AFRL-AFOSR-JP-TR-2022-0052

A Longitudinal Study of Trust Calibration Methods with Individual Differences

Chen, Fang
COMMONWEALTH SCIENTIFIC AND INUSTRIAL RESEARCH ORGANISATION
LIMEEONE AVE
CABERRA, , 2612
AU

06/21/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**	
20220621	Final	**START DATE** 20180710	**END DATE** 20210709

4. TITLE AND SUBTITLE
A Longitudinal Study of Trust Calibration Methods with Individual Differences

5a. CONTRACT NUMBER	**5b. GRANT NUMBER**	**5c. PROGRAM ELEMENT NUMBER**
FA2386-18-1-4091		61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Fang Chen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
COMMONWEALTH SCIENTIFIC AND INUSTRIAL RESEARCH ORGANISATION
LIMEEONE AVE
CABERRA 2612
AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2022-0052

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 calibration” project in the past years. 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 revealed that: 1) AI performance, in particular system errors, has a huge implication on human's trust adjustment and decision making. 2) Humans are able to perceive the detailed system performance not only at system level but also at subsystem level using systematic sampling strategies. 3) Trust knowledge acquired by human could be used to guide their future decision making. 4) The influence information of training data points (functioned as fact-checking) on predictions can benefit trust. 5) Personality traits affect trust in predictive decision making differently under different cognitive load levels. Our on-going work is focusing on refined quantification of human trust, and its implication on perception and decision making in the human-machine collaboration context.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		
U	U	U	SAR	16

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER** *(Include area code)*
ALAN LIN	227-7009

Final Report for AOARD Grant FA2386-18-1-4091

A Longitudinal Study of Trust Calibration Methods with Individual Differences
September 30, 2021

Name of Principal Investigators (PI and Co-PIs) : Fang Chen
- e-mail address : fang.chen@uts.edu.au
- Institution : Data Science Institute, University of Technology Sydney, Australia
- Mailing Address : PO Box 123 Broadway, NSW 2007, Australia
- Phone : +61 2 9514 4538

Period of Performance: July/10/2018 – July/09/2021

Abstract: This report summarizes our major research activities, study results and research accomplishments out of the "trust calibration" project in the past years. 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 revealed that: 1) AI performance, in particular system errors, has a huge implication on human's trust adjustment and decision making. 2) Humans are able to perceive the detailed system performance not only at system level but also at subsystem level using systematic sampling strategies. 3) Trust knowledge acquired by human could be used to guide their future decision making. 4) The influence information of training data points (functioned as fact-checking) on predictions can benefit trust. 5) Personality traits affect trust in predictive decision making differently under different cognitive load levels. Our on-going work is focusing on refined quantification of human trust, and its implication on perception and decision making in the human-machine collaboration context.

List of Publications

[1] J. Zhou, Z. Li, H. Hu, K. Yu, F. Chen, Z. Li, and Y. Wang, "Effects of Influence on User Trust in Predictive Decision Making", CHI 2019 – LBW, 2019.
[2] J. Zhou and F. Chen, "Towards Trustworthy Human-AI Teaming under Uncertainty", IJCAI 2019 Workshop on Explainable AI (XAI), Macau, China, 2019.
[3] J. Zhou, H. Hu, Z. Li, K. Yu, and F. Chen, "Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking", International IFIP Cross Domain Conference for Machine Learning & Knowledge Extraction (CD-MAKE 2019), Canterbury, UK, August, 2019.
[4] J. Zhou, S. Luo, and F. Chen, "Effects of Personality Traits on User Trust in Predictive Decision Making", Journal on Multimodal User Interfaces, 2020.
[5] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen, "Detecting Community Depression Dynamics Due to COVID-19 Pandemic in Australia", IEEE Transactions on Computational Social Systems, Vol. 8, no. 4, 2021.

Papers to be submitted/under review:

[1] K. Yu, S. Yang, T. Lammers, and F. Chen, "Trust, Decisions and Allocation of Efforts in Human-AI Interaction", to be submitted to ACM Transactions on Interactive Intelligent Systems (TIIS).

Contents

1. Introduction

Our previous work, along with cognitive psychology research, have revealed that humans create a mental model of trust when they interact with complex systems to facilitate the usage of those systems. In the case of human-AI systems, a human would create their mental model and build their trust to the AI systems accordingly. A human's subjective experience of an AI system mostly resides on the inter-actions that have occurred and that's how trust is built incrementally. Our earlier study has shown that a system's failure has more impact on trust than its success. The mental trusting picture will help the human decide when and where to rely on the AI teammates in order to maximize task outcomes. However, unlike many other automated systems, AI systems are complicated due to their dependency on the training datasets, limited transparency in the decision making mechanisms, and varied performance in different contexts. How human trust could be shaped in different ways to accommodate these complications and uncertainties remains unclear.

According to the pioneering work conducted by Lee and See, trust is defined as a human's attitude towards an agent or a system which human believes that the agent or system could be able to help achieve the human's goals given the uncertainty and vulnerability. This definition reveals that trust is closely related to uncertainty when humans interact with AI systems. Furthermore, trust roots in the human's mind and is supposed to be related to many internal and external factors including not only the individual's experience, but also the complexity of tasks and the interpretability of the systems. Recent research has further revealed that the one-way relationship of a human's trust to an AI teammate is dynamically affected by different factors and because of this complexity difficult to be accurately assessed. Haring et al. discuss the feasibility of applying a swift trust model, which was traditionally used for trust examination in human-human teams, to the study of human-robot trust. Interestingly, their work shows that swift trust can be associated with contemporary teams with limited collaboration experience, where a trusting relationship is established in a short term and affects collaboration. Along this line, Hoffman et al. argue that the trust dynamics in humans and in automation are qualitatively similar. Meanwhile, as the current AI algorithms become more sophisticated, it is very difficult for humans to understand the technical mechanisms of their AI teammates in a short time frame. Recent studies have shown that after some trials or interactions, humans can achieve a good understanding of the performance of their AI teammate, and further adjust their trust and optimize decision making strategies. However, knowledge on how these adjustments are made during the human-AI interaction process is still very limited.

