ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

.dit/*
Informatics & Telematics

Uncertainty Meets Explainability in Machine Learning

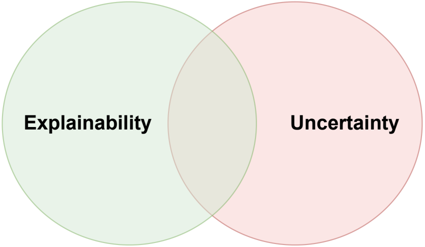
Christos Diou and Vasileios Gkolemis

Department of Informatics and Telematics
Harokopio University of Athens

ECML-PKDD 2023

Welcome

- Welcome to the “Uncertainty Meets Explainability” workshop!



https://xai-uncertainty.github.io/

Workshop schedule

2.30-2.35	Welcome and Introduction	Christos Diou & Vasilis Gkolemis	5 minutes
2.35-2.50	Short introduction on the intersection of uncertainty and explainability in machine learning	Christos Diou	15'
2.50-3.10	Using Stochastic Methods to Setup High Precision Experiments	Kristina Veljković	(17' presentation + 3' questions)
3.10-3.30	Using Part-based Representations for Explainable Deep Reinforcement Learning	Manos Kirtas	(17' presentation + 3' questions)
3.30-4.00	Explaining an image classifier with a GAN conditioned by uncertainty	Adrien Le Coz	7 minutes
	Identifying Trends in Feature Attributions during Training of Neural Networks	Elena Terzieva	7 minutes
	Relation of Activity and Confidence when Training Deep Neural Networks	Valerie Krug	7 minutes
	Temperature scaling for reliable uncertainty estimation: Application to automatic music genre classification	Hanna Lukashevich	7 minutes

Workshop schedule

4.30-4.50	Explainable Learning with Hierarchical Online Deterministic Annealing	Christos Mavridis	(17' presentation + 3' questions)
4.50-5.10	Regionally Additive Models: Explainable-by-design models minimizing feature interactions	Vasilis Gkolemis	(17' presentation + 3' questions)
5.10-5.45	FALE: Fairness aware ALE plots for auditing bias in subgroups	Giorgos Giannopoulos	7 minutes
	Improving the Validity of Decision Trees as Explanations	Jiří Němeček	7 minutes
	Towards Explainability in Monocular Depth Estimation	Vasileios Arampatzakis	7 minutes
	Explaining uncertainty in AI for clinical decision support systems	Elisabeth Heremans	7 minutes
	Designing a Method to Identify Explainability Requirements in Cancer Research	Didier Domínguez	7 minutes
5.45-6.00	Poster session - Poster dimensions (75x200 cm) double-side		15 minutes

A big Thank You to our PC members

- Albert Calvo (i2CAT)
- Carlous Mougan (Univ. of Southampton)
- Dimitrios Gunopulos (NKUA)
- Dimitris Sacharidis (ULB)
- Dimitris Fotakis (NTUA)
- Eirini Ntoutsi (UNIBW)
- Eleni Psaroudaki (NTUA)
- Giorgos Giannopoulos (ATHENA RC)
- Giorgos Papastefanatos (ATHENA RC)
- Giuseppe Casalicchio (LMU)
- Hamid Bouchachia (Bournemouth Univ.)
- Jakub Marecek (CVUT)
- Kostas Stefanidis (TUNI)
- Loukas Kavouras (ATHENA RC)
- Nikos Vryzas (AUTH)
- Maria Tzelepi (AUTH)
- Rahul Nair (IBM)
- Theodore Dalamagas (ATHENA RC)
- Theodora Tsikrika (CERTH/ITI)

Support

Organizations:

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

ATHENA
Research & Innovation
Information Technologies

Projects:

