of reception, accountant, human resources, communication and legal.

The participants in this experiment had various education with 9 participants who have PhD degrees, 10 participants with master's degrees and the rest 23 with bachelor's degrees or are current bachelor candidates. Figure 7 shows the screenshot of a task performed in the study.



Please make a budget estimation for the length of pipe to inspect.

Figure 7. Screenshot of a task performed in the study.

3.2 Effects of Uncertainty and Cognitive Load on User Trust

Trust and Uncertainty: We found that better presentation and adequate communication of uncertainty inherent in the underlying ML process can improve the trust of the user in the system and lead to better and effective decisions. In our case, we experimented with visualizing and communicating two forms of uncertainty, namely, risk and ambiguity.

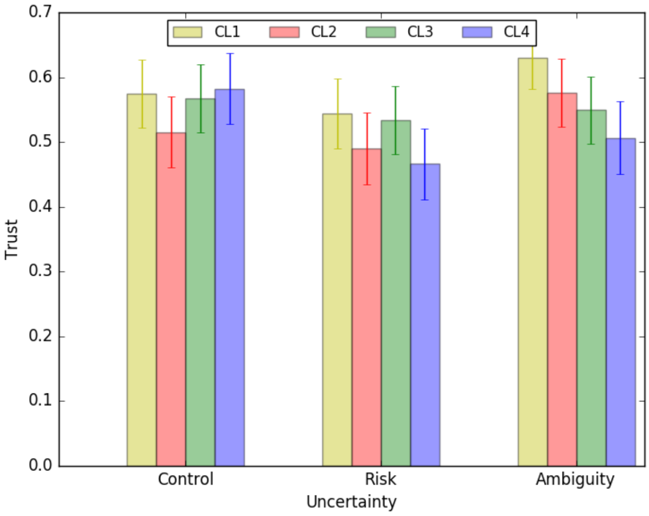


Figure 8. Trust over Uncertainty Presented; Control (No Uncertainty), Risk (Non-Overlapping Uncertainty) and Ambiguity (Overlapping Uncertainty).

Looking at the overall results (see Figure 8), no clear trends can be observed for risk type uncertainty condition, but a clear trend of decreasing trust can be seen for uncertainty of type ambiguity as cognitive load level increases. It can be said that under low cognitive load (implying greater availability of cognitive resources), users felt more confident in analyzing and interpreting the ambiguity type of uncertainty and therefore appeared to trust the judgement/recommendation of the automated predictive assistant as it made more sense to them. However, under high cognitive load, the users might find themselves almost at the edge of their working memory capacity. Limited cognitive resources would result in lower understanding of the ambiguity type of visual. This in turn is indicated by reduced trust in the system and its recommendations. This phenomenon seems to be in line with findings that the better the person understands the system and it's working the greater the person is willing to trust it. Further drilling down this trust (over ambiguity type uncertainty) phenomenon into subject groups (administration, machine learning experts and non-machine learning experts) also leads to an interesting insight (see Figure 9). Clearly the level of trust for ambiguity type uncertainty presentation appears to drop for all subject groups as cognitive load increases. High cognitive load appears to impact the trust in the similar way for all administrative staff and experts (whether they be machine learning or non-machine learning).

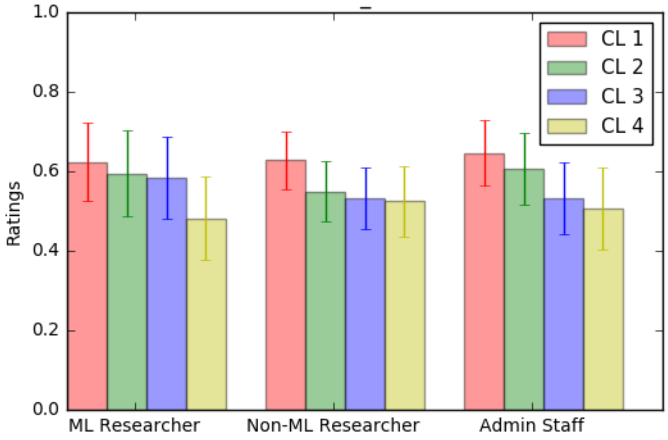


Figure 9. Trust for Ambiguity type uncertainty.

