Abstract

After the tremendous advances of deep learning and other AI methods, more attention is flowing into other properties of modern approaches, such as interpretability, fairness, etc. combined in frameworks like Responsible AI. Two research directions, namely Explainable AI and Uncertainty Quantification are becoming more and more important, but have been so far never combined and jointly explored. In this paper, I show how both research areas provide potential for combination, why more research should be done in this direction and how this would lead to an increase in trustability in AI systems.

Keywords: Responsible AI, Trust, Explainable AI, Uncertainty Quantification

1 Introduction

Deep Learning methods have achieved tremendous success in almost all disciplines in the field of Artificial Intelligence and Machine Learning. Now, there are signs of a shift away from the goal of improving the recognition performance of the models alone. Frameworks like “Responsible AI” are now coming to the fore, postulating properties like fairness, interpretability, privacy and security of AI applications. In this paper, I will focus on the notion of interpretability in particular, expand interpretability to include “Trust” and show why more research should go into combining the methods of Explainable AI and Uncertainty Quantification. To this end, I will take up these two separate research branches, point out their methods, and show their relatedness in contributing towards trustability - the quality or state of being trustable. There are already research fields such as “Planning under Uncertainty” and “Decision making under Uncertainty” which combine known methods with uncertainty but these either
assume that uncertainties already exist or refer to methods for quantifying uncertainty without having properties such as trustability towards humans as a goal.

No one would question that explainability of approaches is conducive to building trust. But uncertainties and awareness of them also have an impact on people. A number of behavioral and electrophysiological studies indicate “[...] a strong relationship between uncertainty and a key component of cognitive control - outcome monitoring. In particular, it appears that highly uncertain environments tend to increase the recruitment of monitoring processes” [21]. So the more uncertain a situation is for us, the more we try to monitor the outcome. Applied to AI systems, this would mean a stronger need for explanation and predictability of the outcome. Both properties that can be achieved through a combination of Explainable AI and Uncertainty Quantification methods. As the coordinator for “Explainable Learning”, one of the main research focus areas at the ADA Lovelace Center for Analytics, Data and Applications I bring those researchers working on explainability of AI approaches together with those working on Uncertainty Quantification.

In the next section, both branches of research with their definitions and methods will be introduced. In the third section, possible combinations of the two research directions and the importance of their combination will be discussed. The outlook will be a suggestion on how to handle trust in the future.

2 Overview of Explainable AI and Uncertainty Quantification

2.1 Explainable AI

There is no clear definition of Explainable AI. The Defense Advanced Research Projects Agency (DARPA, [12]) formulates the goals as “produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners”. Another definition comes from the organisers of xML Challenge, FICO [9], who see XAI as “an innovation towards opening up the black-box of ML” and as “a challenge to create models and techniques that are both accurate and provide good trustworthy explanation that will satisfy customers’ needs”. Both definitions already contain the term “trust” without going further into it. An overview of the individual methods without concrete approaches can be seen in Figure 1 (for an in-depth survey please refer to [2]).

One way to group Explainable AI methods is to divide them into three categories

1https://www.scs.fraunhofer.de/en/focus-projects/ada-center.html
according to their chronological order of use: “Pre-modelling explainability”, “Explainable Modelling” and “Post-modelling Explainability” [15]. We will leave out the “Pre-modelling explainability” category for this paper, as this deals with understanding the data itself before actually training a model.

- **Explainable modelling**
  For explainable modelling, a component for comprehensibility is already implemented in the model during training. Two types of approaches can be distinguished: “Representation Explaining” and “Explanation Producing Systems” (see lower block in Figure 1). An example for Representation Explaining are Concept Activation Vectors [16], which can be used to show for example how sensitive a prediction of “zebra” is to the presence of stripes. Approaches like Beta VAE [13], from the field of Explanation Producing Systems, try to introduce interpretable latent representations, which means that the latent representations contain human-interpretable information already.

- **Post-Explainability**
  In Post-Explainability, after the decision of the model, the processing is examined up to the decision (see upper block in Figure 1). The different approaches use different types of and representations for explanations. An example for a graphical representation is “Layerwise Relevance Propagation” [18] and is often used with images and shows the user which pixels of the input made what contribution to the decision. An example for a different type of representation is another approach named ANN-DT [22]. It extracts decision trees from neural networks to make the underlying decision processes more comprehensible for humans.

### 2.2 Uncertainty Quantification

One comprehensive description [20] of Uncertainty Quantification (UQ) from the field of applied mathematics is: “UQ involves the quantitative characterisation and management of uncertainty in a broad range of applications. It employs both computational models and observational data, together with theoretical analysis. UQ encompasses many different tasks, including uncertainty propagation, sensitivity analysis, statistical inference and model calibration, decision making under uncertainty, experimental design, and model validation. UQ thus draws upon many foundational ideas and techniques in applied mathematics and statistics (e.g., approximation theory, error estimation, stochastic modelling, and Monte Carlo methods) but focuses these techniques on
complex models, for instance, of physical or socio-technical systems that are primarily accessible through computational simulation. [...]"