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

- The implication of trust on human decisions and how they would allocate efforts between AI and themselves.
- Effects of personality traits on user trust in human-machine collaborations.
- Effects of influence on user trust in predictive decision making.

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

2. Trust, Decisions and Allocation of Efforts in Human-AI Interaction

A typical case of human-AI collaboration is the allocation of tasks between human and AI, in which the strategy for task allocation is to a great extent dependent on the subjective trust on the AI. In this study we designed a human-AI collaborative task for image annotation in a way that the AI could conduct it with high efficiency while the human could conduct it with high accuracy. Using this experiment, we aim to identify the trust dynamics of humans, their progress of perceiving and understanding the AI, and how they apply task allocation strategies based on their prior experience.

2.1 Experiment

An experiment involving humans and AI collaboratively annotating images was designed to induce trust dynamics and varied human decisions accordingly. Adding to previous studies, we intentionally involved subsystems in the AI decision support system as another dimensionality in the experimental design, so that diversified AI performance could be introduced even in the same task condition. Via examining the human perceptions, we expect to see whether people will be able to notice the sub-system differences besides the overall performance of the AI. The image annotation task is considered to be a convenient means to manipulate the system parameters, track the human responses, as well as incorporate different subsystems of varied performance. In addition, image annotation services with AI technique involvement have been becoming increasingly popular and hence the insights of this research will benefit practical human-machine collaboration directly.

The experiment design simulated a human-AI collaboration scenario where human and AI work together to annotate a set of images, i.e. to classify the images into different categories. Four categories of images were involved in the dataset for annotation, including flower, bird, dog and berry. The participants were asked to work with their AI teammates to complete the annotation for a set of images with the goal of maximizing their earnings from the image annotation task. The AI image annotation systems were designed with predefined overall accuracy in image annotation, but with a variance of accuracy on different image categories in order to simulate more closely the complexity of AI systems in the real-world. The experiment process involved two phases, Phase I for AI trials and Phase II for task allocation between human and AI. During Phase I for AI trials, a participant could use different images to test the performance of the AI. In Phase II, the participant would need to determine which images will be annotated by human or AI.

2.2 Procedure

The images used for the annotation task were selected from Linnaeus 5 . The images belong to four categories of labeled images – bird, dog, berry and flower. 400 images were chosen from each category for the experiment, resulting in an image pool of 1,600 images. The experiment was conducted in a university laboratory with a developed graphical user interface and was organized in blocks of trials. As mentioned earlier, the tasks in the experiment were segmented into two phases.

In Phase I, 60 images are shown on the left panel of the GUI (Fig. 1). Each time the participants pick one image from the 60 images, the AI will generate an annotation result, i.e. providing one of the four labels (berry, dog, bird or flower). Based on the annotation outcome, the participant will be able to assess the performance of the AI system. The selected image will be replaced by another image of the same category after the AI annotation result is shown.

To visually facilitate the participants' judgment, the output texts from the AI systems will be shown in green if an annotation by the AI is correct, an in red in case of an incorrect annotation. Fig. 1 illustrates the interface for users to test their AI team-mate. During the trial process, users were asked to rate their perceived accuracy (from 1% to 100%) and their trust (7-point Likert scale, in which 1 corresponds to the lowest trust level while 7 corresponds to the highest trust level) in the AI systems.

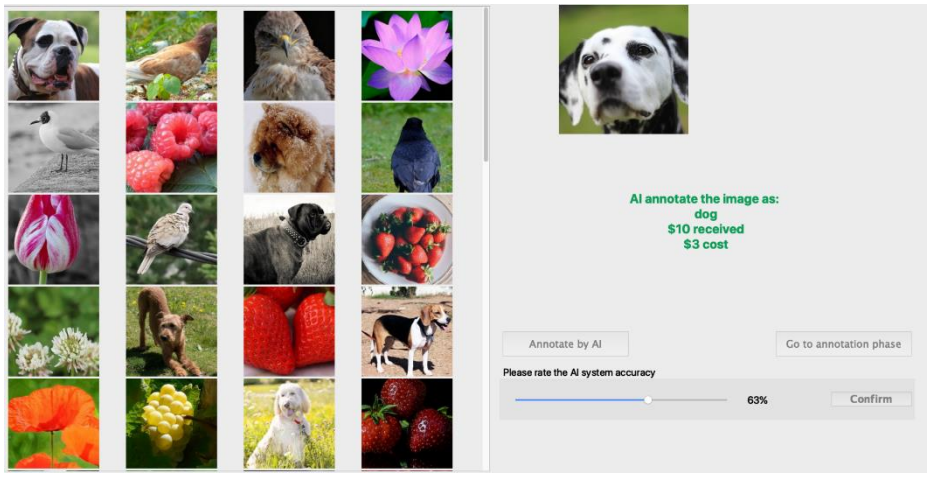

Figure. 1. The interface for participants to test the performance of an AI annotation system in Phase I. The left panel of the interface displays 60 images from four categories. After choosing an image for the AI to annotate, the annotation result of AI is shown on the right panel. The participants need to provide their perceived system accuracy and trust rating after the annotation result is shown.

At least 20 trial tasks need to be conducted before the participants proceed to Phase II of the experiment, the image annotation assignment (shown in Fig. 2). In this phase, the participants are requested to annotate 60 images. The participants can assign some image annotation tasks to the corresponding AI, or conduct the annotations themselves.

Similar to existing human-machine trust investigation experiments by Hoff & Bashir, we introduce virtual cost and gain for individual task outcomes: for each correctly annotated image, a virtual reward of $10 will be received. If the annotation is conducted by the AI, the cost will be $3 per image. However, the AI, as experienced by the participant, might make mistakes. In our setting, the humans are not expected to make mistakes in the annotation task. Their annotation cost will be higher than that of the AI at $5 per image. The arrangement for cost and gain aims to encourage the usage of AI in the annotation process, however when the AI is not reliable the humans can rely on themselves.