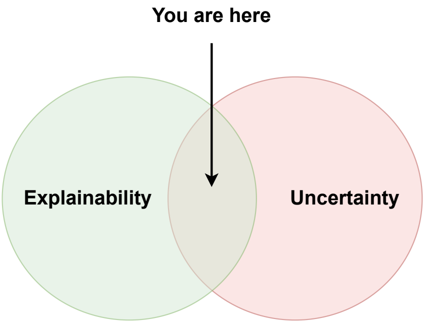
XMANAI
MAKING AI UNDERSTANDABLE

ExtremeXP

HUMANCOMPATIBLE.ORG
HUMAN-COMPATIBLE AI WITH GUARANTEES

RELEVIUM

Uncertainty ∩ Explainability

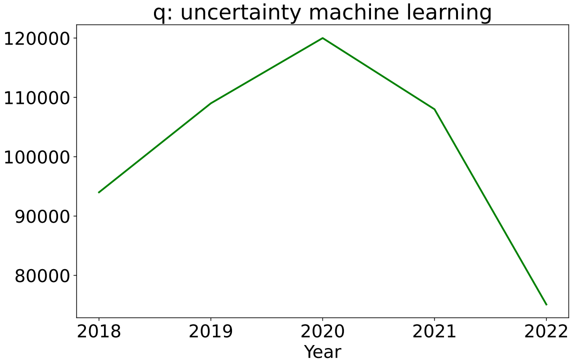


Interesting questions:

- What are the ways in which these fields interact?
- What is the current research interest in the combination of these two fields?
- How can we facilitate research in this area?

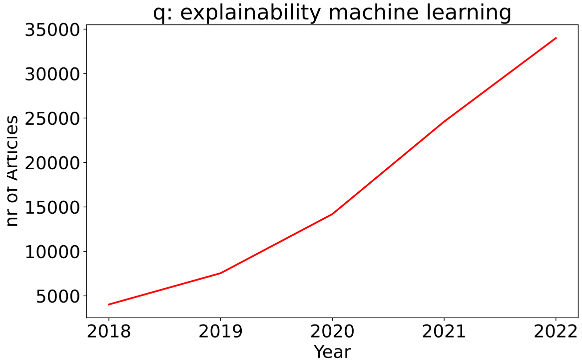
Publications in the last 5 years

Publications containing all query terms in the period 2018–2022, based on Google Scholar



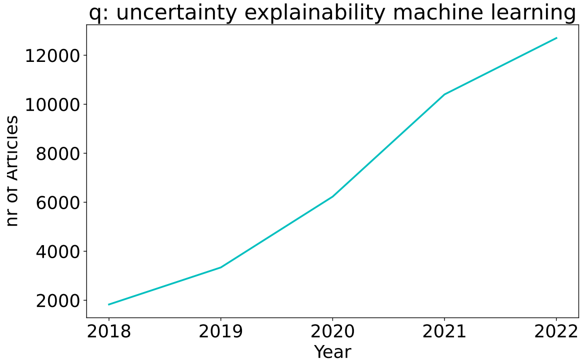
Publications in the last 5 years

Publications containing all query terms in the period 2018-2022, based on Google Scholar



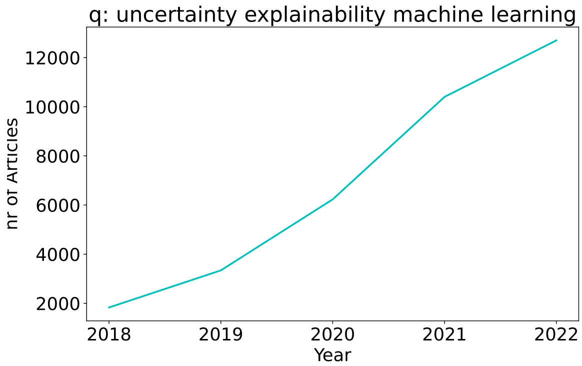
Publications in the last 5 years

Publications containing all query terms in the period 2018-2022, based on Google Scholar



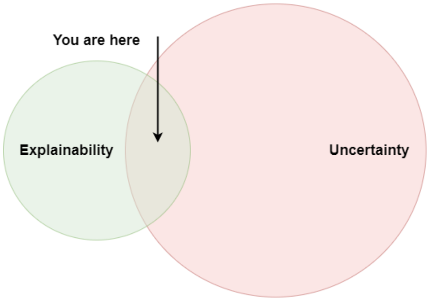
Publications in the last 5 years

Publications containing all query terms in the period 2018-2022, based on Google Scholar



Important note: These numbers indicate the number of papers where these terms co-ocur, not the number of papers that focus on the interaction of uncertainty and explainability.