This definition also holds for the field of machine learning, since we do almost always deal with complex models. There are two main sources of uncertainty, aleatoric uncertainty and epistemic uncertainty [14]. The former is caused by noise in data or labels, and the latter is caused by data sparsity or out of data distribution. The scope of this paper is not to give a comprehensive overview of uncertainty quantification methods, but to provide a categorisation of selected approaches (see Figure 2) to show their combinability with Explainable AI methods. For a more extensive treatise on uncertainty quantification methods, see [1]. In AI research, Uncertainty Quantification can be used to pursue two goals. First, to calibrate neural networks so that the output confidence reflects the empirical accuracy, and second, for out-of-distribution detection, i.e., to detect whether a new classification example really corresponds to the distribution with which the network was trained.

Uncertainty quantification methods can be divided into two types, discriminative methods and generative methods. Generative methods are used, among other things, to perform out-of-distribution detection [11], or to reconstruct data, such as image reconstruction from noisy and incomplete images [6]. Discriminative methods - the focus of this work - are used to perform e.g. classification, i.e. to assign an input to one or more or no output classes. Here, one can distinguish between model-agnostic approaches, Bayesian and Non-Bayesian methods.

- Model-agnostic methods
Figure 2: Overview of Uncertainty Quantification Methods in Machine Learning

This includes approaches that are independent of the method used. They generate e.g. through data augmentation, better calibrated DNNs (Mixup [23]) or alleviate the over-confidence of the model (CutMix [24]).

- **Bayesian methods**
  They use Bayesian Inference, in which Bayes’ theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Two well-known approaches of this type are Bayes by Backprop [4] and MC Dropout [10].

- **Non-bayesian methods**
  These include approaches that are not based on Bayesian inference. Reasons for this are, e.g., that Bayesian approaches require significant modifications to the training procedure and are computationally expensive compared to non-Bayesian neural networks. One well-known approach is to estimate uncertainty using deep ensembles [17]. In this context, the quantification of Uncertainty comes from sampling from several models, where each model is trained separately on the same data or subsets. The Consensus of the models gives confidence in overall model ensemble.
3 Discussion

In this section, I will address the question of why more research should be done in the direction of joint approaches for Explainable AI and Uncertainty Quantification and will also discuss the desired properties of such an approach. Then two short exemplary possible combinations of methods from Explainable AI and Quantification of Uncertainties will be shown. Finally, I will discuss the concept of trust and highlight why both research areas can jointly contribute to improving trustability.

3.1 Combination of Explainable AI and Uncertainty Quantification

When one speaks of Interpretability, one automatically thinks of Explainable AI methods. The research field of Uncertainty Quantification is almost not present here. In my opinion, however, current research needs to be more concerned with merging the two research directions, as in a sense they show two sides of the same coin. While the methods of Explainable AI try to show the way to the decision, the methods of Quantification of Uncertainty try to give a realistic evaluation regarding the reliability of the decision. Thus, through their respective statements, “I know why I made a decision and can show it to you” and “I know what I don’t know or how certain I am about my decision”, both research directions contribute to paint a more complete picture and thus should end up in their combination being more than the sum of their parts.

Sections 2.1 and 2.2 have shown different methods that differ in their complexity and in their integrability, but can form a basis for combinations. On a meta-level, the result of such a combination should be a model that combines both Explainable AI and Uncertainty Quantification, so that one gets back not only an explanation but also an uncertainty in the decision including an indication of where the uncertainty comes from. This must be done in a way that is appropriate for humans: The nature of the explanations must be both understandable to humans and delivered in an appropriate manner. A few possibilities for this are:

- Concrete explanations: The user is directly provided with the explanation in the form of decision → certainty → explanation. An example of this in an emotion classification task would be the feedback “The person is classified as sad to 30% due to a covering of the mouth.”

- Counterfactual Explanations: The user is shown under which conditions the system would have decided differently and with what degree of certainty. Applied to the example just presented, the feedback would be “If the person’s mouth
was not covered, I would classify the person as 80% happy”. User-studies have shown, for example, that people prefer counterfactual explanations over case-based reasoning, i.e., solving new problems based on the solutions of similar past problems.

The type of representation also has to be understandable to humans. It is very input-specific and must be appropriate and expectable for the user depending on the problem. Some forms of representation that could be considered are:

- **Textual**
  The system provides an explanation in the form of a text in which the uncertainty of the decision is also verbalised. For example, a combination of approaches for generating interpretable latent representations, presented in section 2.1 and “Bayes by Backprop”, presented in section 2.2 would be conceivable.

