Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making

Jianlong Zhou(✉), Syed Z. Arshad, Simon Luo, and Fang Chen

DATA61, CSIRO, 13 Garden Street, Eveleigh, NSW 2015, Australia
{jianlong.zhou, syed.arshad, simon.luo, fang.chen}@data61.csiro.au

Abstract. Rapid increase of data in different fields has been resulting in wide applications of Machine Learning (ML) based intelligent systems in predictive decision making scenarios. Unfortunately, these systems appear like a 'black-box' to users due to their complex working mechanisms and therefore significantly affect the user's trust in human-machine interactions. This is partly due to the tightly coupled uncertainty inherent in the ML models that underlie the predictive decision making recommendations. Furthermore, when such analytics-driven intelligent systems are used in modern complex high-risk domains (such as aviation) - user decisions, in addition to trust, are also influenced by higher levels of cognitive load. This paper investigates effects of uncertainty and cognitive load on user trust in predictive decision making in order to design effective user interfaces for such ML-based intelligent systems. Our user study of 42 subjects in a repeated factorial design experiment found that both uncertainty types (risk and ambiguity) and cognitive workload levels affected user trust in predictive decision making. 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.

Keywords: Trust · Uncertainty · Cognitive load · Predictive decision making

1 Introduction

Trust has been found to be a critical factor driving human behavior in human-machine interactions with autonomous systems [1] and more recently in modern complex high-risk domains such as aviation and military command and control [2]. It is also one of the most important factors in management and organizational behavior for all personal and business decision making as well as for efficiency and task performance [3, 4]. Trust is influenced by the types and format of information received by humans, their individual approaches to develop and determine trust, and aspects such as system capability and reliability [5].

Various definitions of trust have been used. One of the most widely cited definition of trust is from Lee and See [6], which defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and

R. Bernhaupt et al. (Eds.): INTERACT 2017, Part IV, LNCS 10516, pp. 23–39, 2017.
DOI: 10.1007/978-3-319-68059-0_2

vulnerability". This definition shows that uncertainty is tightly coupled to trust. Uncertainty indicates it is impossible to determine whether the information available is true or not. There are many variants of uncertainty. The term could refer to statistical variability, noise in the information, nondeterministic relationship between action and consequences, or the psychological reaction to difficult problems. In human-machine interactions, uncertainty often plays an important role in hindering the sense-making process and conducting tasks: on the machine side, uncertainty builds up from the system itself; on the human side, these uncertainties often result in "lack of knowledge or trust" or "over-trust". Such human's biased interpretation can be partially resolved if we can make uncertainty transparent to users. Furthermore, system transparency is regarded as one vital aspect in maintaining human's trust in and reliance on autonomous systems [7, 8]. A user might be risking too much by completely ignoring uncertainties and having complete faith in autonomous systems. On the other hand, trivializing autonomous systems or having high uncertainty perception on autonomous systems could possibly dismiss the incredible potential of autonomous systems. Adobor [9] showed that a certain amount of uncertainty is necessary for trust to emerge. Beyond that threshold, however, increase in uncertainty can lead to a reduction in trust. This midrange proposition suggests that there may be an optimal balance between uncertainty and trust.

Moreover, Parasuraman et al. [10] showed that human cognition constructs such as Cognitive Load (CL) and trust are often invoked in considerations of function allocation and the design of automated systems. For example, in task situations of modern complex high-risk domains, users often need to make decisions in limited time. Therefore, they often make decisions under high cognitive load besides trust issues in such task situations. It was found that a higher cognitive load worsens the situation in relation to trust building [11]. However, it is still not clear how trust varies under both high cognitive load and various uncertainty conditions.

Next we look at decision making, which is now an important research topic in HCI with the fast growing use of intelligent systems [12]. Rapidly increasing data in fields such as finance, infrastructure and society has motivated users to try integrating "Big Data" and advanced analytics into business operations - in order to become more analytics-driven in their decision making. Much of machine learning (ML) research is inspired by such expectations. As a result, we continuously find ourselves coming across ML-based appealing viewgraphs and other predictions that seem to work (or have worked) surprisingly well in practical scenarios (e.g. AlphaGO's beating with professional GO players in 2016 and 2017). So far these machine learning success stories originate from ML technical experts or computing professionals (e.g. Google DeepMind). For many of non-ML users, ML-based predictive analytics software is like a "black box", to which they simply provide their source data and (after selecting some menu options on screen) colorful viewgraphs and/or recommendations are displayed as output. The "black box" approach has obvious drawbacks: it is difficult for the user to understand the complicated ML models [13, 14]. It is neither clear nor well understood that how trustworthy is this output, or how uncertainties are handled by underlying algorithmic procedures. As a result, the user is more or less unconfident in the ML model output when making decisions based on the ML model output and thus also unconfident in the ML models themselves.

From this perspective, Winkler [15] emphasized the importance of communicating uncertainties in predictions (as imprecision and uncertainty are unavoidable in predictive analytics). He believed that the consideration of uncertainty is greatly necessary in making rational decisions. It was also found that the presentation of automation uncertainty information helped the automation system receive higher trust ratings and increase acceptance of the system [16]. This display might improve the acceptance of fallible systems and further enhances human–automation cooperation. However, it remains unclear whether different types of uncertainty (e.g. risk and ambiguity) affect trust building, and if yes how they affect trust building, especially in predictive decision making. Here risk refers to situations with a known distribution of possible outcomes, and ambiguity is the situation where outcomes have unknown probabilities.