Trust and Cognitive Load: It is well known that human performance can be significantly affected by high cognitive or mental workload. Cognitive workload is the load on working memory that the user experiences when engaged in a cognitive problem. In our case, the trust in decision making is influenced by a cognitive phase where user tries to make sense of the model data/visuals presented. Since the decision making task was soft time bound, the user must make efficient use of available cognitive resources in order to complete the task. Here we look at the two extreme conditions where most cognitive resources were expected to be available (CL1) and where least cognitive resources were expected to be available (CL4). As stated earlier in the results section, Friedman test for both these extreme conditions (CL1 & CL4) turned out to be significant.

In low load condition (CL1), trust for ambiguity type uncertainty was significantly higher than risk type uncertainty. The trend seems to be the same for all subject subgroups (see Figure 10). Trust, under low load conditions, seems to be consistently higher for all groups whenever uncertainty of ambiguity type is presented. However, on further testing, only non-ML group (middle group of columns in Figure 10) yielded significantly (p<.006) higher trust level uncertainty types ambiguity to that of risk. A limitation here could be the lower number of subjects in groups other than non-ML experts. These findings go on to support the idea discussed earlier that uncertainty of type ambiguity can be readily processed by users only under low cognitive load conditions.

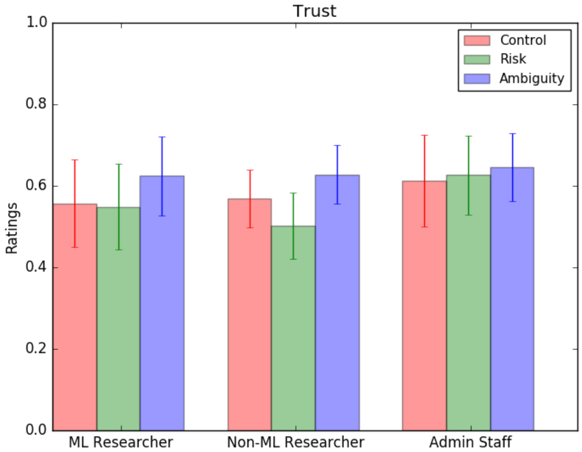


Figure 10. Trust over Subject Groups (Low CL).

Likewise in high load condition (CL4), trust for risk type uncertainty was significantly lower than control condition of no uncertainty presentation. The trend seems to be similar for all subject subgroups (see Figure 11). Trust, for both uncertainty conditions, seems to be consistently lower for all groups with respect to control condition. On further testing, only non-ML group (rightmost group of columns in Figure 11) yielded significantly (p<.003) lower trust level uncertainty type risk to that of control. A limitation here could be the lower number of subjects in other groups than non-machine learning experts.

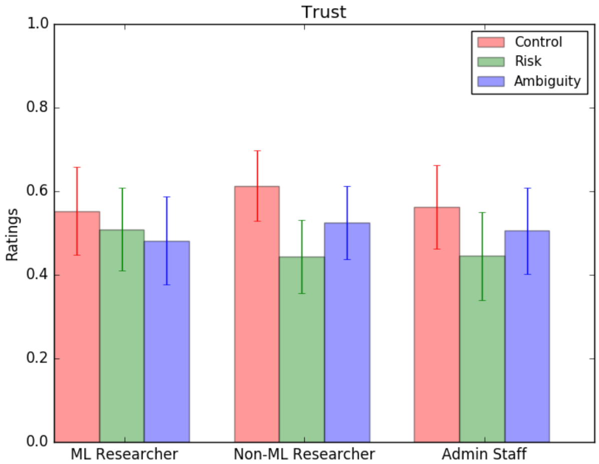


Figure 11. Trust over Subject Groups (High CL).

In summary, our analysis found that uncertainty presentation leads to increased trust but only under low cognitive load conditions when users had sufficient cognitive resources to process the information. Presentation of uncertainty under high load conditions (when cognitive resources were short in supply) leads to a decrease of trust in the system and its recommendations.

3.3 Analysis of GSR Responses

GSR responses from subjects were collected and analysed in this study to find relations between human physiological signals and trust. Figure 12 shows an example of a GSR signal during a task time.