- **Visual**
  The system provides an explanation in the form of a visualisation with an appropriate representation of uncertainties. One possibility here would be, for example, to link the method LRP, presented in section 2.1 with an elliptic coding of uncertainties obtained by Monte Carlo dropout procedure, presented in section 2.2.

### 3.2 The Role of Trust

Early in the introduction, I talked about trust and its psychological role and necessity to humans. In other disciplines, especially in human-machine interaction, trust has played a major role from the very beginning. Parts of these definitions from other research branches can also be transferred to the requirements of Responsible AI.

In robotics, Lewis et al. distinguish two properties of systems: System predictability and system intelligibility and transparency. The former is important because “[...] knowing that the automation may fail reduces the uncertainty and consequent risk associated with use of the automation. In other words, predictability may be as (or more) important as reliability.” This in combination with “Systems that can explain their reasoning will be more likely to be trusted, since they would be more easily understood by their users [...]” describe in my opinion exactly this combination of Explainable AI (“Systems that can explain their reasoning [...]”) and Uncertainty Quantification (“[...] predictability may be as (or more) important as reliability”).


A system that can tell the user that it is uncertain in combination with a reasoning for it corresponds to my propositions from section 3.1 and should have as a goal an enhancement of trustability in AI systems.

4 Outlook

It is remarkable that both definitions of Explainable AI in section 2.1 address trust, even if it only seems more like trust is a pleasant side effect of explainability. At this point, I wonder if this has led to less attention being paid to linking it to quantification of uncertainty, on the one hand, and whether this classification is appropriate, on the other hand. It would also be conceivable to rename the “Interpretability” pillar in the Responsible AI framework to “Trust” and to downgrade Explainable AI to a method for increasing the trustability of an AI system. Even if the actual idea behind the point “Interpretability” is perhaps to check whether AI systems function as intended, this goal can also be subordinated to the term “Trust” or trust-building measure. Perhaps it will also come about in the near future that one will focus even more on the human component and human acceptance in this case and carry out such a renaming.

Acknowledgements

The author wants to thank Teena Hassan for the helpful discussions.

References

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nakhavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges (2020)

[2] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access 6, 52138–52160 (2018)

[3] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: ‘it’s reducing a human being to a percentage’: Perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI ’18, p. 1–14. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3173574.3173951. URL https://doi.org/10.1145/3173574.3173951
[4] Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: F. Bach, D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 37, pp. 1613–1622. PMLR, Lille, France (2015). URL http://proceedings.mlr.press/v37/blundell15.html

[5] Blythe, J.: An overview of planning under uncertainty. In: Artificial intelligence today, pp. 85–110. Springer (1999)

[6] Böhm, V., Lanusse, F., Seljak, U.: Uncertainty quantification with generative models. arXiv preprint arXiv:1910.10046 (2019)

[7] Dimitrakakis, C., Ortner, R.: Decision making under uncertainty and reinforcement learning (2018)

[8] Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 275–285 (2019)

[9] FICO: Explainable Machine Learning Challenge (2018 (accessed December 28, 2020)). URL https://community.fico.com/s/explainable-machine-learning-challenge

[10] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (2016). URL http://proceedings.mlr.press/v48/gal16.html

[11] Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one (2020)

[12] Gunning, D.: Darpa’s explainable artificial intelligence (xai) program. pp. ii–ii (2019). DOI 10.1145/3301275.3308446

[13] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
[14] Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods (2020)

[15] Khaleghi, B.: The How of Explainable AI: Pre-modelling Explainability (2019 (accessed December 28, 2020)). URL https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4

[16] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, pp. 2668–2677. PMLR (2018)

[17] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems, pp. 6402–6413 (2017)

[18] Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10, e0130140 (2015). DOI 10.1371/journal.pone.0130140

[19] Lewis, M., Sycara, K., Walker, P.: The role of trust in human-robot interaction. In: Foundations of trusted autonomy, pp. 135–159. Springer, Cham (2018)

[20] Marzouk, Y.M., Willcox, K.E.: Uncertainty Quantification, vol. II, chap. 34, pp. 131–134. Princeton University Press (2015)

[21] Mushtaq, F., Bland, A.R., Schaefer, A.: Uncertainty and cognitive control. Frontiers in psychology 2, 249 (2011)

[22] Schmitz, G.P.J., Aldrich, C., Gouws, F.S.: Ann-dt: an algorithm for extraction of decision trees from artificial neural networks. IEEE Transactions on Neural Networks 10(6), 1392–1401 (1999). DOI 10.1109/72.809084

[23] Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems 32, 13888–13899 (2019)

[24] Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6023–6032 (2019)