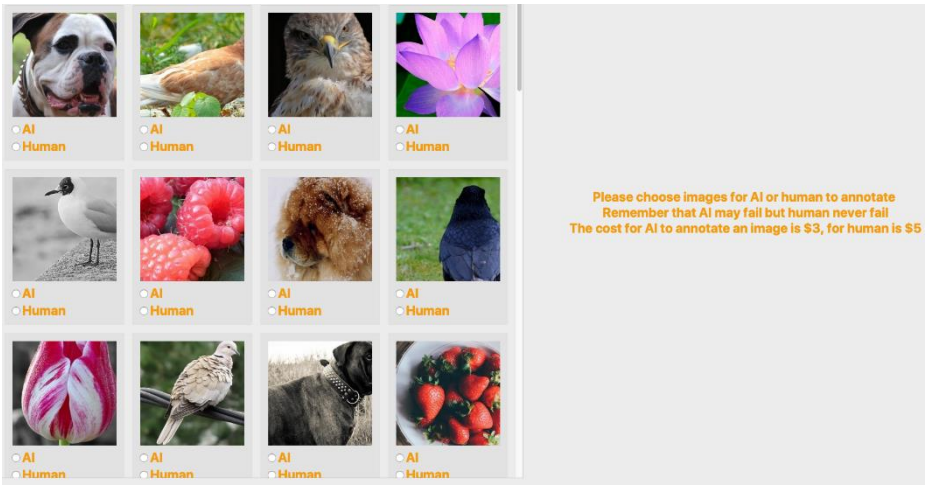

Figure. 2. Phase II interface for participants to allocate the image annotation task between human and AI. Once all the images have been assigned, an overall annotation outcome will be shown on the right panel.

2.2.1 Block Assignment

Each participant was asked to work with eight different AI systems on image annotation tasks. To simulate the complexity of AI systems in realistic tasks, the image an-notation

system was designed to involve four different subsystems, each being re-sponsible for one category of image annotation. We set different overall and subsystem performances.
Table 1 shows the AI systems with various overall accuracy and subsystem accuracy for respective image categories. For example, system 1 and system 2 have the same overall accuracy, but with different accuracies when annotating bird and dog images. We did not involve systems or subsystems with accuracies lower than 50%, which, according to existing examinations, would be easily distrusted by humans. In the experiment, the participants worked with the annotation systems in a randomized order.

2.2.2 Data Collection and Processing

For each participant, we collected multiple categories of data from the human-AI teaming interactions, as listed below:

- The sequence of images picked for test (e.g. flower, dog, dog, dog, berry…)
- AI system annotation outcome (correct or wrong)
- Human perceived AI accuracy (0% to 100%)
- Human subjective trust (1 to 7)
- The images users selected to annotate themselves

Similar to our previous work, the trust levels are normalized to the [0, 1] range for each participant.

2.2.3 Participants

In total, 25 university students and IT professionals participated in this experiment including 6 females and 19 males. After considering violations to the experimental process and unregistered data, we finally confirmed completeness and legitimacy of the data from 20 participants. On average, the participants spent 45 minutes to finish the experiment. There was no specific knowledge background or preparation required to complete the experiment. Ethics approval was acquired for the experiment conducted.

2.3 Results

Based on the data collected, our focus of investigation is the trust dynamics and corresponding decision-making patterns demonstrated by the participants during the experiment.

2.3.1 Trust Dynamics

Based on the normalized trust records, the dynamics of human trust in Phase I of the experiment is illustrated in Fig. 3. Although the initial trust was at similar levels for the different systems, as the trials continued, the human trust demonstrated a significant difference between the different AI systems towards the end of the trials based on an ANOVA examination ($F(3,76) = 145.48$, $p < 0.05$). The users' trust for AI systems with 80% and 70% accuracy showed a similar pattern with a t-test examination ($t = 3.92$, $p > 0.05$). But trust differed significantly between the 90% and 80% AI systems ($t = 11.91$, $p < 0.05$). Similarly, the human perceived system accuracy showed similar patterns to the human trust. A significant difference in the perception of accuracy was observed between the four AI systems ($F(3,76)=147.12$, $p<0.05$). The results in our experiments are consistent with our earlier findings.

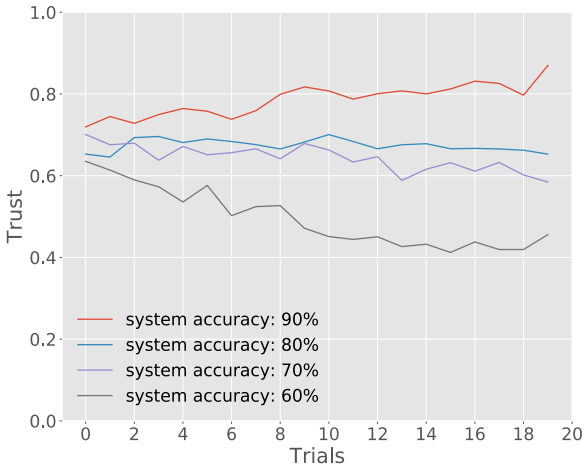

Figure. 1. The dynamics of human trust on different AI systems.

As a refinement of the investigation, the human trust in the different subsystems is shown in Fig. 4. At the sub-system level, a decreasing trend of trust could be identified as subsystem performances decrease.

