Rationale Discovery and Explainable AI

Steging, Cor; Renooij, Silja; Verheij, Bart

Published in:
Legal Knowledge and Information Systems - JURIX 2021

DOI:
10.3233/FAIA210341

Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210341

Rationale Discovery and Explainable AI

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

Abstract. The justification of an algorithm's outcomes is important in many domains, and in particular in the law. However, previous research has shown that machine learning systems can make the right decisions for the wrong reasons: despite high accuracies, not all of the conditions that define the domain of the training data are learned. In this study, we investigate what the system *does* learn, using state-of-the-art explainable AI techniques. With the use of SHAP and LIME, we are able to show which features impact the decision making process and how the impact changes with different distributions of the training data. However, our results also show that even high accuracy and good relevant feature detection are no guarantee for a sound rationale. Hence these state-of-the-art explainable AI techniques cannot be used to fully expose unsound rationales, further advocating the need for a separate method for rationale evaluation.

Keywords. Machine Learning, Explainable AI, Knowledge, Data

1. Introduction

Explanations are essential in AI and law [1,2]. Not only is argumentative reason-based discussion inherent to legal reasoning, but both parties in a court of law also have the right to an explanation when a decision is made [3]. In brief, proper justice requires decisions based on sound rationales. Various types of explanation [4] have been applied in the field of AI and law, ranging from contrastive explanations [5,6,7] to selective explanations [8, 9], and from probabilistic explanations [10] to social explanations [11,12,8].

Many state-of-the-art AI systems, however, are black-box systems that reason without transparency. As long as they cannot explain their decision making, they are inherently unsuitable in the context of AI & law. This is unfortunate, as the performance of in particular deep learning systems is often second to none when it comes to tasks such as image, speech or text classification. The subfield of Explainable AI (XAI) aims to bridge the gap that black-box machine learning systems have created, by providing explanations for these opaque systems [13]. Methods such as LIME [14] and SHAP [15] allow us to look 'inside' the black-box by demonstrating which parts of the input are important in the system's decision making process. In a recent study, lawyers tasked with assessing both LIME and SHAP in a legal text classification task graded both explanation methods similarly, and said to look forward to more explainable systems to assist their work [16]. Other recent research on the use of machine learning in law, on the other hand, emphasizes that explaining the outcome using a list of important input features is insufficient in a legal setting

[17]. In addition to instilling trust in the end user, explainable AI can also expose unsound decision making. In the LIME paper, for example, a husky is misclassified as a wolf because there is snow in the background of the image [14]. In adversarial attacks, small perturbations to an image, imperceptible to the human eye, can cause drastic changes in a model's prediction [18]. These systems had high accuracy scores, but they performed well for the wrong reasons, as their decision making was unsound.

Based on work by Bench-Capon in AI & Law [19], we introduced a preliminary method for rationale evaluation [20]. This method tests the decision making of a system and evaluates to what extent the learned rationale of the system is sound, given high accuracy scores. This method was applied in a set of experiments dealing with legal domains, and it was shown that high accuracies are by no means a guarantee for a sound rationale [21]. Different studies using the same domain and datasets showed similar results when it comes to learning the complete rationale [22,23]. Knowing whether a system learned what it is supposed to learn is essential in pursuing responsible AI. In this paper we investigate whether state-of-the art explainable AI techniques provide relevant insight into the evaluation of the rationale of a machine learning system. In particular, we will use SHAP [15] and LIME [14] to extract explanations from neural networks trained on the fictional welfare benefit domain [19]. We then compare these explanations to the results of the rationale evaluation from previous research.

2. Background

The domain used in both this study and in previous studies [20,22,23] is the welfare benefit domain, as introduced by Bench-Capon [19]. It defines a fictional set of conditions, that must all be satisfied in order for a pensioner to receive benefits for visiting their spouse in the hospital. Eligibility for a welfare benefit depends on six conditions and is formalized as follows:

$$
\begin{aligned}
&Eligible(x) \iff C_1(x) \wedge C_2(x) \wedge C_3(x) \wedge C_4(x) \wedge C_5(x) \wedge C_6(x)\\
&C_1(x) \iff (Gender(x) = female \wedge Age(x) \geq 60) \vee\\
&\qquad\quad (Gender(x) = male \wedge Age(x) \geq 65)\\
&C_2(x) \iff |Con_1(x), Con_2(x), Con_3(x), Con_4(x), Con_5(x)| \geq 4\\
&C_3(x) \iff Spouse(x)\\
&C_4(x) \iff \neg Absent(x)\\
&C_5(x) \iff \neg Resources(x) \geq 3000\\
&C_6(x) \iff (Type(x) = in \wedge Distance(x) < 50) \vee (Type(x) = out \wedge Distance(x) \geq 50)
\end{aligned}
$$