Uncertainty ∩ Explainability



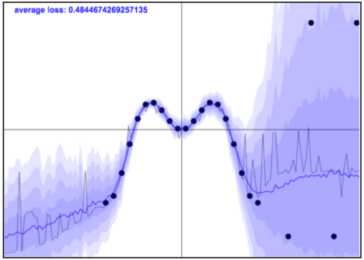
Uncertainty ∩ Explainability

It turns out, not surprisingly, that other people have the same ideas.
Examples:

- G. Scafarto, N. Prosocco, A. Bonnefoy, "Calibrate to Interpret", ECML 2022
- D. Folgado, M. Barandas, L. Famiglini, R. Santos, F. Cabitza, H. Gamboa, Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series, Information Fusion, 2023

The importance of uncertainty in ML

- An integral part of ML
- Sources of uncertainty (Hüllermeier and Waegeman, 2020):
  - Uncertainty inherent in the process
    - $p(y|\mathbf{x})$ even for the best possible model, f^*
  - Uncertainty due to the selected type of model
    - The best model h^*, from the selected family of models may be different from f^*
  - Uncertainty due to our approximation of the best model
    - Our approximation h may be different from h^*
- Aleatoric and epistemic uncertainty



Source: Y. Gal, PhD thesis

Uncertainty - questions

- How certain are we about our predictions?
    - Can you provide a set $\mathbf{C}(\mathbf{x})$ where the value of y lies with probability 0.95?
- What causes this uncertainty? Is it reducible? How?
- How certain are we that we have selected the correct model?
- Can we quantify aleatoric and epistemic uncertainty?

Recent papers discussing uncertainty in ML

Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods

Eyke Hüllermeier [a] and Willem Waegeman [b]

[a] Paderborn University
Heinz Nixdorf Institute and Department of Computer Science
eyke@upb.de

[b] Ghent University
Department of Mathematical Modelling, Statistics and Bioinformatics
willem.waegeman@ugent.be

Abstract

The notion of uncertainty is of major importance in machine learning and constitutes a key element of machine learning methodology. In line with the statistical tradition, uncertainty has long been perceived as almost synonymous with standard probability and probabilistic predictions. Yet, due to the steadily increasing relevance of machine learning for practical applications and related issues such as safety requirements, new problems and challenges have recently been identified by machine learning scholars, and these problems may call for new methodological developments. In particular, this includes the importance of distinguishing between (at least) two different types of uncertainty, often referred to as *aleatoric* and *epistemic*. In this paper, we provide an introduction to the topic of uncertainty in machine learning as well as an overview of attempts so far at handling uncertainty in general and formalizing this distinction in particular.

Uncertainty Quantification in Scientific Machine Learning: Methods, Metrics, and Comparisons

Apostolos F Psaros[a,*], Xuhui Meng[a,*], Zongren Zou[a], Ling Guo[b], George Em Karniadakis[a,c,**]

[a] *Division of Applied Mathematics, Brown University, Providence, RI 02906, USA*
[b] *Department of Mathematics, Shanghai Normal University, Shanghai, China*
[c] *Pacific Northwest National Laboratory, Richland, WA 99354, USA*

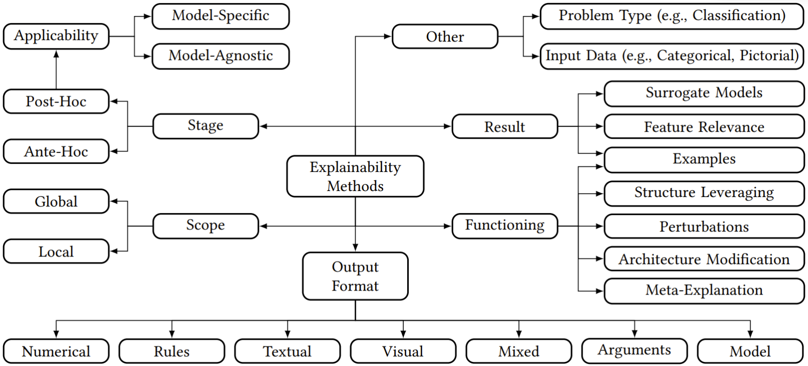
Abstract

Neural networks (NNs) are currently changing the computational paradigm on how to combine data with mathematical laws in physics and engineering in a profound way, tackling challenging inverse and ill-posed problems not solvable with traditional methods. However, quantifying errors and uncertainties in NN-based inference is more complicated than in traditional methods. This is because in addition to aleatoric uncertainty associated with noisy data, there is also uncertainty due to limited data, but also due to NN hyperparameters, overparametrization, optimization and sampling errors as well as model misspecification. Although there are some recent works on uncertainty quantification (UQ) in NNs, there is no systematic investigation of suitable methods towards quantifying the *total uncertainty* effectively and efficiently even for function approximation, and there is even less work on solving partial differential equations and learning operator mappings between infinite-dimensional function spaces using NNs. In this work, we present a comprehensive framework that includes uncertainty modeling, new and existing solution methods, as well as evaluation metrics and post-hoc improvement approaches. To demonstrate the applicability and reliability of our framework, we present an extensive comparative study in which various methods are tested on prototype problems, including problems

