Towards Trustworthy Human-AI Teaming under Uncertainty

Jianlong Zhou[*] , **Fang Chen**
School of Computer Science, University of Technology Sydney, Australia
{fisrtName.lastName }@uts.edu.au

Abstract

This paper proposes to incorporate uncertainty and performance into AI-advised human decision making to boost user trust in AI. Motivated by the trial-to-trial approach in the adjustment of decision boundaries by humans in decision making, this paper presents a novel framework of human-AI teaming for trustworthy AI under uncertainty. The framework employs the trial-to-trial approach to simulate the trial to trial process for adjusting decision boundaries and provide an estimation of uncertainty in AI. An Uncertainty-Performance Interface (UPI) in the framework is then proposed to allow users access the quality and performance of machine learning models in AI at the same time. The proposed framework allows human and AI collaborate in a teaming environment for trustworthy decisions under uncertainty.

1 Introduction

With the rapid advancement of Machine Learning (ML) techniques, we continuously find ourselves coming across ML-based Artificial Intelligence (AI) systems that seem to work or have worked surprisingly well in practical scenarios (e.g. the self-driving cars, and the conversational agents for self-services). One of major applications of AI is to help human make decisions in complex situations [Zhou *et al.*, 2015]. In such situations, human and AI form a team to handle tasks: AI recommends an action to the human and the human can choose to trust and accept the recommendation or take a different action from its partner AI based on human's experiences with the system and domain knowledge. This kind of interaction is called predictive decision making [Zhou *et al.*, 2017] or AI-advised human decision making [Bansal *et al.*, 2019]. However, AI-advised human decision making still faces 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 methods for domain experts when offering AI solutions [Zhou and Chen, 2018]. For example, for many non-ML users, they simply provide source data to an AI system, and after selecting some menu options, the system displays colorful viewgraphs and/or recommendations as output. It is neither clear nor well understood *why ML algorithms made this prediction*, or *how trustworthy this output or decision based on the prediction was*. These questions demonstrate that user trust plays significant roles in affecting the impact of AI in practical applications.

Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [Lee and See, 2004]. This definition shows that uncertainty is tightly coupled to trust. Uncertainty is a common phenomenon in machine learning, which can be found in every stage of learning from input data and its preprocessing, algorithm design, feature selection, to model evaluation and others. For example, the input data to ML algorithms is often historical data recorded from different events. It is not perfect and often ambiguous because any sensors or humans only observe incomplete portions of the world or shadows of reality at any given time. There is also inherent uncertainty in ML algorithms due to statistical nature of them. 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. 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 [Adobor, 2006] 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.

Besides uncertainty, the performance of an ML model is usually measured with different metrics such as accuracy, F-score, AUC (Area Under the ROC Curve) [Ferri *et al.*, 2009]. Yin et al. [Yin *et al.*, 2018] found that the stated model accuracy has a significant effect on the extent to which people trust the model, suggesting the importance of communication of ML model performance for user trust. Zhou et al. [Zhou *et al.*, 2019] demonstrated that human has higher trust in ML models with higher performance than

[*] Contact Author

that with lower performance. Therefore, in the AI-advised human decision making, ML model performance is another key factor that affects human trust in AI [Zhou *et al.*, 2019; Bansal *et al.*, 2019].

This paper proposes to incorporate both uncertainty and performance of ML models into AI-advised human decision making to boost user trust in AI. Motivated by the trial-to-trial approach in the adjustment of decision boundaries by humans in decision making [Qamar *et al.*, 2013], the trial-to-trial is introduced to simulate the trial to trial process in AI by humans. An interface showing relations between uncertainty and model performance is presented, which allows human and AI to collaborate in a teaming environment for tasks. A framework of human-AI teaming for trustworthy AI under uncertainty is proposed to demonstrate the partnership between human and AI.

