Explaining the Uncertainty in AI-Assisted Decision Making

Thao Le

School of Computing and Information Systems
The University of Melbourne
thaol4@student.unimelb.edu.au

Abstract

The aim of this project is to improve human decision-making using explainability; specifically, how to explain the (un)certainty of machine learning models. Prior research has used uncertainty measures to promote trust and decision-making. However, the direction of explaining why the AI prediction is confident (or not confident) in its prediction needs to be addressed. By explaining the model uncertainty, we can promote trust, improve understanding and improve decision-making for users.

Introduction

In Machine Learning (ML), the *confidence score* indicates how *certain* the model is in its prediction; or inversely, how uncertain it is. Prior research (Zhang, Liao, and Bellamy 2020) has used uncertainty (confidence) as a measure for trust calibration, which is a key factor in decision-making to help people decide if they should *trust* or *distrust* the model (i.e., *trust calibration*).

In my PhD project, I aim to apply both explainable AI techniques and uncertainty measures to promote people's trust and improve decision-making. Further, I also explore effective and informative explanation designs to help people better understand and make better decisions when interacting with the AI.

Research Questions

- **RQ1** Can explanations of uncertainty help users better **understand** and **trust** the AI model?
- **RQ2** How do we improve people's **decision making** with explanations and uncertainty?
- **RQ3** What design options are needed to distinguish explanations of aleatoric uncertainty (data uncertainty) versus epistemic uncertainty (model uncertainty)?

Progress to Date

Explaining Model Confidence Using Counterfactuals

I introduce a model that explains model confidence using counterfactuals (CF) (Le et al. 2022). A counterfactual explanation is described as the possible smallest changes in input values in order to change the model prediction to the desired output, which has been increasingly used in explainable AI (XAI) to facilitate human interaction with the AI model (Miller 2019). We formalise counterfactual explanations of confidence by extending the prior counterfactual model (Russell 2019). The difference between Russell (2019)'s model and our proposed approach is explained in Table 1. I then generate these explanations in two different presentation forms: (1) example-based counterfactuals and (2) visualisation-based counterfactuals.

To evaluate the explanation, we conduct user studies because it is increasingly accepted that explainability techniques should be built on studies in philosophy, psychology and cognitive science (Miller 2019) and that the assessment process of explanations should involve user studies. We recruited a total of **180** participants for two different domains. To evaluate **understanding**, we use *task prediction* (Hoffman et al. 2018, p11). Participants are given some instances and their task is to decide for which instance the AI model will predict a higher confidence score. Thus, task prediction helps evaluate the user's mental model about their understanding in model confidence. To evaluate **trust**, we use 10-point Likert *Trust Scale* from (Hoffman et al. 2018, p49). For **satisfaction**, we use the 10-point Likert *Explanation Satisfaction Scale* from (Hoffman et al. 2018, p39).

The results show that both forms of counterfactual explanations increase trust and understanding over a baseline of no explanation. Notably, there is minimal difference between visualisation-based and example-based in improving understanding, trust and satisfaction. Using qualitative analysis, we observe some limits of both approaches as follows:

- People use case-based reasoning to understand the *example-based explanation*. That is, they find the closest example in the example-based presentation and overlook the linear correlation between the confidence score and the feature values. This result suggests that we should be careful when using example-based explanations to interpret continuous variables.
- Although using *visualisation-based explanation* is easier to interpret the correlation, when not all counterfactual points are shown in the explanation, people are not willing to extrapolate the correlation beyond the lowest and highest values. Thus, all counterfactual points should be shown in the explanation to mitigate this issue.

	Original CF Model (Russell 2019)	CF Model for explaining model confidence
Method	Search for CF inputs of another class	Search for CF inputs of the same class, but with a different confidence score
Example Question	Why does the model predict this employee will *leave* instead of will *stay* in this company?	The model predicts that this employee will *leave*. Why is the model 70% confident instead of 40% confident or less?
Example Explanation	You could have got a prediction of *stay* instead if Age had taken the value of 45 rather than 25	You could have got a confidence score of 40% instead if Daily Rate had taken the value 400 rather than 300

Table 1: The difference between the CF model from prior research and the CF model from our approach

Remaining Work and Timeline

Improving Decision Making with Evidence-Based Explanation

We focus on how the explanation and uncertainty can improve people's decision-making. There are two methods often used to support decision-making: (1) using uncertainty or confidence measures (Zhang, Liao, and Bellamy 2020) and (2) using explanation AI techniques (Riveiro and Thill 2021). We aim to incorporate people's hypotheses into the decision-making process when interacting with the AI system based on the idea of abductive reasoning (Krawczyk 2018). Further, the explanations need to be informative rather than convincing, which is essential in case the AI recommendation is incorrect.

In recent research, Gajos and Mamykina (2022) suggest that providing only the AI explanation and *no AI recommendation* can help people process the AI explanation more carefully and therefore, improve their knowledge and make better decisions. Following this paper, we want to provide explanations *for humans' predictions* rather than providing explanations for the AI prediction. Furthermore, the explanation design needs to give both positive evidence (support the human's hypothesis) and negative evidence (against the human's hypothesis), which will help users to have adequate information to make a judgement. We also expect this design to reduce the over-reliance on the AI system when making the decision. This design can be further applied to explain the model uncertainty in the decision-making process. This is a work in progress and I anticipate having a new decision-making model by the workshop date (February 7, 2023).

Future directions of this work can address the uncertainty in provided explanations. Explanations are often generated based on probabilistic models. Therefore, they have some degree of uncertainty. If we can communicate effectively the uncertainty in the explanation, it is a promising direction to help people make better decisions.

Aleatoric Uncertainty versus Epistemic Uncertainty

In our future work, I want to explore the differences between aleatoric uncertainty (data uncertainty) and epistemic uncertainty (model uncertainty) in supporting decision-makers. There is limited empirical work on these differences (Bhatt et al. 2021) and therefore, I plan to investigate this by setting up a user study to find how people make use of aleatoric and epistemic uncertainty to make a decision. Moreover, we can define separate explanation models for aleatoric and epistemic uncertainty.

References

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; Nachman, L.; Chunara, R.; Srikumar, M.; Weller, A.; and Xiang, A. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.

Gajos, K. Z.; and Mamykina, L. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*, 794–806.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

Krawczyk, D. C. 2018. Chapter 11 - Decision Making and Abductive Reasoning. In *Reasoning*, 255–282. Academic Press.

Le, T.; Miller, T.; Singh, R.; and Sonenberg, L. 2022. Improving Model Understanding and Trust with Counterfactual Explanations of Model Confidence. arXiv:2206.02790.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Riveiro, M.; and Thill, S. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298.

Russell, C. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 20–28.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of Conference on Fairness, Accountability, and Transparency*, 295–305.