In this domain a pensioner is therefore eligible for a benefit iff he or she is of pensionable age (60 for a woman, 65 for a man), has paid four out of the last five contributions Con_i, is the spouse of the patient, is not absent from the UK, does not have capital resources amounting to more than £3,000, and lives at a distance of less than 50 miles from the hospital if the relative is an *in*-patient, or beyond that for an *out*-patient.

Artificial datasets were generated based on this domain and used to train neural networks. These networks are then tasked with classifying new, unseen instances from the welfare benefit domain. In addition to measuring the performance of the system in terms of accuracy, the main interest lies in evaluating the rationale that the networks have learned; were they able to internalize the six conditions that define eligibility? To answer

that question, a preliminary method for rationale evaluation was introduced and applied to the welfare benefit domain [20]. The rationale evaluation method prescribes designing dedicated test sets that target specific elements of the desired rationale, based on expert knowledge of the domain. Learning systems can only perform well on these dedicated test sets if they have learned a particular element of the rationale.

2.1. Datasets

This study builds on the same datasets used in previous work [19,21], generated from the welfare benefit domain[1]. Every dataset contains 64 features, made up of the 12 features from the domain and 52 noise features. We use both relatively small training datasets of 2,400 instances and larger datasets that consist of 50,000 instances.

Two types of datasets are used: type A and type B. Both types of datasets have a balanced label distribution, wherein 50% of the instances are eligible, satisfying *all* six conditions, and 50% are ineligible. Eligible instances are generated randomly from uniform distributions in both type A and type B datasets, such that all six conditions are satisfied. In type A datasets, ineligible instances are generated such that an equal number of them fail on each condition. In other words, each condition is responsible for the ineligibility of an equal number of instances. The values of the remaining features are generated completely randomly from a uniform distribution. As a result, multiple conditions can be unsatisfied if an instance is ineligible. Type A datasets are the most realistic dataset, since very little of the distribution is controlled, just as in non-artificial datasets. In type B datasets, ineligible instances only fail due to a single condition. It is therefore not possible to have multiple unsatisfied conditions in a type B dataset. For this more controlled version of the domain, it is harder for a network to achieve high performance scores, since it has to learn each condition individually. In a type A dataset, it is possible to achieve a theoretical accuracy of 98.95% while only knowing four out of the six conditions that define the domain [19], because ineligible instances almost always fail on multiple conditions. This is not possible in type B datasets.

The method for rationale evaluation prescribes the design of dedicated test sets for rationale evaluation, that target specific components of the rationale based on expert knowledge of the domain. For the welfare benefit domain, to measure how well any given condition C_i is learned, a dedicated test set is created in which every condition is satisfied except for C_i. The values of the features that define C_i are then generated across their full range of values. That way, the eligibility in this dedicated test set is determined solely by condition C_i: condition C_i is satisfied iff the instance is eligible. Networks are only able to classify the instances from this dedicated test set correctly if they have learned condition C_i. The particular components of the rationale that we investigate are the Age-Gender condition (condition C_1) and the Patient-Distance condition (condition C_6). The dedicated test sets to evaluate these conditions are referred to as the Age-Gender and Patient-Distance datasets, respectively.

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

Table 1. The accuracies obtained by the neural networks in the welfare benefit domain.