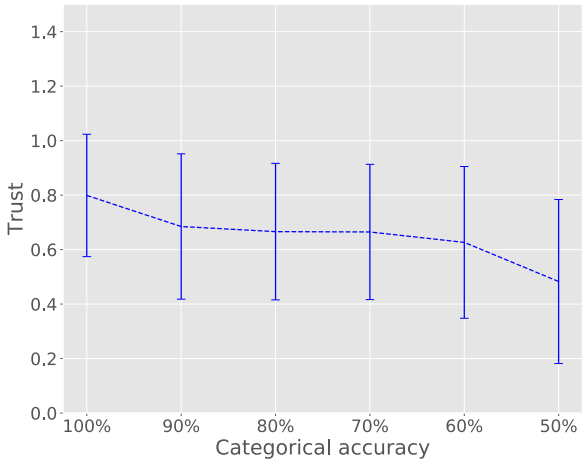

Figure. 2. Human trust in the AI subsystems with different accuracies.

2.3.2 Trust and Perception Adjustment

The examination above has shown that AI system performances affect trust and subjective perceptions. In our next examination, we will reveal how the trust and perceived system accuracy is adjusted specifically along with the trials.

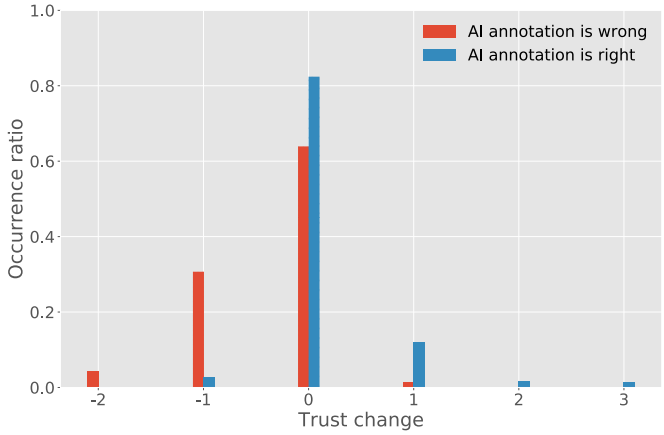

Figure. 3. Human trust change when the AI systems make correct/incorrect annotations.

Implications of the AI systems' performance on human trust change is shown in Fig. 5. Overall, when systems make a correct annotation, the human trust value would remain stable or increase slightly, while when an incorrect annotation was made by the system, the trust level would remain stable or be reduced. Comparing these two cases, we find there is a higher chance that the trust rating is affected more by the incorrect AI annotations than by the correct ones. This means that a mistake made by AI is more likely to lower the trust level than a correct annotation to increase the trust level.

A similar pattern could be observed in terms of human adjustments of perceived accuracy, as shown in Fig. 6. The perception is more affected by negative AI system performance than positive performance, although in most cases, minimal adjustments to the perception are observed.

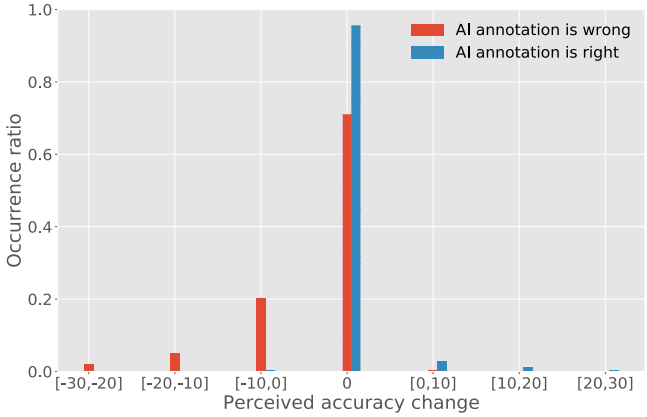

Figure. 4. The change of human perception with AI system making correct/incorrect annotations

2.3.3 User Strategy to Test AI

We also investigate the users' strategies to work with the AI. In particular, we investigate the users' response, when the AI fails or succeeds in annotating one specific category of image. This is examined via the length of consecutively tested images of the same category. For example, when the user tested three dog images in a row and a berry image afterwards, if the first image was annotated incorrectly by the AI, it would be counted as a sequence length of 3 beginning with a wrong AI annotation. As shown in Fig. 7. After the AI made a correct annotation, for around 65% of the cases an image from another category was chosen for the next trial. However, when AI made incorrect annotations, participants conducted more consecutive trials on the same category of images.

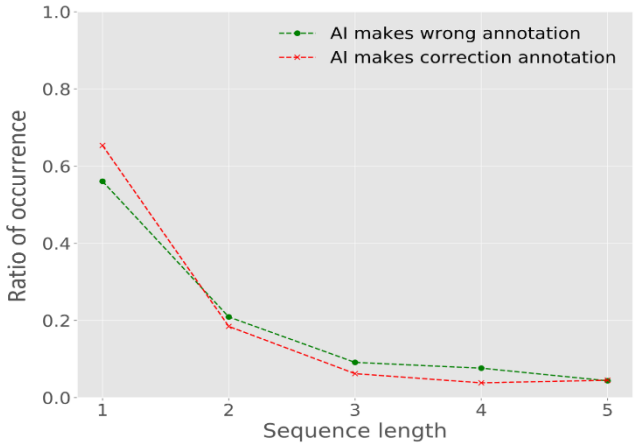

Figure. 5. Human tested image sequence length within the same image categories.

As illustrated in Fig. 8, the users adopted interesting trial strategies when AI makes correct or incorrect annotations in Phase I. For subsystems of different accuracies, although they are mixed together in the trials, the trend is still visible that fewer images were tested for the more accurate subsystems, while more images were tested for the less accurate systems.