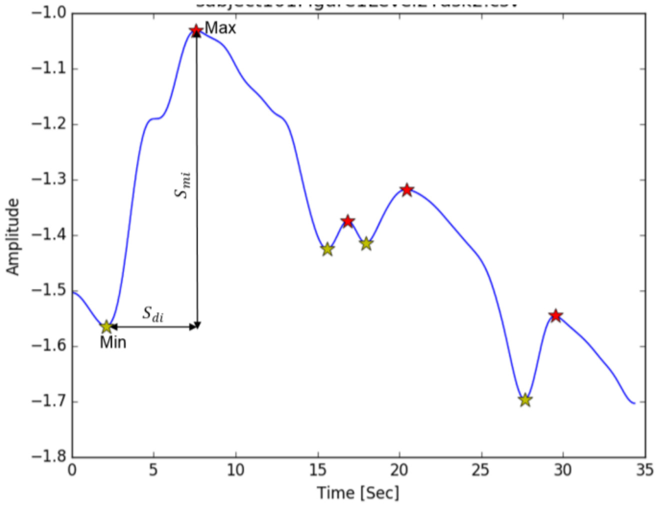


Figure 12. An example of GSR signal, extremas and extrema features of GSR.

GSR Features and Uncertainty: One-way ANOVA tests with post-hoc analysis using *t*-test were performed to evaluate differences of each GSR features among tasks of four CL levels under a fixed uncertainty type. Statistically significant differences of all GSR features among CL levels have not been found under control and risk uncertainty except ambiguity uncertainty. When participants experienced ambiguity uncertainty, one-way ANOVA tests showed a statistically significant difference in GSR values among four CL levels for GSR features of sum of magnitude S_m ($F_{3,112}$=4.111, p=.008), sum of magnitude per second S_{ms} ($F_{3,112}$=3.0, p=.033), average gradient G_a ($F_{3,112}$=2.697, p=.048), and maximum gradient $G^{\max}$ ($F_{3,112}$=3.81, p=.012). Figure 13 shows the example of GSR feature of sum of magnitude S_m over four CL levels under different uncertainty types.

The results found that the increase of cognitive load made GSR features such as sum of magnitude S_m, sum of magnitude per second S_{ms}, average gradient G_a and maximum gradient $G^{\max}$ values increased significantly under ambiguity uncertainty. Therefore, these GSR features can be used to index trust variations among tasks of various CL conditions.

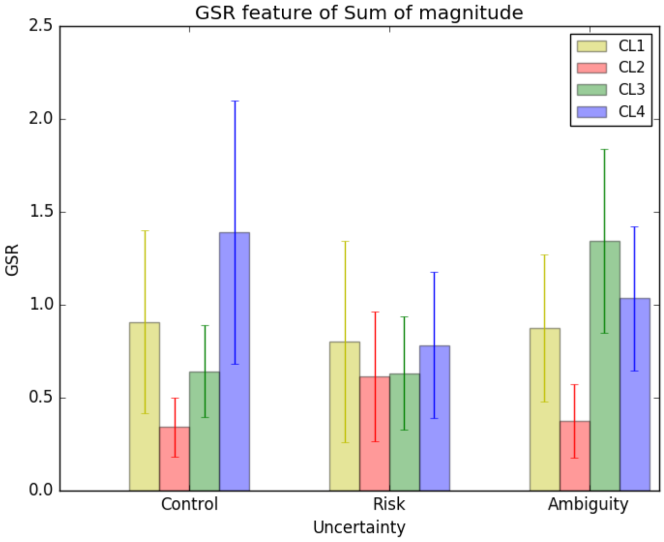


Figure 13. GSR feature of sum of magnitude over four CL levels under various uncertainty types.

GSR Features and Cognitive Load: One-way ANOVA tests with post-hoc analysis using *t*-test were performed to evaluate differences of each GSR features among tasks of various uncertainty types under a fixed CL condition. Under the low CL condition (CL1), one-way ANOVA tests did not find significant differences of GSR features among tasks of various uncertainty types. However, under high CL conditions (e.g. CL3 or CL4), one-way ANOVA tests found significant differences among tasks of three uncertainty conditions in various GSR features: σ_G, S_m, S_{ms}, $S_m^{\max}$, G_a, and $G^{\max}$. Figure 14 shows the example of $S_m^{\max}$ over four CL levels under three uncertainty conditions.