	Test set A	Test set B	Age-Gender	Patient-Distance
Trained on A (2,400)	98.97±0.19	72.39±1.66	52.14±4.01	50.05±0.09
Trained on B (2,400)	96.13±0.66	90.51±1.25	86.4±1.33	85.77±5.21
Trained on A (50,000)	99.8±0.03	80.98±1.47	60.22±3.87	64.44±2.87
Trained on B (50,000)	99.64±0.17	98.53±0.15	98.51±0.47	97.17±0.46

2.2. Neural networks

Multilayer perceptrons (MLP) with a single, two and three hidden layers were used, as in the initial study [19]. The goal of the study was not to achieve the highest accuracies, but rather to investigate and evaluate the learned rationale. Since MLP's are simple, black-box systems, they are ideal candidates for research into rationale evaluation. The details regarding the architecture and parameters of the networks can be found in [20]. Each of the three networks is trained separately on both type A and type B datasets and evaluated using separate type A, type B, Age-Gender and Patient-Distance datasets.

2.3. Previous results

Table 1 shows the accuracies of the neural networks with a single hidden layer from previous research, trained and tested on the various datasets. Networks with two or three hidden layers scored similarly. When training on a type A dataset, accuracies on the 'normal' test set A are high, while performance on test set B, the Age-Gender and Patient-Distance dataset is much lower. This shows that the correct rationale has not been learned, since a correct rationale would lead to high performance on all data sets. It is therefore clear that high accuracies are no guarantee for a sound rationale. When training on a type B dataset, the accuracies remain high on type A test sets, but improve drastically on test set B, the Age-Gender and the Patient-Distance dataset; training on a type B dataset therefore yields a better learned rationale. The table shows that with more data, the same pattern can be observed, while training on a type B test set then leads to even better accuracy scores on all data sets.

3. Explainable AI

3.1. Previous results

Our current objective is to investigate whether explainable AI techniques can be used to discover the unsound rationale as it is actually used by the trained system. Earlier research attempted to discover what rules a network trained on this domain had learned by inverting the trained network, in the sense that the output node node becomes the input node and the input nodes become the output node [19]. Passing a value of 1 through the new 'input' node would yield a list of features and their relevance. Since all of the features were normalized between 0 an 1, relevance was described as its deviation from 0.5.

However, this approach cannot account for the impact that a combination of features has, as pointed out in the original study [19]. A clear example of the shortcomings of this method is the Patient-Distance condition C_6, which is a variation of a XOR problem.

A XOR function yields true if and only if exactly one of its two variables is true and the other is false. If the first variable is true, the output will therefore be true in 50% of the cases (whenever the other variable is false). In the remaining 50% of the cases, the output will be false (whenever the other variable is true). Applying the method of inverting the network to a network that can solve a XOR problem would therefore yield an output value of 0.5 for both features, since each feature yield true (1) half of the time and false (0) in the other cases. This result provides little insight into the structure of a XOR problem, since the two features depend on each other and are meaningless on their own. Inverting the network does not account for such combinations of features, and is therefore unsuitable for our current objective.

3.2. State of the art: SHAP & LIME

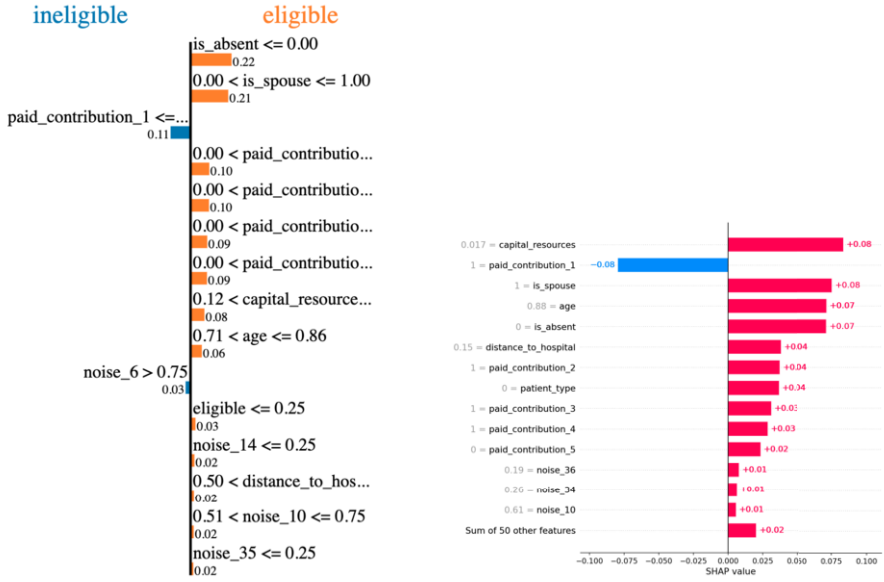
Since the earlier research [19] much progress has been made in the availability of explainable AI methods. Hence now we extend the previous research by investigating the trained neural networks using modern, state-of-the-art explainable AI techniques, to see whether these methods allow us to evaluate the soundness of the rationale. In particular, we will use SHAP [15] and LIME [14], two of the most commonly used explainable AI methods. SHAP (SHapley Additive exPlanation) is an explainable AI framework, that explains the output of a machine learning system based on the idea of Shapley values from game theory. LIME creates explanations by perturbing individual instances and using those to learn interpretable sparse linear models that approximate the system's decision making.