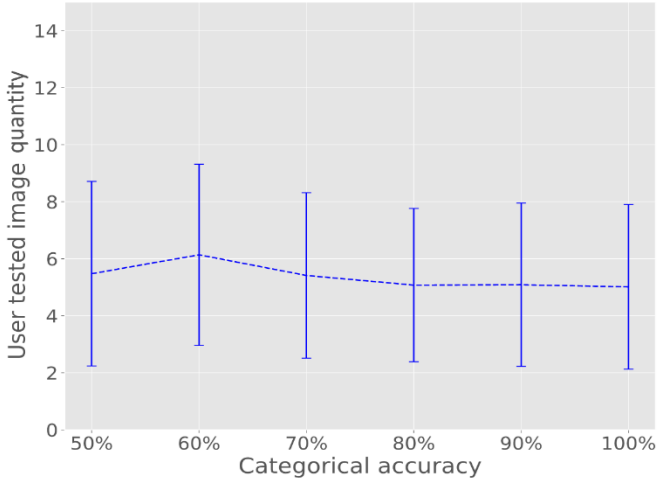

Figure. 6. Quantity of user tested images in different categorical accuracy.

2.3.4 Task Assignment between User and AI

As Phase II of the experiment, the participants were asked to determine which images they would either like to annotate themselves or assign to the AI instead. Fig. 9 presents the overall users' decisions with regard to the different categorical accuracy of the AI subsystems. It could be observed that more users tend to annotate images by themselves if the AI teammates didn't perform well in Phase I. Interestingly, even if the sub-systems were mixed together in the experiment, the overall trend for users to make decisions in the task assignments is still identifiable: when AI subsystem accuracy is 50%, the users would annotate the highest number of images by themselves even though the cost of human annotation is higher. When the AI subsystem accuracy increases, the general trend is that users would annotate fewer images by them-selves and assign more images to the AI systems. Comparing the results from Fig. 9 and Fig. 8, a linear correlation could be identified ($r = 2.84$, $p < 0.05$).

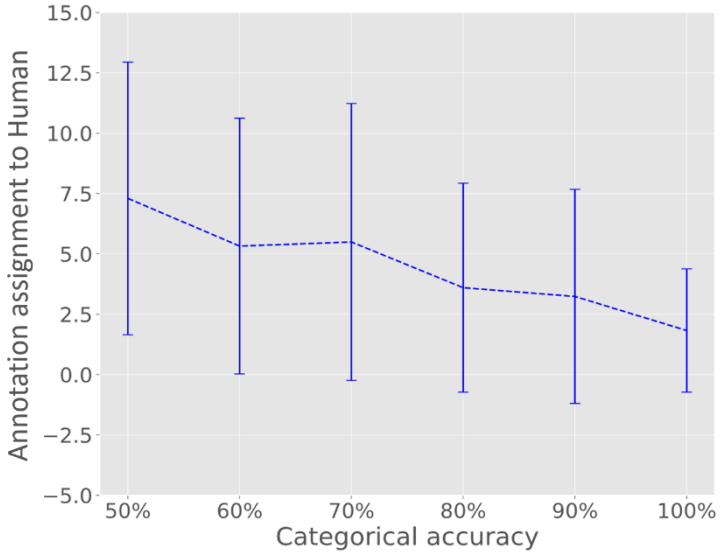

Figure. 7. Quantity of images assigned to human for annotation in Phase II, with respect to subsystems of different categorical accuracy.

2.4 Discussions

The general aim of this study is to identify the relationship between human perception, trust and task allocation strategies in a human-AI collaboration environment. Consistent with the findings in our prior work., the trust levels correlated with the accuracy of the respective AI systems as the user trials proceeded, This validates the experimental design as a feasible means to induce varied levels of trust in human-AI collaboration. Human perception and trust change in different ways when the AI systems made correct or incorrect annotations. Overall, correct annotations from the systems help to increase human trust and perceived accuracy, while incorrect annotations from the system result in decreased trust and accuracy perception, which is consistent with our intuition. However, we have observed that for many cases trust remained stable irrespective of the AI performance.

We found two factors contributing to this observation. First, if the AI system was highly trusted, the trusting level could reach the top level, and hence high system performance would not result in further increased trust because of the limit of the trust scale. The same rule applies to the least trusted system, on which the rated trust level could not decrease further. Secondly, the outcome of a single trial may not be able to change the trust level, especially if the outcome is positive. However, the different implications of correct and incorrect system performances should be noticed for trust change: an incorrect AI annotation has a higher chance to lead users to conduct more trials on the same subsystem. In comparison, when a correct AI annotation is encountered, the users will more likely switch to another category of images. In that sense, AI system failures will affect the time and trials to explore other sub-systems involved in the AI.

Using the different categories of images involved in the experiment design, the investigation was fine-tuned to examine the different levels of human perception. One interesting finding is that humans are not only able to perceive the different performances at a whole system level, but are also capable of identifying the detailed an-notation performances for individual image categories at subsystem level and adjusting their trust accordingly. It should be noted that during the experiment, the participants were never told that different performances may occur for different image categories within a given AI system. However, many of them were able to identify this pattern and used it to adjust their trust on the system performance on individual categories of images, and support their decision making.

For a subsystem with higher accuracy, the number of images tested will be lower, which

implies that participants spend less time examining the performance of the better subsystems. However, for the less accurate subsystems, more trials were conducted for a better understanding of the performance. This implies that the users apply systematic sampling theory in their trials with their specific adjustments, i.e. to focus on those less accurate subsystems while considering the over-all performance of the whole system.

As the ultimate aim of this study, we identify how people make decisions based on their prior experience. The understanding of how the results are produced by techno-logical systems affects the trust in those systems by the users. A positive perception of the system's performance enhances fuller use and reliance on the systems by the users. Furthermore, the correlation of annotated image quantity between the trial phase and annotation task allocation phase implied that prior experience and knowledge at a refined subsystem level guides future decision making. As a consequence, the participants assigned more tasks to the more accurate AI subsystems, while leaving more tasks for themselves if the AI subsystem's performance was considered low. Obviously, this can be seen as an optimal strategy that could help the human-AI team to gain the best performance. Although we endeavor to conduct quantitative examinations for the trust dynamics, user trials and task allocations, there are a few limitations that should be mentioned. Firstly, we used system performance as a single factor to induce different trust levels and user responses. In practice there could be many other factors affecting the trust and user decisions, and the user might be able to get information from multiple channels besides observing AI performance. As a consequence, user trust and decision making could be the outcome of multiple factors, although if any factor is quantifiable, methods similar to the one applied in this research could be applied. For instance, we didn't take into consideration system characteristics such as convenience and fairness, whose impact should be examined in future work. Secondly, in each image annotation AI, four different sub-systems are involved, each with distinguishable performance. In realis-tic practice the structure and performance of AI could be much more complicated, and their performances could be unstable and hence the trust dynamics could be even more difficult to characterize. However, we believe the current research sheds light on the design of practical AI systems, especially when different components of such systems are examined individually or at different time scales. Finally, in our experiment the trust was shaped by short-term interactions between human and AI, which has been noted as swift trust in existing literatures. Although in our earlier work it has been observed that the trusting level would converge within 30 trials, more investigations are expected to con-firm the generalizability of current findings to other human-AI collaboration contexts.