This paper aims to investigate the effects of uncertainty on user trust under various cognitive load levels in predictive decision making. 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 cognitive systems, as reviewed by Candello [17], and used in [18]. This user study was deployed as a simulation (derived from the case study) of water pipe failure history analysis for future pipe failure prediction. It shows that both uncertainty presentation and cognitive load levels affect user trust in predictive decision making. The investigation results can be used to design effective user interfaces for ML-based intelligent systems and improve the acceptability of ML techniques by users.

2 Related Work

The research in human-machine trust and similar cognitive engineering constructs has a rich history [10]. Several of the efforts in this area can be traced back to Rouse's [19] ideas about adaptive aiding, that later, among other things evolved into more advanced HCI techniques. Also that psychophysiology was proposed for adaptive automation [20] and then trust and self-confidence argued into adaptive automation [21].

Winkler [15] demonstrated, with the help of several effective examples (from different fields), that probabilities are needed to understand the risk associated with potential decisions as well as to determine measures such as expected payoffs and expected utilities. LeClerc and Joslyn [22] successfully demonstrated that adding a probabilistic uncertainty estimate in public weather forecasts improved both decision quality and compliance (to evacuation instructions in cases of severe weather threats). Uggirala et al. [23] studied humans using systems that include uncertainties by having the users rate their trust at each level through questionnaires. Their study showed that trust relates to competence and an inverse relation to uncertainty, meaning that an increase in uncertainty decreases trust in the systems.

Allen et al. [24] investigated the effects of communicating uncertainty information with users using different representations on cognitive tasks. Uncertainty information is typically presented to users visually, most commonly in graphical format [24, 25]. Edwards et al. [26] compared different graphical methods from presenting quantitative uncertainty in decision making tasks. The representation of uncertainty can have significant impact on human performance. It was shown that when the representation of uncertainty for a spatial task better matches the expert's preferred representation of the problem even a non-expert can show expert-like performance [27]. This is actually a very good motivation for trying to figure out the preferred representation of uncertainty by different user groups.

Decision making under uncertainty is widely investigated in decision theory [28], where uncertainty is usually considered as probabilities in utility functions. de Visser and Parasuraman [29] conducted two experiments to examine the effects of automation reliability and adaptive automation on human-system performance with different levels of task load by using a high-fidelity multi-UV (uninhabited vehicles) simulation involving both air and ground vehicles. User trust and self-confidence were higher and workload was lower for adaptive automation compared with the other conditions. It was found that human-robot teams can benefit from imperfect static automation even in high task load conditions and that adaptive automation can provide additional benefits in trust and workload.

However, little research has been done on the effects of uncertainty, especially different types of uncertainty such as risk and ambiguity uncertainty, on user trust in predictive decision making under various cognitive load levels. With the use of a case study of predictive decision making for the water pipe failure budget planning, this paper investigates user trust changes under variations of both uncertainty types and cognitive load levels. Two types of uncertainty presentations (risk and ambiguous) and four cognitive load levels are introduced in the study to learn their effects on user trust in predictive decision making.

3 Experiment

3.1 Experiment Data

This research used water pipe failure prediction as a case study for predictive decision making (replicated in lab environment). Water supply networks constitute one of the most crucial and valuable urban assets. The combination of growing populations and aging pipe networks requires water utilities to develop advanced risk management strategies in order to maintain their distribution systems in a financially viable way [30]. Pipes are characterized by different attributes, referred to as features, such as laid year, material, diameter size, etc. If pipe failure historical data is provided, future water pipe failure rate is predictable with respect to the inspected length of the water pipe network [30]. Such models are used by utility companies for budget planning and pipe maintenance. However, different models with various uncertainty conditions may be achievable resulting in different possible budget plans. The experiment is then set up to determine what uncertainty conditions may influence the user's trust during the decision process. The prediction models were simulated following the models such as Weibull and Hierarchical Beta Process (HBP) [30].

3.2 Experimental Data

In this study, models are simulated and based on different pipe features (e.g. size or laid year). The model performance curve was presented to let the participants evaluate different models. The model performance is the functional relationship between the inspected length of the network and the percentage of failures detected by the model. Figure 1 shows the performances of two sample models, where the "blue model" outperforms the "red model", because the former detects more failures than the latter for a given pipe length.

ML models are usually imperfect abstractions of reality. As a result, imprecision can occur in the prediction through model uncertainty. Model uncertainty here refers to an interval within which the true value of a measured quantity would lie. For example, in Fig. 2(a), in order to inspect 20% of the pipes in length, the uncertainty interval of the failure rate is $[46\%, 60\%]$ for the blue model, and about $[15\%, 25\%]$ for the red model: the red model is said to have less uncertainty in prediction than the blue model because the red model has smaller uncertainty interval than the blue model.