2 Related Work

Human trust in ML has been investigated from different perspectives. Explainable ML aims to interpret ML arrived at a specific decision with algorithms or visualizations to enable human users to understand, appropriately trust, and effectively manage the ML-based solutions [Zhou and Chen, 2018]. For example, feature contributions and influence of each training data point on predictions are typical approaches for ML explanations [Koh and Liang, 2017; Zhou *et al.*, 2019]. Ribeiro et al. [Ribeiro *et al.*, 2016] explained predictions of classifiers by learning an interpretable model locally around the prediction and visualizing importance of the most relevant features to improve user trust in classifications. Other studies that empirically tested the importance of explanation to users, in various fields, consistently showed that explanations significantly increase users' confidence and trust [Symeonidis *et al.*, 2009]. Explanation also shows the ability to correctly assess whether a prediction is accurate [Biran and McKeown, 2017].

Besides explanations, model performance such as accuracy is investigated to understand how variations of model performance affect user trust. Bansal et al. [Bansal *et al.*, 2019] found that an update of increased model accuracy may reduce human-AI teaming performance (also decrease human trust) when the update is not compatible with prior user experience. Yu et al. [Yu *et al.*, 2019] explored how user trust in the intelligent system varies with its accuracy. It was found that users are able to correctly perceive the accuracy and stabilize their trust to a level correlated with the accuracy, though system failures have a stronger impact on trust than system successes.

Moreover, various researches have been investigated to learn user trust variations in ML. Ye and Johnson [Ye and Johnson, 1995] experimented with three types of explanations (trace, justification and strategy) for an expert system, and found that justification (defined as showing the rationale behind each step in the decision) was the most effective type of explanation in changing users' attitudes towards the system. Kizilcec [Kizilcec, 2016] proposed that the transparency of algorithm interfaces can promote awareness and foster user trust. It was found that appropriate transparency of algorithms through explanation benefited the user trust. However, too much explanation information on algorithms eroded user trust.

On the other hand, various approaches have been proposed to represent uncertainty in machine learning. Fuzzy logic was the fashionable way of dealing with uncertainty in machine learning before Bayesian approaches came to the front [Zadeh, 1983; Hammer and Villmann, 2007]. Bayesian methods provide a natural probabilistic representation of uncertainty [Blundell *et al.*, 2015]. Gal et al. [Gal and Ghahramani, 2016] used a spike and slab variational distribution to view dropout at test time as approximation of uncertainty. Ritter et al. [Ritter *et al.*, 2018] constructed a Kronecker factored approximation to the Hessian matrix for Laplace approximation to the posterior over the weights of a trained network to estimate uncertainty of ML models. Maddox et al. [Maddox *et al.*, 2019] used the information contained in the Stochastic Gradient Descent (SGD) trajectory to approximate the posterior distribution over the weights of the neural network to estimate uncertainty of models. However, the work on uncertainty in ML mostly focuses on the evaluation of uncertainty in machine learning but not how uncertainty affects user trust.

Furthermore, the study of neural representation of uncertainty demonstrates that different types of uncertainty are encoded in different areas of the brain [Ma and Jazayeri, 2014]. Zhou et al. [Zhou *et al.*, 2017] also investigated the effects of uncertainty on trust under different cognitive load conditions. Therefore, it is possible to investigate human trust in machine learning by examining human physiological and behavioural responses in a AI-advised decision making scenario under uncertainty.

These previous work motivate us to consider both uncertainty and model performance in the AI-advised human decision making, aiming to set up effective human-AI teaming for trustworthy decision making.

3 A Framework of Human-AI Teaming

A framework of human-AI teaming for trustworthy AI under uncertainty is proposed in this section (see Figure 1). In this framework, a test data is input to an AI system (ML model) in order to get predictions/recommendations for decision making by humans (AI-advised human decision making). A trial-to-trial approach, which simulates human's trial to trial method in the adjustment of decision boundaries in decision making [Qamar *et al.*, 2013; Honig *et al.*, 2018], is introduced into this framework. Furthermore, from the trial-to-trial approach, uncertainty of the AI system can be estimated besides model performance. Therefore, another component of Uncertainty-Performance Interface (UPI) which shows relations between uncertainty and performance of the ML model is introduced into the framework. The UPI acts as an agent between the human and the AI: the human interacts with the UPI to understand and address the uncertainty/performance trade-off, while the AI provides responses on ML performance based on interaction requests by the human. With such interactions, the human obtains trust perceptions in the AI and makes trustworthy decisions.