Overall this study revealed a number of key findings that should be considered in interaction system design and analytics. First, the design of AI systems should take into consideration the user's perception process, and some prior trials would help the user to make better decisions. Secondly, system failures are the main driver to shape user's trust change. Whenever a system error occurred, users used more trials to con-firm if the AI subsystem was unreliable, which is a typical example of systematic sampling theory in practice. This study also suggests that users are able to perceive the AI performance at both system and subsystem level, and use the different levels of knowledge to guide future task allocation. This finding implies that the observation of user decision behaviors can be helpful for AI system diagnosis.

3. Trust in Predictive Decision Making

ML technologies face prolonged challenges with low user acceptance as well as seeing system misuse, disuse, or even failure. These fundamental challenges can be attributed to the nature of the "black-box" of ML for domain experts when offering ML-based solutions. As a result, recent research suggests model explanation as a remedy for the "black-box" ML. Taking the influence of training data points on predictions as an example, the explanation with influence allows to capture the weight/contribution of each training data point on the

prediction. However, these explanations are highly biased towards ML experts' views, while domain users are more interested in what influence information affect and how these influence information are presented to them to boost their trust in predictions. Besides explanation, the ability to provide justifiable and reliable evidences for ML-based decisions would increase the trust of users.

Fact-checking, which provides "evaluation of verifiable claims made in public statements through investigation of primary and secondary sources", is increasingly used to check and debunk online information because of credibility challenges. Motivated by this, our work introduces fact-checking into Machine Learning (ML) explanation by referring training data points as facts to users to boost user trust. These training data points are selected based on their influence values on predictions. We aim to investigate what influence of training data points and model performance, and how they affect user trust in order to enhance ML explanation and boost user trust. We tackle this question by allowing users check the training data points that have the higher influence and the lower influence on the prediction.

3.1 Experiment

We present a visualization approach called fact-checking visualization for presenting multiple data attributes based on parallel coordinates as shown in Figure 10. It demonstrates how similar the training pipes are with the testing pipe in red color. The pipe attributes visualized in Figure 1 include pipe size (diameter), pipe length, pipe age, failure times during the observation period, and whether it was failed in the checked year (0 means not failed and 1 means failed).

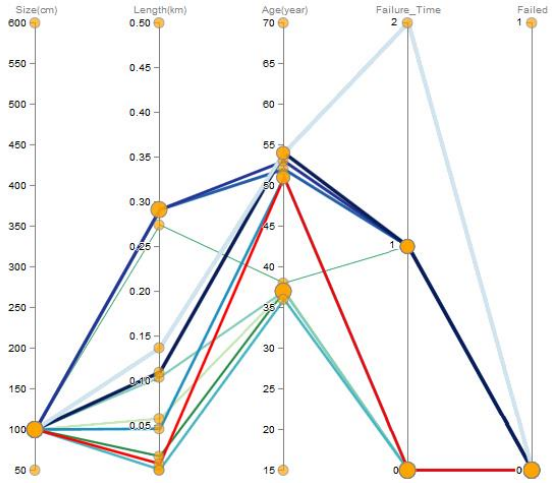

Figure 10: Fact-checking visualization.

We present a framework of fact-checking for boosting user trust in a predictive decision making scenario (see Figure 11). A user study was conducted to investigate the effectiveness of fact-checking in boosting user trust.

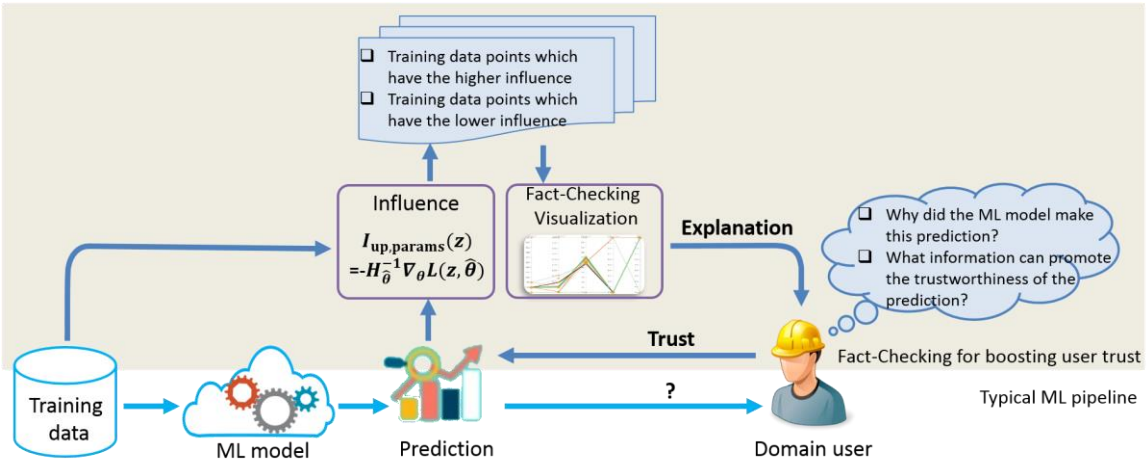

Figure 11: A Framework of Fact-Checking for Boosting User Trust.