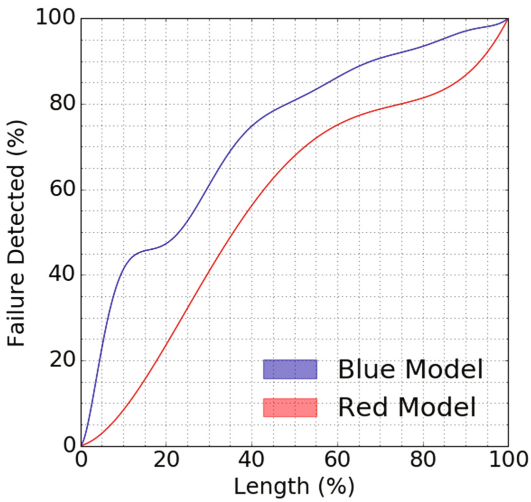

Fig. 1. Performance curves of ML models. (Color figure online)

Model uncertainty usually spans as a band in the model performance diagram as shown in Fig. 2. By considering model uncertainty, the relationship between two models may have two cases as shown in Fig. 2: non-overlapping models (see Fig. 2a), and overlapping models (see Fig. 2b). In Fig. 2b, the interval of the model with lower uncertainty is subsumed in the interval of the model with higher uncertainty, whereas in Fig. 2a, the two bands are disjoint.

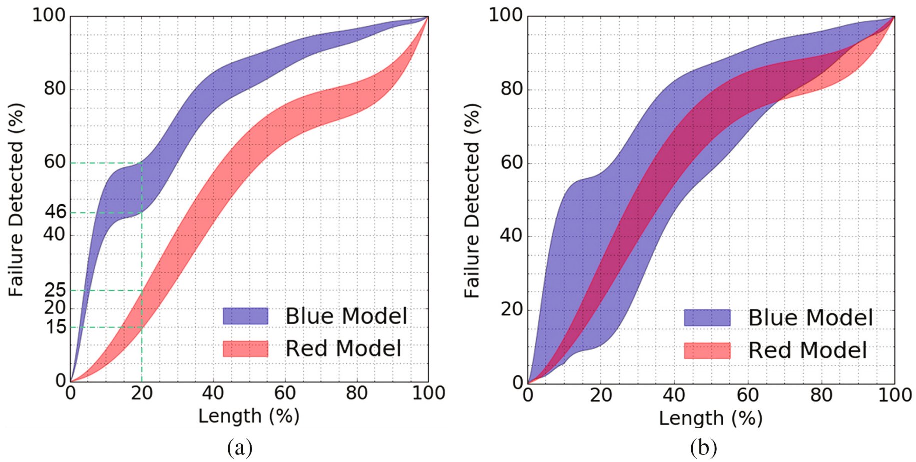

Fig. 2. Predictive models with uncertainty: (a) non-overlapping models, and (b) overlapping models. (Color figure online)

3.3 Task Design

According to the water pipe failure prediction framework, we investigated the decisions made by users under varied conditions. Each user was asked to make a budget plan, i.e. a budget in terms of network length to be inspected, using the failure prediction models learned from the historical pipe failure records. Two ML models were provided for each estimation task. Participants were required to make decisions by selecting one of two presented ML models and then making a budget estimate based on the selected ML model. The budget estimate needs to meet the following requirements:

- To inspect as short length of pipes as possible (low cost);
- To be as precise in budget estimate as possible (higher accuracy would reflect greater confidence in estimation).

In this study, a module named Automatic Predictive Assistant (APA) is introduced to the user as a new module ‘under testing’ phase. The APA is a simulated module which reads in the information provided by the ML models, and then recommends a typical decision (of average accuracy) for the participant. Users can choose to trust, modify or totally ignore the recommendations of APA. The participant needs to evaluate whether she trusts the estimation recommended by the APA. If she does not trust the APA, she is asked to provide her own estimation. Figure 3 shows the screenshot of a task performed in the study.

The actions a participant needed to do during a task session include answering questions to validate understanding of machine learning performance and uncertainty, and to validate cognitive load levels introduced, as well as making decisions according to information presented. In summary, each task is divided into following major steps:

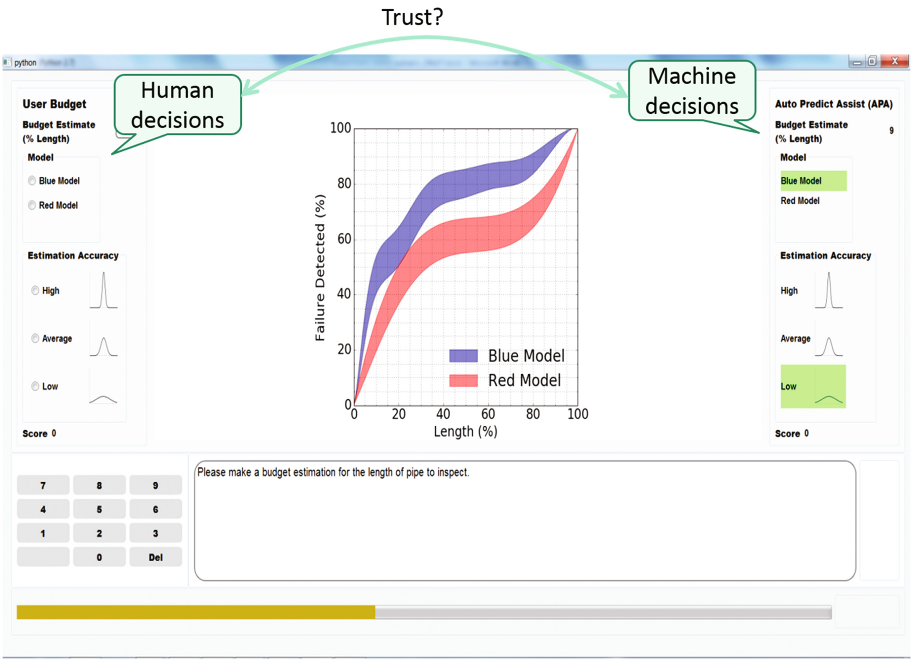

Fig. 3. Screenshot of a task performed in the study. (Color figure online)

(1) The participant is firstly asked to study the ML performance diagram (in the middle of the screenshot in Fig. 3) and answer questions on model performance and uncertainty to validate her understanding of the information presented.
(2) Next, the APA recommendations (at the right side of the screenshot in Fig. 3) are displayed and again the user understanding validated.
(3) Finally, the participant is required to estimate the budget by selecting an ML model and its estimation accuracy. If he/she does not trust the recommendations from the APA, he/she is required to provide his/her own estimations (at the left side of the screenshot in Fig. 3) based on the ML performance displayed at the first step. Subjective trust ratings are obtained immediately after this step.

Participants were encouraged to reach the best budget estimates they could as quick as possible.

In this study, cognitive load was introduced by asking participants to remember a random number digit sequence for the duration of task time and reciting it after the task. This dual-task load inducing technique is quite popular in decision making scenarios [31]. The cognitive load level was determined based on the number of digits being remembered. Four cognitive load levels were applied in this study – from low to high, the number of random digits to be remembered ranged from three, five, seven and nine. Three-digit number is for lowest load condition and nine-digit is for the highest load.

There were three different uncertainty visualizations (no uncertainty, non-overlapping uncertainty, and overlapping uncertainty). Each condition was performed under four different cognitive load levels. Each task was performed for three rounds. All together 36 estimation tasks (3 uncertainty conditions × 4 cognitive load levels × 3 rounds) were conducted by each subject. Three additional training tasks were also conducted by each subject before the formal tasks. The order of tasks was randomized during the experiment to avoid any bias. Slides-based instructions on the concepts of predictive models and uncertainty as well as predictive decision making were presented to each participant before the task time.

3.4 Participants and Apparatus

Forty-two participants were recruited from three groups with different background, with the ages ranging from 20 to 57 years: Fourteen participants were ML researchers (experts in ML or data mining research), nineteen were non-ML researchers (participants who were researchers but not in ML or data mining), and nine administrative staff. These three user groups constituted the majority of the users for this particular type of predictive decision making scenario we considered. All were requested to make predictive decisions (using historical data visualized on screen) about the optimal length of pipe (thus budget estimation) to be checked in order to minimize water pipe failures. Information was presented on a 21-inch Dell widescreen monitor. Figure 3 presents a screenshot of a task performed in the study.

3.5 Data Collection

After each decision making task, participants were asked to rate their trust in APA recommendations and also confidence in their own decisions (using a 9-point Likert scale where 1 = least trust, and 9 = most trust). Besides trust subjective ratings, cognitive load rankings for each task from subjects were also collected using a 9-point Likert scale (1 = least mental effort, and 9 = most mental effort) for load validation purposes.

4 Results

Figures 4 and 5 depict the summary visualization of trust measured via subjective responses of all 42 subjects. Trust values were normalized with respect to each subject to minimize individual differences in rating behavior. Since we had more than two dependent samples, we first performed Friedman test and then followed it up with post-hoc analysis using Wilcoxon signed-rank tests (with a Bonferroni correction) to analyze differences in participant responses of trust for various conditions.

Trust and Uncertainty: Figure 4 shows normalized trust values over the uncertainty treatments. *Control task* had only point prediction lines (refer to Fig. 1) and no uncertainty was presented. *Risk* uncertainty was presented by models with non-overlapping uncertainty (see Fig. 2a) and *ambiguity* by overlapping uncertainty models (see Fig. 2b).

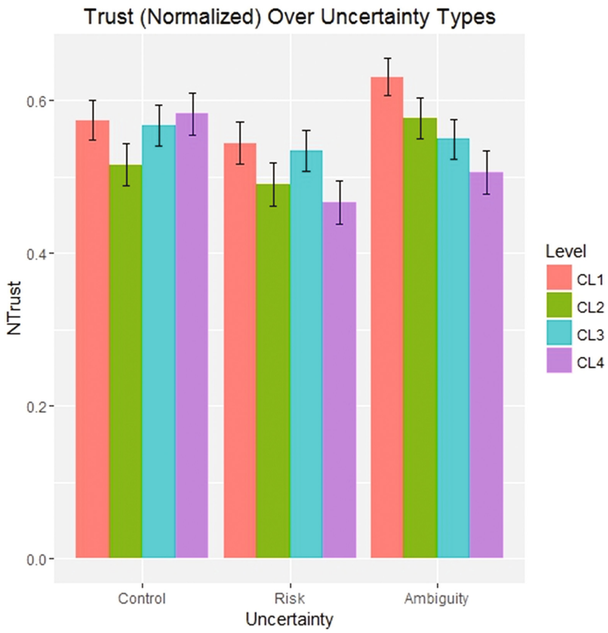

Fig. 4. Trust over uncertainty presented; control (No Uncertainty), risk (Non-Overlapping Uncertainty) and ambiguity (Overlapping Uncertainty).