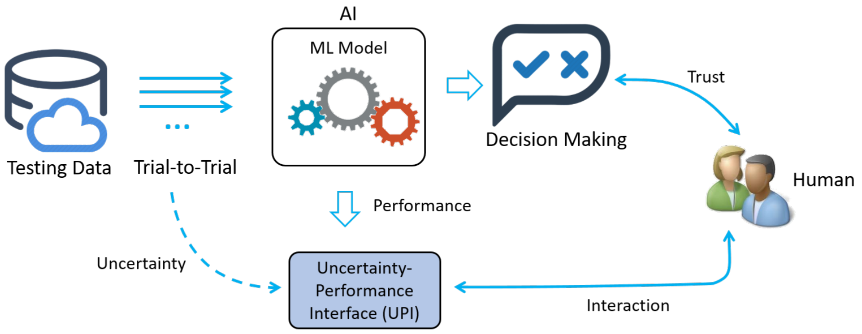

Figure 1: A framework of human-AI teaming for trustworthy AI under uncertainty.

3.1 Trial-to-Trial

The trial-to-trial approach is often used to understand uncertainties inherent and adjust decision boundaries gradually in decision making. Research found that adjusting decision boundaries from the trial-to-trial based on the current level of uncertainty is a better strategy than using uncertainty-independent decision boundaries [Qamar *et al.*, 2013]. Yu et al. [Yu *et al.*, 2019] found that trust can be derived from a series of user's decisions (i.e. different trials) rather than from a single one. Motivated by these observations, this section proposes to use trial-to-trial approach to:

- Simulate the trial to trial process for adjusting decision boundaries in AI-advised human decision making;
- Provide an estimation of uncertainty in AI for human decision making;
- Set up the bridge between uncertainty and trust in AI.

In practical applications especially high-stakes domains, it is not possible to allow users to try the AI system from time to time to get optimal decisions. The trial-to-trial approach in the proposed framework simulates the trial to trial process by randomly sampling different part of testing data as input to the AI system (see Figure 2). Each randomly sampled data are input to the AI, and therefore we get a series of outputs from the AI system for all trials. The overall uncertainty and performance of the AI can be obtained statistically from these series of AI outputs. For example, the standard deviation of the series of AI outputs can be regarded as the uncertainty.

3.2 Uncertainty-Performance Interface

It was found that humans can incorporate uncertainty inherently into their decision making and report their confidence in decisions meaningfully to reflect the uncertainty in their decisions [Honig *et al.*, 2018]. Furthermore, model performance has a significant effect on the extent to which people trust the ML model in AI-advised human decision making. The Uncertainty-Performance Interface in the proposed framework incorporates uncertainty and performance of AI into a single interface. Figure 3 presents an example of UPI which shows the relations between uncertainty and performance of an AI system. In this example, each dot represents one pair of uncertainty and performance of the AI system, the histogram and regression of uncertainty are displayed on the top side, the histogram and regression of performance are displayed on the right side, and the regression of both uncertainty and performance are represented as a blue line inside the figure. Therefore, distributions of both uncertainty and performance of the ML model can be accessed by humans in the UPI. These distributions help humans easily address the uncertainty/performance trade-off interactively in AI-advised human decision making. For example, for the given testing data and the AI system, the trials whose uncertainty and performance are the mean value respectively (e.g. the model on point C in Figure 3) may give humans an overall perception of uncertainty and performance for their decisions. The UPI acts as an agent between human and AI communications. It allows humans to perceive different trials at the same time and compare them in one place. From such comparison, we can at least get following information from the UPI:

- Model quality: the distribution of dots of uncertainty-performance pairs in the UPI may reflect the ML model quality. If the dots of uncertainty-performance pairs are scattered throughout the overall UPI, it shows the low quality of the ML model. The more compact the dots are, the higher the ML model quality is. Therefore, it is expected that humans have higher trust in the AI system when the dots of uncertainty-performance pairs are more compact in the UPI.
- Performance with a specific testing data: the region, where the most of compact dots of uncertainty-performance pairs are located in the UPI, may reflect the overall AI performance with the testing data. For example, in the UPI, the bottom-right region represents trials with high uncertainty and low performance, while the top-left region represents trials with low uncertainty and high performance. Therefore, it is expected that the most of compact dots of uncertainty-performance pairs are located in the top-left region in the UPI. Such information will help to enhance user trust in AI during decision making.