These two explanation methods were chosen due to their inherent fidelity [14] to the decision making of the black-box model, meaning that their explanations should accurately reflect the opaque decision making of the model. Since our aim is to investigate what the model has actually learned, fidelity is the most important criterion. Though it is often impossible to create completely faithful explanations, local fidelity (how a model responds to a given instance) can be achieved. Both LIME and SHAP are local explanation methods that provide an explanation for the output produced given a single input instance. Additionally, SHAP includes methods to aggregate a set of local explanations into a global interpretation of a system's decision making. Both methods are additive methods, meaning that summing up the effects of all feature attributions should approximate the prediction of the network. We will apply both LIME and SHAP to our trained networks, in order to ensure that our results are explainer-independent.

The same three neural networks mentioned in Section 2.3 are trained on type A and type B datasets separately, using both 2,400 instances and 50,000 instances. LIME and SHAP are then used to extract explanations from networks using a test set of 500 instances, sampled from a separate type A testing dataset.

Table 2. Example instance

Age	**Gender**	**Con$_1$**	**Con$_2$**	**Con$_3$**	**Con$_4$**	**Con$_5$**	**Spouse**	**Absent**	**Resources**	**Type**	**Distance**
84	female	0	1	1	1	1	1	0	1569	out	74



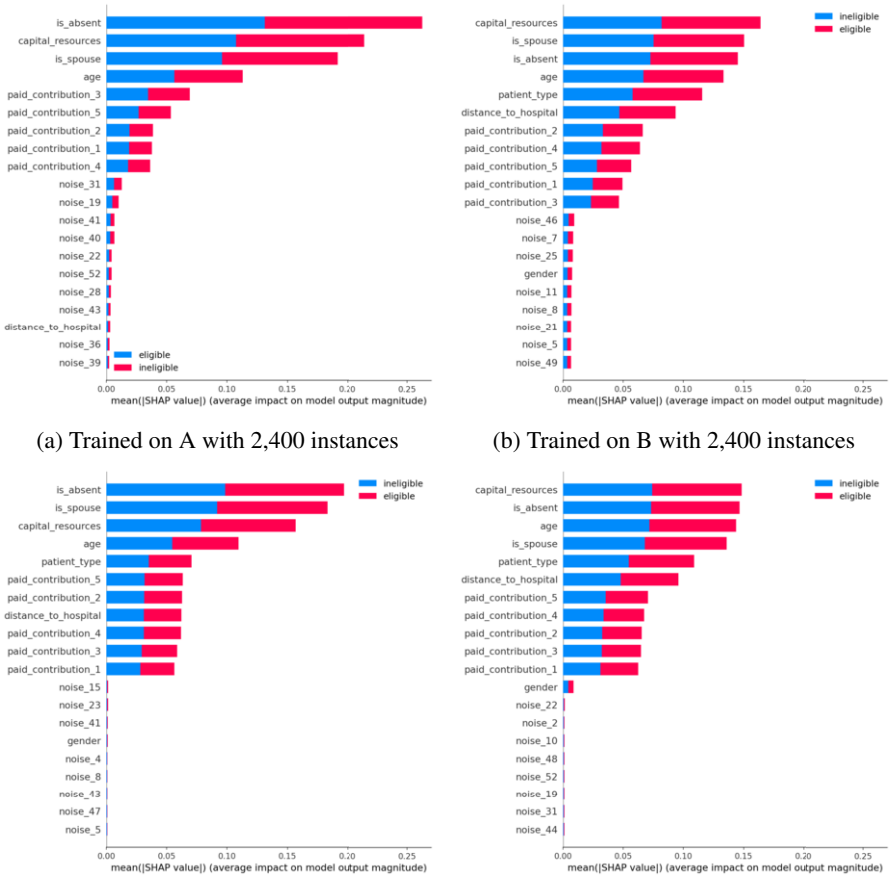
(a) LIME (b) SHAP

Figure 1. LIME and SHAP bar plots of the network trained on large type A dataset, displaying the impact of each feature in the classification process towards the 'ineligible' label (blue) or the 'eligible' label (orange/red) of the instance in Table 2.

