HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220208

Discovering the Rationale of Decisions

Extended Abstract

Cor STEGING [a] Silja RENOOIJ [b] Bart VERHEIJ [a]
[a] *University of Groningen, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence*
[b] *Utrecht University, Department of Information and Computing Sciences*

In order to create human-centered intelligence, the AI systems that assist humans should behave responsibly and make the right decisions based upon a sound rationale. Previous work has shown that machine learning systems can make the right decisions for the wrong reasons [1,2]. Conventional explainable AI methods are not always able to correctly detect such unwanted behavior [3]. For evaluating the soundness of the decision-making rationale, therefore a new approach is required. In this extended abstract, we summarize and illustrate our proposed hybrid solution: a knowledge-driven, model-agnostic method for rationale evaluation.[1] The method consists of three distinct steps:

1. Measure the accuracy of a trained system, and proceed if the accuracy is sufficiently high;
2. Design dedicated test sets for rationale evaluation targeting selected rationale elements based on expert knowledge of the domain;
3. Evaluate the rationale through the performance of the trained system on these dedicated test sets.

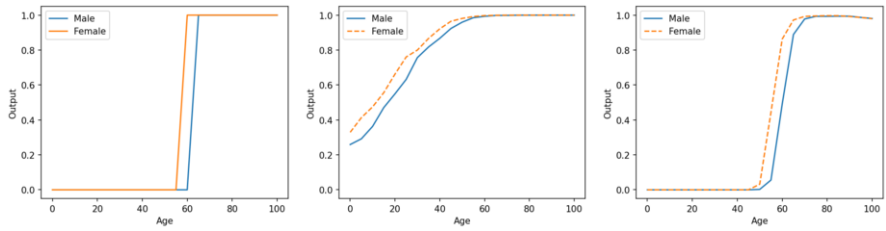
This method is hybrid in two ways: it combines knowledge with data and uses human-in-the-loop domain experts to strengthen the decision-making rationale of the AI. The first step ensures that further efforts are only made if the general performance of the model is already considered good enough by the domain expert. The second step in our method depends on domain knowledge for identifying the rationale elements for which test sets are to be designed. In the third step, performance is again evaluated by considering both accuracy and a comparison between model output and expected output in terms of the dedicated test sets. Information gained from the rationale evaluation method can subsequently be used to improve the decision rationale of the system.

We have applied our method to neural networks trained on artificial datasets for a number of (legal) domains [4,5]. Here we consider one of our experiments [5], which deals with the fictional welfare-benefit domain introduced by Bench-Capon [6], in which pensioners are eligible for a welfare benefit if they satisfy a set of conditions. Using these conditions, artificial datasets are generated; multi-layer perceptrons (MLPs) are trained on these type **A** datasets, and tasked with predicting the eligibility for welfare based on the personal information of fictional pensioners from a similarly generated test set.

[1] Details of the experiments can be found in our ICAIL 2021 and XAILA 2021 contributions [4,5], on which this research abstract is based. This research was funded by the NWO Hybrid Intelligence Gravitation Project.

Table 1. The mean accuracies (with standard deviations) of the MLPs trained and tested on various datasets.

	Regular test set	Test set $\mathbf{C}_1$
Trained on **A**	99.79 ± 0.04	63.24 ± 4.86
Trained on **B**	99.29 ± 0.36	97.66 ± 0.79



(a) $\mathbf{C}_1$: ideal output. (b) $\mathbf{C}_1$: trained on type **A**. (c) $\mathbf{C}_1$: trained on type **B**.

Figure 1. The mean network output on test set $\mathbf{C}_1$ versus the age (a)-(c).

Measuring the accuracy of the MLPs on this regular test set corresponds to the first step of our rationale evaluation method. We now consider one of the conditions that define the domain, condition C_1: "the person must be of pensionable age (60 for women and 65 for men)". To investigate the rationale underlying the trained MLPs, we generate dedicated test set $\mathbf{C}_1$, relying on the knowledge that we, as humans, have about the domain. The test set is designed such that an instance can only be correctly classified if the MLP has learned condition C_1. Further details regarding the MLPs and datasets are described in the original publications [4,5].[2]

Table 1 (top row) displays the average accuracies of the MLPs, trained on a type **A** dataset, tested on the two different test sets. We find a high accuracy on the regular test set when training MLPs on the original type **A** dataset. However, the accuracy on the dedicated test set $\mathbf{C}_1$ created in step 2 of our method is poor: 63.24%. The performance of the MLPs on the dedicated test set is also represented graphically in Figure 1b: the mean output of the MLP is shown versus the age for both men and women. Ideally, these mean outputs are 1 if the condition is satisfied, and 0 otherwise, as depicted in Figure 1a. Condition C_1 is quite poorly approximated by the MLPs when compared to the ideal output. We conclude that the MLPs are unable to learn this condition, despite a high accuracy on the regular test set.

After applying our three-step method we can design an additional type of dataset based upon the results of the evaluation. This type **B** dataset has a different distribution of instances and is used to re-train the MLPs. Training the MLPs on a type **B** dataset, as seen in the bottom row of Table 1, causes a minimal decrease in performance on the regular test set, yet the accuracy on the dedicated test set increases significantly (see also Figure 1c, which is similar to the ideal output). We conclude that the MLPs trained on a type **B** set are able to learn condition C_1.

Our experiments affirm that systems can make the right decisions for the wrong reasons and that our proposed hybrid, knowledge-driven, model-agnostic method can be used to detect and improve an unsound rationale.

[2]The datasets and the Jupyter notebooks used for data generation can be found in a Github repository: https://github.com/CorSteging/DiscoveringTheRationaleOfDecisions

References

[1] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016; 2016. p. 1135-44.

[2] Goodfellow IJ, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples. CoRR. 2015;abs/1412.6572.

[3] Steging C, Renooij S, Verheij B. Rationale Discovery and Explainable AI. In: Schweighofer E, editor. Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021. vol. 346 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2021. p. 225-34. Available from: https://doi.org/10.3233/FAIA210341.

[4] Steging C, Renooij S, Verheij B. Discovering the Rationale of Decisions: Towards a Method for Aligning Learning and Reasoning. In: Maranhão J, Wyner AZ, editors. ICAIL 2021: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021. ACM; 2021. p. 235-9. Available from: https://doi.org/10.1145/3462757.3466059.

[5] Steging C, Renooij S, Verheij B. Discovering the Rationale of Decisions: Experiments on Aligning Learning and Reasoning. In: 4th EXplainable AI in Law Workshop (XAILA 2021). ACM; 2021. Available from: https://arxiv.org/abs/2105.06758.

[6] Bench-Capon T. Neural Networks and Open Texture. In: Proceedings of the 4th International Conference on Artificial Intelligence and Law. ICAIL 1993. New York, NY, USA: ACM, New York; 1993. p. 292-7.