When participants experienced uncertainty type ambiguity (rightmost group of columns in Fig. 4), Friedman test for cognitive load level conditions showed a statistically significant difference in trust among four CL levels, $\chi^2(3) = 12.363$, $p < .006$. Then post-hoc Wilcoxon tests (with a Bonferroni correction under a significance level set at $p < .013$) was applied to find pair-wise differences between levels in trust. The adjusted significance alpha level of .013 was calculated by dividing the original alpha of .05 by 4, based on the fact that we had four load level conditions to test.

The post-hoc tests found that for uncertainty condition of ambiguity, participants had significantly lower trust under high cognitive load (CL4), with $p < .001$, compared to that of low load (CL1). More details of this result and its implications for subject groups appear in discussion ahead.

Trust and Cognitive Load: Figure 5 shows normalized trust values over cognitive load levels. Here we are interested only in the extreme load levels administered, namely CL1 (the lowest) and CL4 (the highest), as they are the most relevant for automated cognitive load management [32]. Friedman's test of cognitive load level conditions of the lowest (CL1) and highest (CL4) both gave statistically significant differences in trust among three uncertainty conditions, $\chi^2(2) = 11.227$, $p < .004$ and $\chi^2(2) = 10.356$, $p < .006$ respectively. Then post-hoc Wilcoxon tests (with a Bonferroni correction under a significance level set at $p < .017$) was applied to find pair-wise differences between uncertainty conditions. The adjusted significance alpha

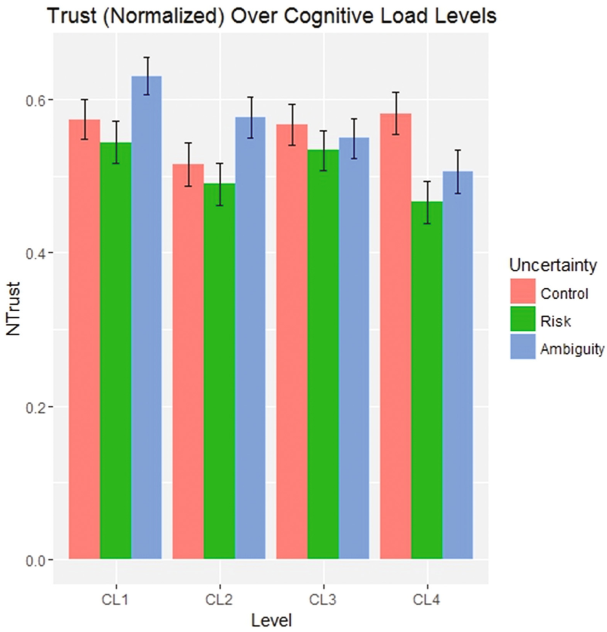

Fig. 5. Trust over cognitive load levels.

level of .017 was calculated by dividing the original alpha of .05 by 3, based on the fact that we had three uncertainty conditions to test.

The post-hoc tests found that for low cognitive load condition (CL1), trust in the condition of ambiguity was significantly higher ($p < .006$) than that of risk condition (Fig. 5, leftmost group of three columns). Whereas, for high cognitive load (CL4) condition, trust in risk condition was significantly lower ($p < .003$) than control condition. More details of this result and its implications for subject groups appear in discussion ahead.

5 Discussion

As discussed in earlier sections, trust is a challenging concept to study and investigate in, therefore, we opted to study human-machine trust in a specialized predictive decision making scenario. Predictive decision making support and automated aids have become quite popular with the advent of new machine learning based intelligent applications. Now that machines are becoming more intelligent – human-computer interaction must also evolve accordingly. Only by working together as a trusted team can humans and machines improve efficiency and productivity.

Trust and Uncertainty: Generally, in an automated predictive decision making scenario, humans are required to make future oriented decisions based on the information or recommendation presented on the screen by a machine learning (and data crunching) model/algorithm that mostly works on historical data behind the scenes (appearing like a black box to user). Since these decisions are about the future, there can be no absolutely correct answers - but only better and more appropriate ones based on a more precise understanding of the underlying data presented during the decision making process. Therefore, 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. Risk is a form of uncertainty where all probabilities related to outcomes are known. The user, with the help of these known probabilities, can be expected to make better and well-informed decisions quickly. Such risk type uncertainty was represented by non-overlapping models (see Fig. 2a). The other type of uncertainty we experimented with was ambiguity, which was represented by overlapping models (see Fig. 2b) and where probabilities of outcomes were either unknown or not clearly stated. The control condition was the case where models were presented without any uncertainty component.

Looking at the overall results (Fig. 4), no clear trends can be observed for risk type uncertainty condition, but a clear trend of falling 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 analyzing and interpreting the ambiguity type of uncertainty and therefore appear 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 lesser 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 [27].