3.2 Analytics result

Influence and Trust: Figure 12 shows mean normalized trust values over different influence settings under high model performance. Friedman's test gave statistically significant differences in trust among four influence conditions, $\chi^2(3) = 21.675$, $p = .000$. Then post-hoc Wilcoxon tests (with a Bonferroni correction under a significance level set at $p < .013$ (0.05/4)) was applied to find pair-wise differences between influence conditions. The results suggest that the presentation of influence of training data points on predictions significantly increases the user trust in predictions, but only for training data points with higher influence values under the high model performance condition.

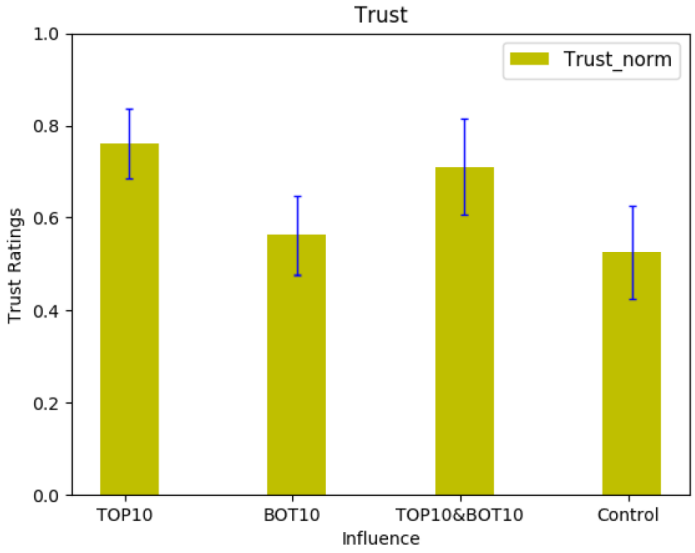

Figure 12: Trust levels by influence under high model performance.

Model Performance and Trust: Figure 13 shows mean normalized trust values over two model performance conditions (high and low) under different influence settings. The results suggest that high model performance together with influence information result in the higher user trust in predictions.

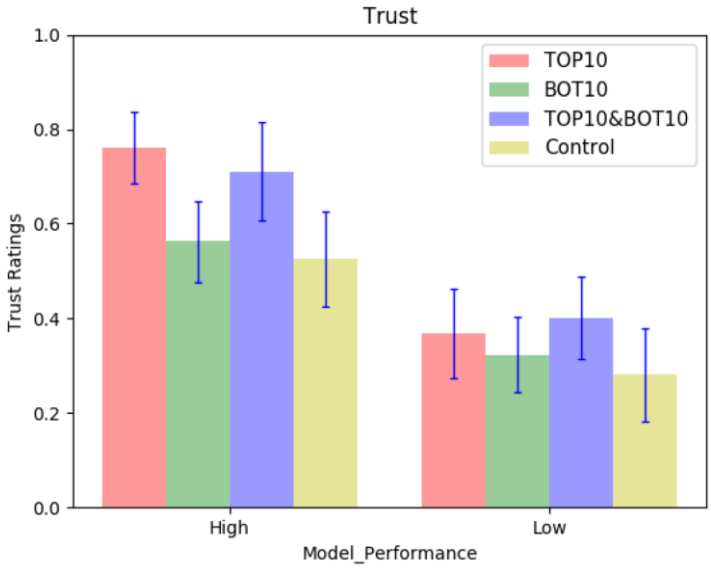

Figure 13: Normalized trust by model performance under different influences.

Overall, we can say that the influence information of training data points (functioned as fact-checking) on predictions can benefit trust, where users can check the facts that are similar to the testing data point. Presentation of influence information of training data points having the higher influence values can lead to increased trust but only under the high ML model performance condition, where users can justify the action with more similar facts and fit their general understanding of the problem.

4. Personality Traits and Trust

Recent studies showed that individual differences in personality traits contribute to differences in trust. For example, a probability model is proposed to examine the effect of personality traits on trust in automation. While predictive decision making involves much human's cognitive effort, understanding the effects of different personality traits on user trust will help to design more effective personalized intelligent user interface for human–machine collaborations. However, little work is done on the effects of personality traits on trust in predictive decision making especially under uncertainty and cognitive load conditions.

Since the development of trust is affected by an interplay of characteristics of human, machine, and operational environment, Hancock et al. and Schaefer et al. proposed a conceptual organization of trust influences highlighting crucial influence factors in trust development. We adapt this conceptual organization into the predictive decision making scenario and examines the effects of human, machine, and environment factors on trust in human–machine collaborations. We specifically focus on the investigation of personality traits and uncertainty of ML models as key human and machine factors respectively in predictive decision making. Furthermore, cognitive load is introduced from a second cognitive task in our predictive decision making scenario and it is used as an environment factor to examine how these three factors affect trust in predictive decision making.

4.1 Research method

The Big Five personality model is used to identify personality traits of users. Two uncertainty types of risk and ambiguity are presented with predictive model results in a decision making scenario. This follows the user method and approach to design a cognitive system which uses a simulation of water pipe failure prediction as a case study. It shows that different personality traits affect user trust in predictive decision making differently under both uncertainty presentation and cognitive load levels. A framework of user trust feedback loop

is proposed to incorporate study results into human–machine collaborations (Figure 14).