3.3 Human-AI Teaming for Decision Making under Uncertainty

Humans and AIs have complementary strengths and abilities. In the AI-advised human decision making, humans and AIs form a team in which AIs give recommendations to humans

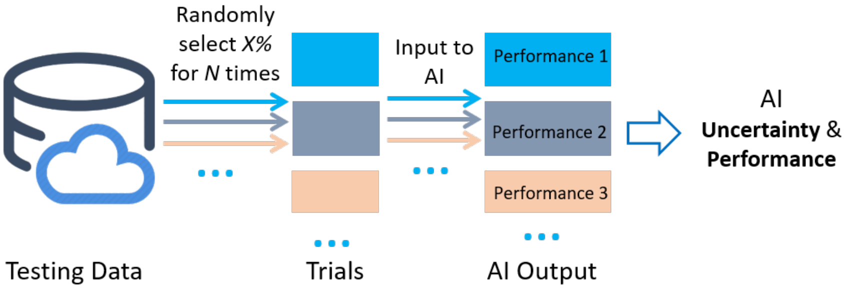

Figure 2: Trial-to-trial approach.

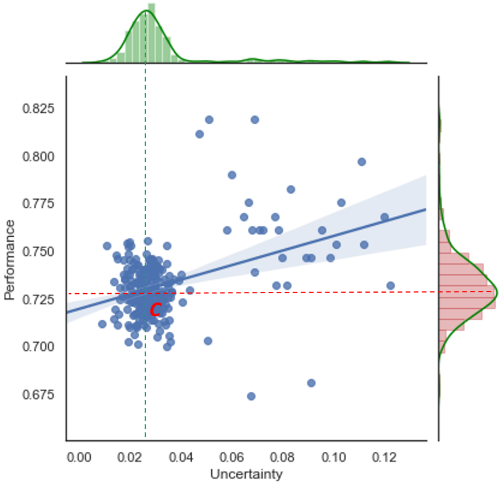

Figure 3: The UPI shows relations between uncertainty and performance.

and humans make final decisions after reviewing AI recommendations to achieve mutual enhancement between them. However, the effective teaming between the human and the AI for trustworthy decision making is still a challenge especially under uncertainty.

In the proposed framework of human-AI teaming for trustworthy AI under uncertainty, the UPI incorporates uncertainty and performance of AI into the decision making pipeline and allows humans to access them at the same time. In such new teaming framework (see Figure 1), the trial-to-trial simulates different trials of the AI system with the test data. These trials are then presented to the human at the same time in the UPI. The human interacts with different trials in the UPI to make decisions. Since the UPI shows distributions of uncertainty and performance of the AI system at the same time based on the trials, the interaction in the UPI helps the human to optimise decision boundaries under uncertainty [Qamar *et al.*, 2013] and results in trustworthy decisions. The UPI simplifies the interaction between human and AI.

4 Discussions and Research Challenges

Generally, it is assumed that the AI is more efficient (e.g. higher accuracy, faster) than the human in some instances in decision making, while the human can recognise when the AI is capable of doing so [Bansal *et al.*, 2019]. In the framework of human-AI teaming for trustworthy AI under uncertainty, the UPI is the place where the human accesses uncertainty and performance of the AI system meaningfully. Given different trials of the AI system in the UPI, the compactness and location of the dots of trials in the UPI provide the human meaningful information such as the quality of the model and overall performance with the testing data. The human interacts with these trials in the UPI to refine decision boundaries and make final decisions.

However, we still have various research challenges with the proposed framework. Two of significant challenges are:

- Human trust variations in AI-advised human decision making in the proposed framework. We need to examine human behaviour in the UPI and understand human trust variations given different trials with uncertainty-performance pairs in the UPI.
- New trust measures for AI. New trust measures needs to be developed by incorporating uncertainty and performance of the AI system at the same time.