Further drilling down deeper into 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 Fig. 6). 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 same for all administrative staff and experts (whether they be machine learning or non-machine learning).

Important lessons here for improved trust can be to either avoid ambiguity type of uncertainty representation altogether or present it only when condition of low cognitive load. Also another direction could be to look for alternate ways of visualizing ambiguity type of uncertainty.

Trust and Cognitive Load: It is well known that human performance can be significantly affected by high cognitive or mental workload [33]. 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

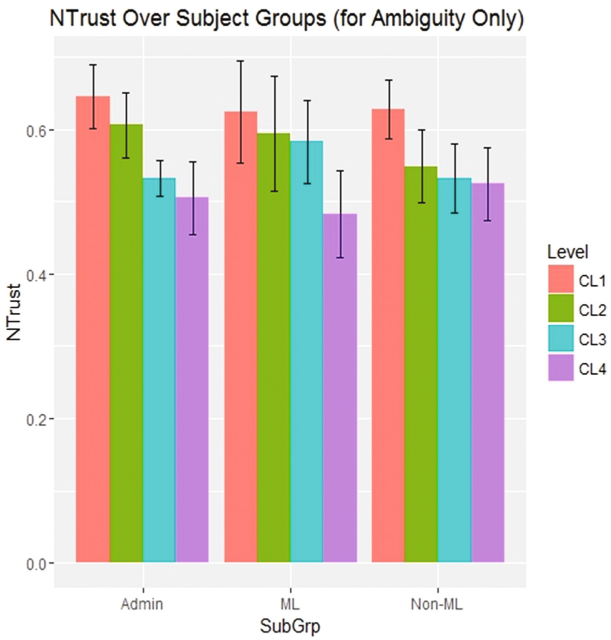

Fig. 6. Trust for ambiguity type uncertainty.

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 (see leftmost group of columns in Fig. 5). The trend seems to be the same for all subject subgroups (see Fig. 7). 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 (rightmost group of columns in Fig. 7) 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.

Likewise in high load condition (CL4), trust for risk type uncertainty was significantly lower than control condition of no uncertainty presentation (see rightmost group of columns in Fig. 5). The trend seems to be similar for all subject subgroups (see Fig. 8). 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 Fig. 8) 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. Of the total 42, there were 9 administrative staff, 14 machine learning experts and 19 non-machine learning experts.

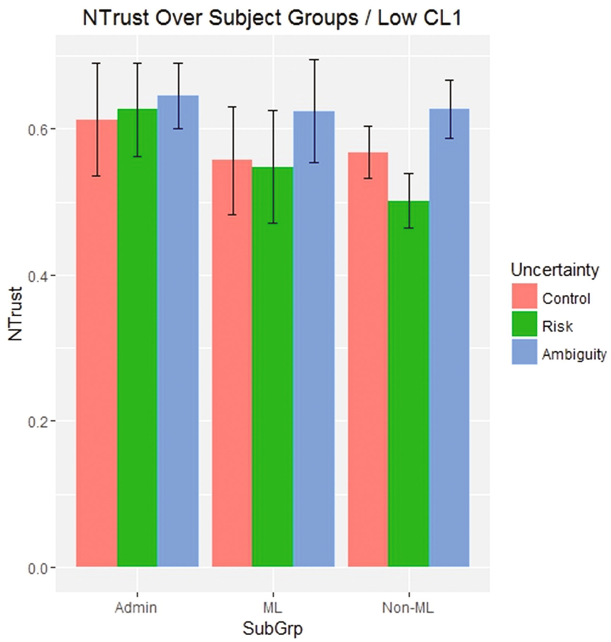

Fig. 7. Trust over subject groups (Low CL).

These findings go on to support the idea that people trust less what they have not had time to understand. Visuals presented in control condition are straightforward with no complication – however, they are simple, but only at the cost of hiding away the uncertainty inherent in the ML models. Once attempts are made to communicate the uncertainty – the trust seems to increase from control to risk and ambiguity uncertainty only under conditions of low cognitive load (see Fig. 7) and decrease under conditions of high cognitive load (see Fig. 8).

User Groups: The three user groups of administrative staff, ML experts and Non-ML experts constitute the majority of the users of such predictive decision making interfaces [18]. Of the people involved in predictive decision making, administrative staff is expected to be least knowledgeable of statistical probability and its representations. Non-ML experts may or may not be knowledgeable of statistical probability but they can be expected to be familiar with model representations. Finally the ML experts are expected to be the most knowledgeable of statistical probability and also of uncertainty inherent in ML models. In our case, from the total of 42, there were 9 administrative staff, 14 machine learning experts and 19 non-machine learning experts. The administrative group by itself is too small in number for any meaningful inference. ML-experts were reasonably good in number but nothing significant would be inferred. Only the Non-ML group appears to significant results in certain conditions as discussed above.