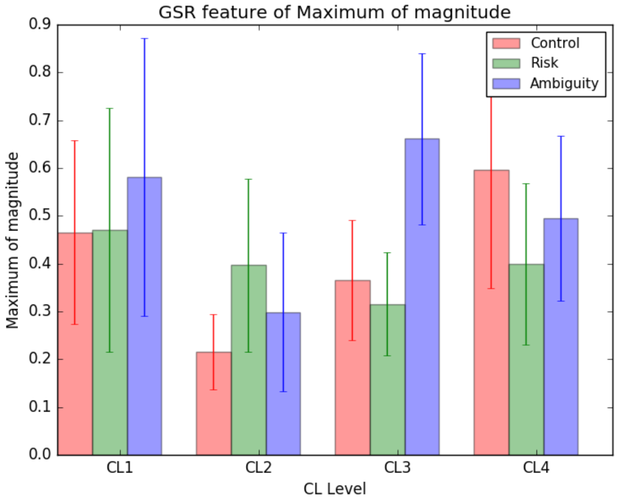


Figure 14. GSR feature of maximum estimate of gradient ($S_m^{\max}$) over four CL levels.

Despite no significant differences in GSR values found among tasks with various uncertainty conditions under extreme CL levels (CL1 and CL4), GSR values still showed a trend among tasks of various conditions, for example, the mean value of $S_m^{\max}$ was higher in ambiguity uncertainty than that in risk uncertainty under CL1, whereas the mean value of $S_m^{\max}$ was lower in risk uncertainty than that in control condition under CL4 as shown in Figure 14. These findings of GSR variations are consistent with trust differences among tasks of three uncertainty conditions. The results suggested that various uncertainty conditions made GSR features such as $S_m^{\max}$ values significantly different. Therefore, these GSR features can be used to index trust variations among tasks of various uncertainty conditions.

4. User Confidence and Uncertainty in Predictive Decision Making

4.1 Experiment

Water pipe failure prediction is used as case study for this research. Water supply networks constitute one of the most crucial and valuable urban assets. If high-risk pipes can be identified before failure onset, it is likely that repairs can be completed with minimal service interruption, water loss and negative reputational community impacts. Identifying an accurate predictive measure for 'imminent failure' allows utility companies to prioritize preventive repairs that would be significantly lower than the cost of full-scale failures. Thus, utility companies use outcomes from failure prediction models, to make renewal plans based on risk levels of pipes and also reasonable budget plans for the pipe maintenance. Here 'uncertainty' simply refers to an interval within which the true value of a measured quantity would lie with a given probability.

The nature of task was on-screen budget estimation with expected variation to be noted as upper and lower limits based on different uncertainty conditions as shown in Figure 15.

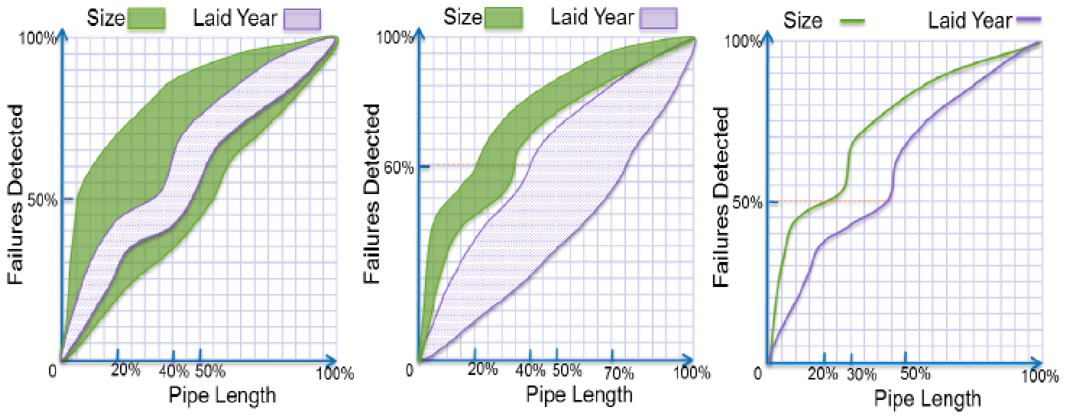


Figure 15. Exemplary figures used in the experiment, as a means to induce different level of uncer tainty.