These challenges motivate future research directions of the proposed framework.

5 Conclusion

This paper presents an overview of the conceptual framework of human-AI teaming for trustworthy AI under uncertainty, which aims to achieve trustworthy decisions in AI-advised human decision making scenarios. Three major components of the framework were explored in details: trial-to-trial, uncertainty-performance interface, and human-AI teaming for decision making under uncertainty. Our work demonstrates a new approach to get trustworthy AI under uncertainty by simulating human's trial to trial method in the adjustment of decision boundaries in decision making. Our future work will focus on examining human trust variations given uncertainty-performance pairs of an AI system and developing new trust measures for AI by incorporating uncertainty and performance of the AI system.

Acknowledgements

This work is supported in part by AOARD under No. 216624.

References

[Adobor, 2006] Henry Adobor. Optimal trust? uncertainty as a determinant and limit to trust in inter-firm alliancesnull. *Leadership & Organization Development Journal*, 27(7):537–553, 2006.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of AAAI 2019*, 2019.

[Biran and McKeown, 2017] Or Biran and Kathleen McKeown. Human-centric justification of machine learning predictions. In *Proceedings of IJCAI 2017*, pages 1461–1467, 2017.

[Blundell *et al.*, 2015] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proceedings of ICML 2015*, volume 37, pages 1613–1622, Lille, France, 2015.

[Ferri *et al.*, 2009] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, January 2009.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 2016.

[Hammer and Villmann, 2007] Barbara Hammer and Thomas Villmann. How to process uncertainty in machine learning? In *Proceedings of WEuropean Symposium on Artificial Neural Networks (ESANN'2007)*, 2007.

[Honig *et al.*, 2018] Maija Honig, Wei Ji Ma, and Daryl Fougnie. Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. pages 306–225, 2018.

[Kizilcec, 2016] René F. Kizilcec. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of CHI 2016*, pages 2390–2395, 2016.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of ICML 2017*, pages 1885–1894, Sydney, Australia, 2017.

[Lee and See, 2004] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[Ma and Jazayeri, 2014] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37(1):205–220, 2014.

[Maddox *et al.*, 2019] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *CoRR*, abs/1902.02476, 2019.

[Qamar *et al.*, 2013] Ahmad T. Qamar, R. James Cotton, Ryan G. George, Jeffrey M. Beck, Eugenia Prezhdo, Allison Laudano, Andreas S. Tolias, and Wei Ji Ma. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. 110(50):20332–20337, 2013.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, February 2016. arXiv: 1602.04938.

[Ritter *et al.*, 2018] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *Proceedings of International Conference on Learning Representations 2018*, 2018.

[Symeonidis *et al.*, 2009] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. Moviexplain: A recommender system with explanations. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 317–320, 2009.

[Ye and Johnson, 1995] L. Richard Ye and Paul E. Johnson. The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2):157–172, June 1995.

[Yin *et al.*, 2018] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Does stated accuracy affect trust in machine learning algorithms? In *Proceedings of ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018)*, 7 2018.

[Yu *et al.*, 2019] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. Do i trust my machine teammate?: An investigation from perception to decision. In *Proceedings of IUI 2019*, pages 460–468, 2019.

[Zadeh, 1983] Lotfi A. Zadeh. The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems*, 11(1-3):199–227, 1983.

[Zhou and Chen, 2018] Jianlong Zhou and Fang Chen, editors. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer, Cham, 2018.

[Zhou *et al.*, 2015] Jianlong Zhou, Jinjun Sun, Fang Chen, Yang Wang, Ronnie Taib, Ahmad Khawaji, and Zhidong Li. Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface. *ACM Transactions on Computer-Human Interaction*, 21(6):33, 2015.

[Zhou *et al.*, 2017] Jianlong Zhou, Syed Z. Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *Proceedings of 16th IFIP TC 13 International Conference on Human-Computer Interaction - INTERACT 2017*, pages 23–39, Mumbai, India, 2017.

[Zhou *et al.*, 2019] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. Effects of influence on user trust in predictive decision making. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts)*, Glasgow, Scotland UK, May 2019.