Overall, we can say 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,

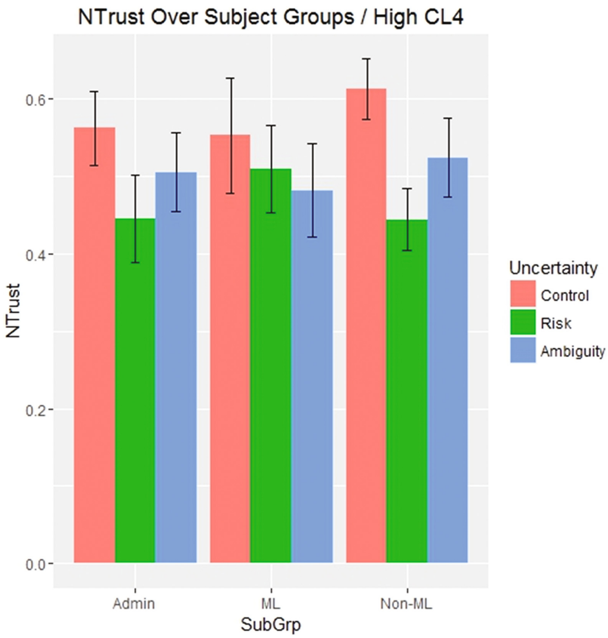

Fig. 8. Trust over subject groups (High CL).

when cognitive resources are short in supply can lead to lowering of trust in the system and its recommendations.

In order to incorporate these findings in HCI and real-world applications, the user interface for an ML-based intelligent system needs to include the following components:

- Components which show uncertainty of ML models. This could help users increase trust in their decisions;
- Feedback on user trust and load levels that allow users interfaces to adapt accordingly.

These components may be incorporated into the framework of adaptive measure-able decision making proposed in [34], thereby introducing trust levels into an adaptive decision making process and allowing for efficient and informed decisions. Therefore, besides decision quality as demonstrated in [35], the revealing of user trust levels in predictive decision making also benefits the evaluation of ML models. From this perspective, this study made “black-box” ML models transparent through revealing user responses to ML models, but not directly explain how ML algorithms process data to get outputs with visualizations or feature contributions [36–38], where domain users still have difficulty to understand those complex visualizations and abstract numbers. The revealing of user trust in a predictive decision making scenario is more meaningful for both ML researchers and domain experts, and therefore help improve the acceptance of ML solutions by users.

In summary, this study showed that uncertainty and cognitive load played significant roles in affecting user trust in predictive decision making. Furthermore, various uncertainty types had different effects on user trust perceptions. These findings have at least two benefits in real-world applications: (1) to design intelligent user interface of predictive decision related applications in HCI. The user interface, which shows user trust in decision making in real-time, would help users make informed decisions effectively; (2) to evaluate ML models in ML research areas by measuring what is the user trust level in decision making based on ML output.

6 Conclusions and Future Work

This paper investigated the effects of uncertainty of ML models and cognitive load on user trust in predictive decision making in order to design effective user interfaces for ML-based intelligent systems. A user study found that both uncertainty types, as well as the cognitive load levels affected user trust in decision making. Furthermore, various user groups showed different trust perceptions under both uncertainty and cognitive load conditions.

Our future work will focus on analyzing physiological signals, such as Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP) signals, as well as behavioral signals (mouse movement) of participants for indexing user trust levels during ML-based decision making. The relationship between the trust and the task performance will also be analyzed. Our ultimate goal is to set up a framework of measurable trust in decision making in order to dynamically adjust trust levels in ML-based intelligent systems.

Acknowledgements. Authors thank all volunteer participants for the experiment. This work was supported in part by the Asian Office of Aerospace Research & Development (AOARD) under grant No. FA2386-14-1-0022 AOARD 134131.

References

1. Hoff, K.A., Bashir, M.: Trust in automation integrating empirical evidence on factors that influence trust. Hum. Factors J. Hum. Factors Ergon. Soc. **57**, 407–434 (2015). doi:10.1177/0018720814547570
2. Marusich, L.R., Bakdash, J.Z., Onal, E., et al.: Effects of information availability on command-and-control decision making performance, trust, and situation awareness. Hum. Factors J. Hum. Factors Ergon. Soc. **58**, 301–321 (2016). doi:10.1177/0018720815619515
3. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manage. Rev. **20**, 709–734 (1995). doi:10.5465/AMR.1995.9508080335
4. Schoorman, F.D., Mayer, R.C., Davis, J.H.: An integrative model of organizational trust: past, present, and future. Acad. Manage. Rev. **32**, 344–354 (2007). doi:10.5465/AMR.2007.24348410
5. Wheeler, S.: Trusted Autonomy: Concept Development in Technology Foresight. Defense Technical Information Center, Defense Science & Technology Group, Australia (2015)

6. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**, 50–80 (2004)
7. Mercado, J.E., Rupp, M.A., Chen, J.Y.C., et al.: Intelligent agent transparency in human-agent teaming for Multi-UxV management. Hum. Factors **58**, 401–415 (2016)
8. Ososky, S., Sanders, T., Jentsch, F., et al.: Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In: Karlsen, R.E., Gage, D. W., Shoemaker, C.M., Gerhart, G.R. (eds.), p. 90840E (2014)
9. Adobor, H.: Optimal trust? Uncertainty as a determinant and limit to trust in inter-firm alliances. Leadersh. Org Dev. J. **27**, 537–553 (2006). doi:10.1108/01437730610692407
10. Parasuraman, R., Sheridan, T.B., Wickens, D.C.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Making **2**, 140–160 (2008)
11. Khawaji, A., Chen, F., Zhou, J., Marcus, N.: Trust and cognitive load in the text-chat environment: the role of mouse movement. In: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design, pp. 324–327 (2014)
12. Smith, P.J., Geddes, N.D., Beatty, R.: Human-centered design of decision-support systems. In: Human-Computer Interaction: Design Issues, Solutions, and Applications (2009)
13. Zhou, J., Li, Z., Wang, Y., Chen, F.: Transparent machine learning — revealing internal states of machine learning. In: Proceedings of IUI 2013 Workshop on Interactive Machine Learning (2013)
14. Zhou, J., Khawaja, M.A., Li, Z., et al.: Making machine learning useable by revealing internal states update — a transparent approach. Int. J. Comput. Sci. Eng. **13**, 378–389 (2016)
15. Winkler, R.L.: The importance of communicating uncertainties in forecasts: overestimating the risks from winter storm Juno. Risk Anal. **35**, 349–353 (2015)
16. Beller, J., Heesen, M., Vollrath, M.: Improving the driver-automation interaction an approach using automation uncertainty. Hum. Factors J. Hum. Factors Ergon. Soc. **55**, 1130–1141 (2013). doi:10.1177/0018720813482327
17. Candello, H.: User methods and approaches to design cognitive systems. In: Marcus, A. (ed.) DUXU 2016, Part I. LNCS, vol. 9746, pp. 231–242. Springer, Cham (2016). doi:10.1007/978-3-319-40409-7_23
18. Arshad, S.Z., Zhou, J., Bridon, C., et al.: Investigating user confidence for uncertainty presentation in predictive decision making. In: Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, pp. 352–360. ACM, New York (2015)
19. Rouse, W.B.: Adaptive aiding for human/computer control. Hum. Factors J. Hum. Factors Ergon. Soc. **30**, 431–443 (1988). doi:10.1177/001872088803000405
20. Byrne, E.A., Parasuraman, R.: Psychophysiology and adaptive automation. Biol. Psychol. **42**, 249–268 (1996)
21. Moray, N., Inagaki, T., Itoh, M.: Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. J. Exp. Psychol. Appl. **6**, 44–58 (2000)
22. LeClerc, J., Joslyn, S.: The cry wolf effect and weather-related decision making. Risk Anal. **35**, 385–395 (2015). doi:10.1111/risa.12336
23. Uggirala, A., Gramopadhye, A.K., Melloy, B.J., Toler, J.E.: Measurement of trust in complex and dynamic systems using a quantitative approach. Int. J. Ind. Ergon. **34**, 175–186 (2004)
24. Allen, P.M., Edwards, J.A., Snyder, F.J., et al.: The effect of cognitive load on decision making with graphically displayed uncertainty information. Risk Anal. **34**, 1495–1505 (2014)

25. Ibrekk, H., Morgan, M.G.: Graphical communication of uncertain quantities to nontechnical people. Risk Anal. **7**, 519–529 (1987)
26. Edwards, J.A., Snyder, F.J., Allen, P.M., et al.: Decision making for risk management: a comparison of graphical methods for presenting quantitative uncertainty. Risk Anal. **32**, 2055–2070 (2012). doi:10.1111/j.1539-6924.2012.01839.x
27. Kirschenbaum, S.S., Trafton, J.G., Schunn, C.D., Trickett, S.B.: Visualizing Uncertainty The Impact on Performance. Hum. Factors J. Hum. Factors Ergon. Soc. **56**, 509–520 (2014). doi:10.1177/0018720813498093
28. Damghani, K.K., Taghavifard, M.T., Moghaddam, R.T.: Decision Making Under Uncertain and Risky Situations. Enterprise Risk Management Symposium Monograph Society of Actuaries-Schaumburg, Illinois, vol. 15 (2009)
29. de Visser, E., Parasuraman, R.: Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload. J. Cogn. Eng. Decis. Making **5**, 209–231 (2011). doi:10.1177/1555343411410160
30. Li, Z., Zhang, B., Wang, Y., et al.: Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. Mach. Learn. **95**, 11–26 (2014)
31. Deck, C., Jahedi, S.: The effect of cognitive load on economic decision making: a survey and new experiments. Eur. Econ. Rev. **78**, 97–119 (2015)
32. Arshad, S.Z., Wang, Y., Chen, F.: Interactive mouse stream as real-time indicator of user's cognitive load. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1025–1030. ACM, New York (2015)
33. Chen, F., Zhou, J., Wang, Y., et al.: Robust Multimodal Cognitive Load Measurement. Springer International Publishing, Cham (2016)
34. Zhou, J., Sun, J., Chen, F., et al.: Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans. Comput. Hum. Interact. **21**, 33 (2015)
35. Helgee, E.A.: Improving Drug Discovery Decision Making using Machine Learning and Graph Theory in QSAR Modeling. Ph.D. thesis, University of Gothenburg (2010)
36. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst., **41**, 647–665 (2014)
37. Zhou, J., Chen, F.: Making machine learning useable. Int. J. Intell. Syst. Technol. Appl. **14**, 91 (2015). doi:10.1504/IJISTA.2015.074069
38. Krause, J., Perer, A., Ng, K.: Interacting with predictions: visual inspection of black-box machine learning models. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, New York, USA, pp. 5686–5697 (2016)