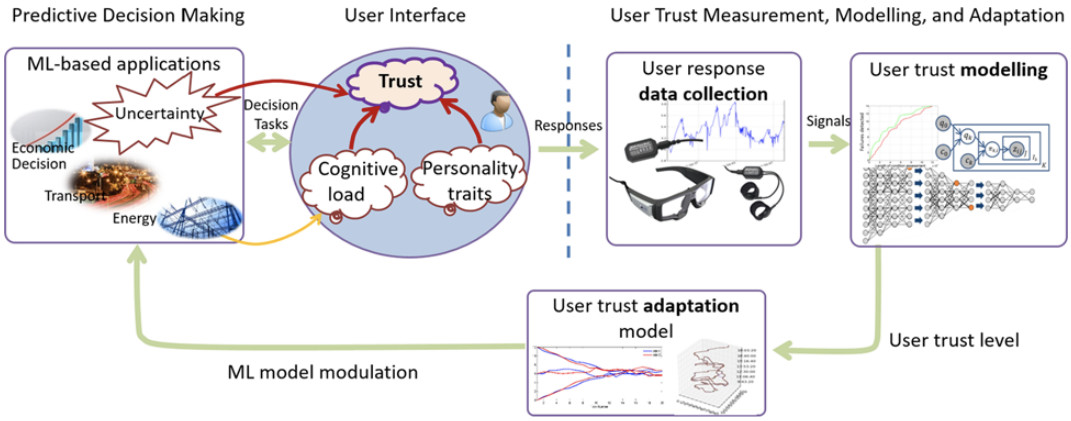

Figure 14: Framework of user trust feedback loop in predictive decision making in human–machine collaborations.

An experiment with the use of water pipe failure prediction as a case study to investigate the effects of personality traits on user trust (Figure 15).

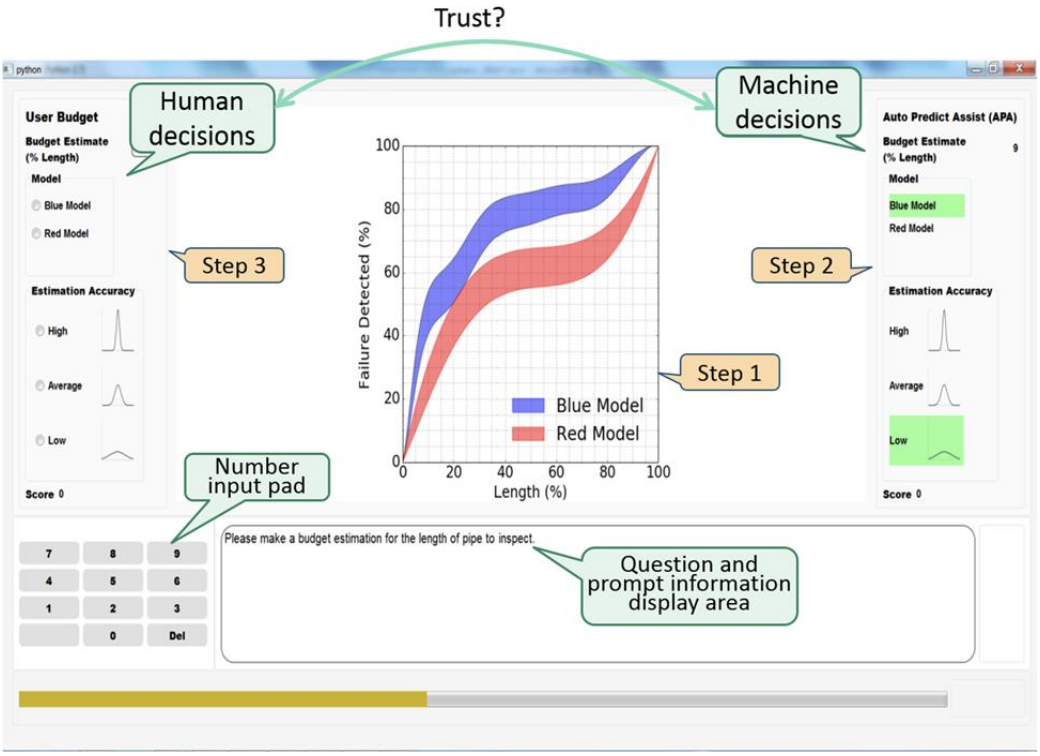

Figure 15: Screenshot of a task performed in the study.

4.2 Analytics result

The experimental results show that uncertainty presentation can lead to the increased trust but only under low cognitive load conditions when users have sufficient cognitive resources to process the information. Presentation of uncertainty under high cognitive load conditions, when cognitive resources are short in supply, can lead to lowering trust in the system and its recommendations. Furthermore, different personality traits affect trust differently under both uncertainty and cognitive load conditions. For predictive decision making tasks with different cognitive load requirements and uncertainty conditions, users should be appointed according to their personality traits. The results of this work provide guidelines for such appointment. For example, under low cognitive load with ambiguity uncertainty, people with low

Agreeableness, low Neuroticism, high Extraversion, high Conscientiousness, and high Openness should be appointed to conduct predictive decision making tasks, while under low cognitive load with risk uncertainty, people with high Neuroticism, low Extraversion, high Conscientiousness and low Openness should be appointed to conduct predictive decision making tasks. Furthermore, under high cognitive load situations, people with high Neuroticism and low Extraversion should be recruited to conduct predictive decision making without uncertainty presentations.

The study suggests that personality traits affect trust in predictive decision making differently under different cognitive load levels. Furthermore, personality traits under different uncertainty types and cognitive load levels showed different user trust perceptions. Therefore, the approaches for the improved trust need to consider both cognitive requirements by tasks and users' personality traits at the same time, e.g. users with high Neuroticism and low Extraversion may be assigned to conduct predictive decision making tasks under highly critical situations without uncertainty presentation.

5. Conclusion

In this project, we have investigated the interplay between human trust, perception and decision making when collaborating with an AI annotation system under different uncertainty conditions. The different studies have revealed the implication of AI performance, especially system errors, on human's trust adjustment and decision making, and how different factors including personality traits will affect the human-AI teamworking. We have also proposed the framework of trust feedback loop to integrate these findings into participatory design in human–machine collaborations. Our findings fill a significant gap in trust in predictive decision making with the introduction of personality traits into ML-based solutions. This project has laid a foundation towards better understanding of human-AI trusting relationship in a collaborative team, and provides important findings that should be considered in future intelligent AI system design.