Explainability - example questions

- Why did a model make a specific decision?
- What would be a minimal change so that the model will make a different decision?
- Can we summarize and predict the overall model's behavior?

Taxonomy of interpretability methods



Timo Speith, “A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods”. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), 2022

Interpretable models

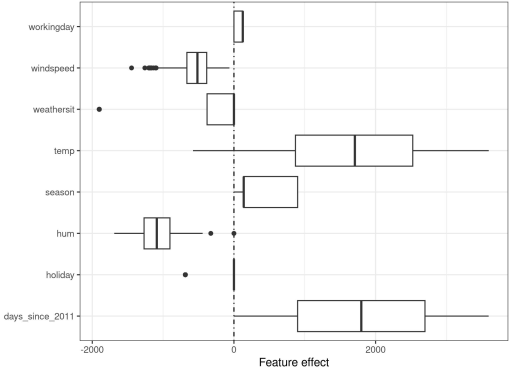
- Some models afford explanations
- Examples, (generalized) linear models, decision trees, k-NN
- Example: Linear regression

$$\hat{y} = w_1 x_1 + \ldots + w_p x_p + b$$

Interpretable models

- Feature effects (visualization) - example in bike sharing dataset

$$effect_j^{(i)} = w_j x_j^{(i)} \tag{1}$$



C. Molnar, IML book

Goal

- Most models do not afford explanations
  - we cannot explain them by looking at their parameters
  - we handle these as “black boxes”
- In this case we apply general interpretability methods
- **Local**: Interpret the model’s output for a particular input instance
- **Global**: Provide a general interpretation of the model’s behavior

Uncertainty $\cap$ Explainability

Explain the uncertainty

Uncertainty of the explanation

Post-hoc uncertainty explanations

Paper summary:

- Use non-parametric bootstrap and SHAP to provide explainable uncertainty estimates
- Use this to estimate model deterioration in the deployment environment (no labels)
- Detect the source of deterioration
- Key ideas:
  - Separate estimations of model variance noise, bias and observation noise terms in model
  - Use Shapley values to estimate the contribution of each feature in uncertainty and deterioration

Example paper:

Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap

Carlous Mougan[*1], **Dan Saattrup Nielsen**[*2]

[1] University of Southampton, United Kingdom
[2] The Alexandra Institute, Denmark
C.Mougan-Navarro@southampton.ac.uk, dan.nielsen@alexandra.dk

Abstract

Monitoring machine learning models once they are deployed is challenging. It is even more challenging to decide when to retrain models in real-case scenarios when labeled data is beyond reach, and monitoring performance metrics becomes unfeasible. In this work, we use non-parametric bootstrapped uncertainty estimates and SHAP values to provide explainable uncertainty estimation as a technique that aims to monitor the deterioration of machine learning models in deployment environments, as well as determine the source of model deterioration when target labels are not available. Classical methods are purely aimed at detecting distribution shift, which can lead to false positives in the sense that the model has not deteriorated despite a shift in the data distribution. To estimate model uncertainty we construct prediction intervals using a novel bootstrap method, which improves upon the work of Kumar and Srivastava (2012). We show that both our model deterioration detection system as well as our uncertainty estimation method achieve better performance than the current

new input data in order to maintain high performance. This process is called continual learning (Diethe et al. 2019) and it can be computationally expensive and put high demands on the software engineering system. Deciding when to retrain machine learning models is paramount in many situations.