All together there were 26 subjects (each one a user of ML predictive systems at local water department). Ages ranged from 23 to 45 with an average age of about 30 years. Educational qualifications were largely postgraduate (13 PhD, 6 Masters, 4 Bachelors, 3 others). Subject subgroups comprised of nine machine learning experts, eight non-machine learning experts and nine administrative staff.

4.2 Revealing User Confidence in Predictive Decision Making

A framework of informed decision making called *DecisionMind* is proposed to show how human's behaviour and physiological signals are used to reveal human cognition states in ML-based decision making (see Figure 16). Our work takes the revealing of user confidence in ML-based decision making as an example to demonstrate the effectiveness of the proposed approach. Based on the revealing of human cognition states during ML-based decision making, the chapter presents a concept of adaptive measurable decision making to show how the revealing of human cognition states are integrated into ML-based decision making to make ML transparent. On the one hand, human cognition states could help understand to what degree humans accept innovative technologies. On the other hand, through understanding human cognition states during ML-based decision making, ML-based decision attributes/factors and even ML models can be adaptively refined in order to make ML transparent.

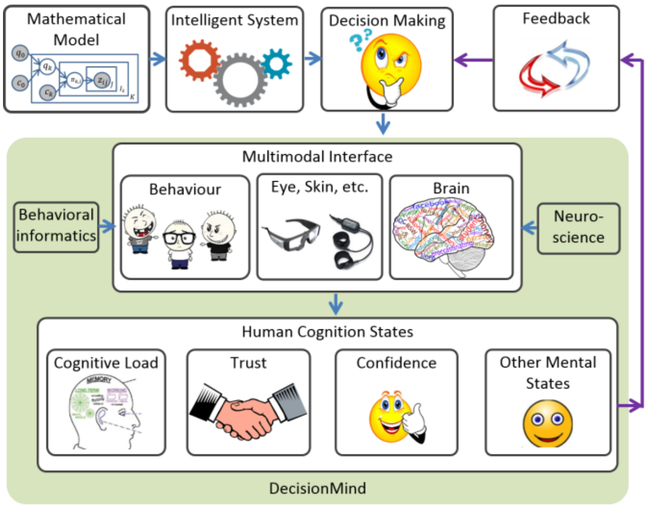


Figure 16. Framework of informed decision making -- DecisionMind.

5. Conclusions and Future Work

5.1 Conclusions

This research carried out trust studies from different perspectives: trust examination with varied system accuracy; trust in text-chat environment; effects of uncertainty and cognitive load on trust in predictive decision making; and effects of uncertainty on user confidences in predictive decision making.

In the study of trust examination with varied system accuracy, we found that users' trust stabilize over time and users could correctly perceive the accuracy of the system and adjust their trust accordingly. We also identified the fact that if a system is featured with a high level of accuracy, then people may tend to rely on them in terms of decisions, however once the system accuracy falls below an acceptance threshold, the reliance may deteriorate as well.

In the study of effects of uncertainty and cognitive load on trust in predictive decision making, it was found that uncertainty presentation can lead to increased trust but only under low cognitive load conditions when user has sufficient cognitive resources to process the information. Presentation of uncertainty under high load conditions, when cognitive resources are short in supply will lead to lowering of trust in the system and its recommendations. The further group-wise analyses have not found significant differences in trust perceptions among subject groups. GSR features were also analyzed to find relations between human physiological responses and trust variations.

This study proposed a framework of informed decision making called *DecisionMind* to show how human's behaviour and physiological signals are used to reveal human cognition states in predictive decision making. The revealing of user confidence is used as a case study based on this framework.

5.2 Future Work

Future work will include analyzing the trust variations based on other physiological signals su ch as pupillary response and Blood Volume Pulse (BVP). Further understanding of the effects of system accuracy variations on trust changes in the system can also benefit the measure ment of user trust and development of intelligent systems.