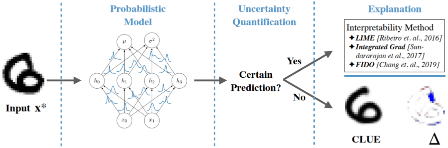
Traditional machine learning systems assume that training data has been generated from a stationary source, but *data is not static, it evolves*. This problem can be seen as a distribution shift, where the data distributions of the training set and the test set differ. Detecting distribution shifts has been a longstanding problem in the machine learning (ML) research community (Shimodaira 2000; Sugiyama, Krauledat, and Müller 2007; Sugiyama and Müller 2005; Tasche 2017; Zadrozny 2004; Stolzenberg and Relles 1997; Heckman 1990; Cortes et al. 2008; Huang et al. 2006; He et al. 2014), as it is one of the main sources of model performance deterioration (Quiñonero-Candela et al. 2008). Furthermore, data scientists in machine learning competitions claim that

Explanations of probabilistic models

Example paper:

Paper summary:

- When a BNN is uncertain about its predictions, the explanation is also affected. It is better to provide an explanation of why it is uncertain instead
- Key ideas:
  - Select a counterfactual in a latent space of a deep generative model such that the estimation of uncertainty is minimized
  - The difference between the original and new data highlights the source of uncertainty



Published as a conference paper at ICLR 2021

GETTING A CLUE: A METHOD FOR EXPLAINING UNCERTAINTY ESTIMATES

Javier Antorán
University of Cambridge
ja666@cam.ac.uk

Umang Bhatt
University of Cambridge
usb20@cam.ac.uk

Tameem Adel
University of Cambridge
University of Liverpool
tah47@cam.ac.uk

Adrian Weller
University of Cambridge
The Alan Turing Institute
aw665@cam.ac.uk

José Miguel Hernández-Lobato
University of Cambridge
The Alan Turing Institute
jmh233@cam.ac.uk

ABSTRACT

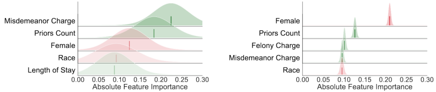
Both uncertainty estimation and interpretability are important factors for trustworthy machine learning systems. However, there is little work at the intersection of these two areas. We address this gap by proposing a novel method for interpreting uncertainty estimates from differentiable probabilistic models, like Bayesian Neural Networks (BNNs). Our method, Counterfactual Latent Uncertainty Explanations (CLUE), indicates how to change an input, while keeping it on the data

Quantifying the uncertainty of post-hoc explanations

Paper summary:

- Bayesian framework to generate local explanations with uncertainty
- Bayesian LIME and KernelSHAP
- Key ideas:
  - Define a generative process and use data to infer its parameters
  - Use the parameters to provide the explanation along with its estimated uncertainty
  - Can use these to select the required number of perturbations for providing reliable explanations

Example paper:



(a) Explanation computed with 100 perturbations (b) Explanation with 2000 perturbations

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability

Dylan Slack UC Irvine dslack@uci.edu

Sophie Hilgard Harvard University ash798@g.harvard.edu

Sameer Singh UC Irvine sameer@uci.edu

Himabindu Lakkaraju Harvard University hlakkaraju@hbs.edu

Abstract

As black box explanations are increasingly being employed to establish model credibility in high stakes settings, it is important to ensure that these explanations are accurate and reliable. However, prior work demonstrates that explanations generated by state-of-the-art techniques are inconsistent, unstable, and provide very little insight into their correctness and reliability. In addition, these methods are also computationally inefficient, and require significant hyper-parameter tuning. In this paper, we address the aforementioned challenges by developing a novel Bayesian framework for generating local explanations along with their associated uncertainty. We instantiate this framework to obtain Bayesian versions of LIME and KernelSHAP which output credible intervals for the feature importances, capturing the associated uncertainty. The resulting explanations not only enable us to make concrete inferences about their quality (e.g., there is a 95% chance that the feature importance lies within the given range), but are also highly consistent and stable.

Quantifying the uncertainty of post-hoc explanations

Paper summary:

- Introduce BayLIME: Another Bayesian version of LIME
- Key ideas:
    - Prior knowledge is introduced by weighting samples based on their proximity to the sample we wish to explain
    - Use these to estimate mean and variance for Bayesian linear regression in LIME

Example paper 2:

BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations

Xingyu Zhao[1,2] **Wei Huang**[1] **Xiaowei Huang**[1] **Valentin Robu**[2,3,4] **David Flynn**[2]

[1]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, U.K.
[2]School of Engineering & Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, U.K.
[3]Centrum voor Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
[4]Delft University of Technology, Algorithmics Group, EEMCS, 2628 XE Delft, The Netherlands

Abstract

Given the pressing need for assuring algorithmic transparency, Explainable AI (XAI) has emerged as one of the key areas of AI research. In this paper, we develop a novel Bayesian extension to the LIME framework, one of the most widely used approaches in XAI – which we call BayLIME. Compared to LIME, BayLIME exploits prior knowledge and Bayesian reasoning to improve both the

Model-agnostic Explanations (LIME) [Ribeiro et al., 2016]. Despite its very considerable success in both research and practice, LIME has several weaknesses, the most significant of which are the lack of *consistency in repeated explanations of a single prediction* and *robustness to kernel settings*. Meanwhile, higher *explanation fidelity* is also expected in many settings. Arguably, these three properties are among the most desirable for an XAI method to have.

The inconsistency of LIME, where different explanations can be generated for the same prediction, has been identified

Quantifying the uncertainty of post-hoc explanations

Paper summary:

- Global effect estimation methods introduce uncertainty due to sample heterogeneity
- Key ideas:
  - Use DALE, a fast and more accurate version of ALE for differentiable models
  - Provide an unbiased estimator of variance
  - Use this to select optimal bin splitting strategy for ALE

Example paper 3:

RHALE: Robust and Heterogeneity-aware Accumulated Local Effects

Vasilis Gkolemis[a, b], Theodore Dalamagas[b], Eirini Ntoutsi[c] and Christos Diou[a]

[a]Harokopio University of Athens
[b]ATHENA RC
[c]Universitat der Bundeswehr Munchen

Abstract. Accumulated Local Effects (ALE) is a widely-used explainability method for isolating the average effect of a feature on the output, because it handles cases with correlated features well. However, it has two limitations. First, it does not quantify the deviation of instance-level (local) effects from the average (global) effect, known as heterogeneity. Second, for estimating the average effect, it partitions the feature domain into user-defined, fixed-sized bins, where different bin sizes may lead to inconsistent ALE estimations. To address these limitations, we propose Robust and Heterogeneity-aware ALE (RHALE). RHALE quantifies the heterogeneity by considering the standard deviation of the local effects and automatically determines an optimal variable-size bin-splitting. In this paper, we prove that to achieve an unbiased approximation of the standard deviation of local effects within each bin, bin splitting must follow a set of sufficient conditions. Based on these conditions, we propose an algorithm that automatically determines the optimal partitioning, balancing the estimation bias and variance. Through evaluations on synthetic and real datasets, we demonstrate the superiority of RHALE compared to other methods, including the advantages of automatic bin splitting, especially in cases with correlated features.

for a complete interpretation of the average effect. Secondly, the way ALE estimates the FE from the instances of the training set (*ALE approximation* at Eq. (3)) relies on a user-defined binning process that often results in poor estimations. Therefore, this paper presents RHALE (Robust and Heterogeneity-aware ALE), a FE method build on-top of ALE that overcomes these issues. To better understand the advantages of RHALE over ALE, consider the following example, which was first introduced in [9]:

$$\begin{aligned} &Y = 0.2X_1 - 5X_2 + 10X_2\mathbb{1}_{X_3>0} + \mathcal{E} \\ &\mathcal{E} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad X_1, X_2, X_3 \overset{i.i.d.}{\sim} \mathcal{U}(-1,1) \end{aligned} \tag{1}$$

where we draw $N = 100$ samples, i.e, $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$. Given the knowledge of Eq. (1), the FE of X_3 is zero because the term $10X_2\mathbb{1}_{X_3>0}$, where X_3 appears, is part of the effect of X_2. In contrast, X_2 relates to Y in two opposite ways, as $-5X_2$ when $X_3 < 0$ and as $5X_2$ otherwise. Therefore, the zero average effect of X_2 after aggregating the two opposites effects, should not erroneously imply that X_2 does not affect Y. However, as demonstrated in Figure 1a (for X_2) and Figure 2a (for X_3) *ALE definition* erroneously indi-

Conclusions

- There seems to be a strong interaction between explainability and uncertainty of ML models
- There also seems to be a growing interest in problems that fall into the intersection of the two fields
- Methods can be roughly grouped into two major categories:
    1. Methods that explain the uncertainty
    2. Methods that quantify the uncertainty of the explanation
- We expect to see further research in this growing field
- Some of them today in our workshop!

Thank you!