5.2.1 BVP Feature Analysis for Intelligent User Interface

In our study, other physiological signals such as BVP signals were collected for analysis. BVP sensor has been becoming increasingly common in devices such as smart phone and smart watches. These devices often use BVP to monitor the heart rate of an individual. There has been a large amount of research linking the mental and emotional changes with the physiological changes. The BVP sensor measures one of these physiological changes known as Heart Rate Variability (HRV). HRV is known to be closely related to Respiratory Sinus Arrhythmia (RSA) which can be used as a measurement to quantify the activity of the parasympathetic activity. However, the BVP sensor is highly susceptible to noise and therefore BVP signals often contain a large number of artefacts which make it difficult to extract meaningful features from the BVP signals (see Figure 17).

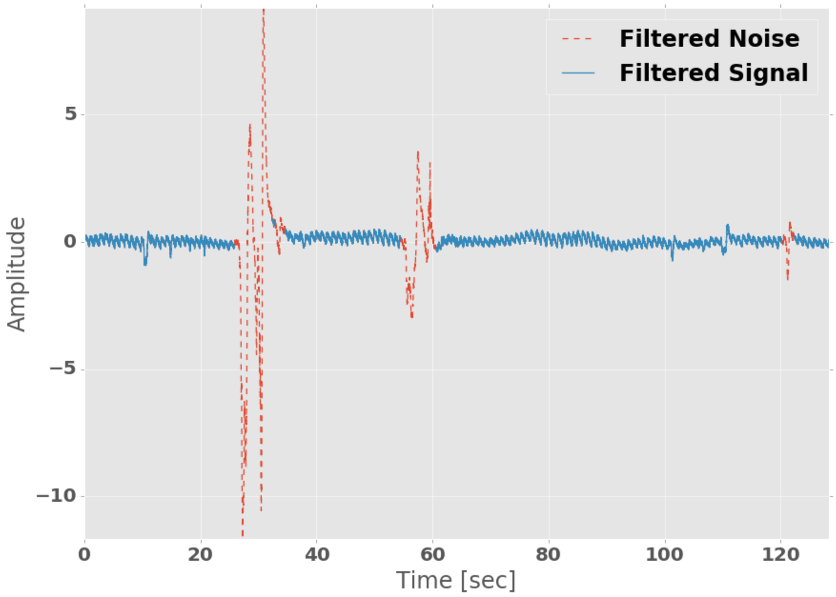


Figure 17. An example of noisy BVP signal and filtered signal with our approach.

We proposed a new algorithm to filter artefacts from BVP signals. The algorithm is comprised of two stages. The first stage is to detect the corrupt signal using a Short Term Fourier Transform (STFT). The second stage uses Lomb-Scargle Periodogram (LSP) to approximate the Power Spectral Density (PSD) of the BVP signal. The algorithm has shown to be effective in removing artefacts which disrupt the signal for a short period of time (see Figure 17). This algorithm provides the capability for BVP signals to be analysed for frequency based features in HRV which traditionally could be done from the cleaner signals from electrocardiogram (ECG) in medical applications.

Our future work will focus on examining trust based on these BVP signal analysis results.

5.2.2 Trust and Human Roles in HCI

Based on our study of trust dynamics, we have identified that trust, perception and human decision making are tightly linked to each other. However, all the findings are based on investigations in one context, i.e. the machine serves as a recommendation system and the human is the decision maker. However, in realistic practice, there are many other occasions when the human have different forms of collaboration with their machine partner. For example, a human can be the task operator while the machine can be the monitor, and it will alert human when a mistake is detected. Under certain circumstances the roles of human and machine may be swapped, i.e. the human monitors the operations of the machine and is able to override it if an error is perceived. Or otherwise, the human and the machine can be responsible for one sub-task respectively and the overall task outcome will depend on the successful implementations of both. For all the different types of human-machine collaboration or interaction, the trusting or trusted relationship can be different, and we would like to expand our experiments as a means to examine them. The proposed study, being quantitative and explorative, are expected to reveal a number of findings that benefit interaction system design and analytics, and help us to build a framework that can be used to characterize, predict, diagnose and further adjust the trusting relationship within a human-machine system.