4. Results

To illustrate the explanation methods, we first examine the classification process and its explanations using the example instance found in Table 2. In this example, all six conditions are satisfied and therefore the instance is eligible for a welfare benefit. The network with a single hidden layer trained on a dataset of 50,000 type A instances correctly predicts this. Figure 1 shows the LIME and SHAP explanations of these networks for the given instance. The bar plots display the impact that each feature had on classifying that instance, ranked from highest to lowest. Blue bars on the left contribute to the 'ineligible' label, whereas the orange and red bars on the right contribute to an 'eligible' label. Running the same experiment using networks with multiple hidden layers showed similar results and are therefore omitted. Note that even though each instance has 64 features, only the top 15 features with the highest impact are shown.

We now consider the explanations for the entire test set of 500 instances to illustrate the global interpretations of the networks. Since LIME does not possess a method for aggregating sets of explanations, we will use SHAP's summary plots. To ensure that the results are explainer-independent, we also examined 10 LIME cases for each scenario (each combination of network, training data type and size; 120 in total). The LIME results do not contradict any of the SHAP results and support its global interpretations of the networks. Figure 2 shows the average SHAP values for networks with a single hidden layer, trained on both training set A (left) and training set B (right) for training sets with 2,400 (top) and 50,000 instances (bottom). As mentioned in Section 2.3, experiments using networks with more hidden layers yielded similar results and are therefore omitted.



(a) Trained on A with 2,400 instances

(b) Trained on B with 2,400 instances

(c) Trained on A with 50,000 instances

(d) Trained on B with 50,000 instances

Figure 2. SHAP bar plots of the network trained on various datasets, displaying the impact of each feature in the classification process towards the 'ineligible' label (blue) or the 'eligible' label (red).

These bar plots display the top 20 features that have the highest average impact on the classification process, ranked from highest to lowest.

5. Discussion

If the networks have learned the correct rationale, the relevant features should have a high impact in Figures 1 and 2. These relevant features are the age, gender, contributions, spouse, absent, resources, distance and patient-type features, since they determine eligibility. The 52 noise variables should have a low impact as they do not contribute anything to the desired outcome.

The example instance in Table 2 satisfies all conditions and is therefore eligible. Since the network correctly predicted its eligibility label, we would expect to see that all relevant features in the SHAP and LIME plots have a high impact, whereas the noise features would have a low impact. In the plots of Figure 1 we indeed see that most rel-

evant features have been attributed a high impact. Noticeable exceptions are, however, the gender, patient-type and distance features. In this particular example, gender is irrelevant, as the person is older (88) than the threshold for both males (65) and females (60). The patient-distance condition is difficult to learn, as made evident from Table 1, which can account for the low impact values of these features. In the example instance, only one feature had a noticeable impact towards the 'ineligible' label, rather than the 'eligible' label. This is the first paid contribution feature, which makes sense, as the first contribution was not paid in this case (see Table 2) and would thus act as evidence against eligibility.

Since each example instance represents a different case, it is evaluated differently in terms of feature importance. To get a broader evaluation of the reasoning of the networks we therefore examine the global interpretations. The SHAP results from Figure 2a show us the impact of each feature in the classification process of a network trained on a type A dataset with 2,400 instances. We can see that most of the relevant features are listed with a high impact, whereas the noise features have a low impact. Given the large amount of noise variables (52 noise variables versus 12 relevant variables), discovering the relevant features is a non-trivial outcome, supporting the value of SHAP. However, the gender and patient-type features are missing from the plot, and the distance feature has an impact value similar to those of the noise features. This supports our previous findings (see Table 1) that the correct rationale was not learned after training on a smaller type A dataset [21]. SHAP has thus managed to detect the unsound rationale: some features that are relevant in the domain do not have high impact in the trained network's decision process.

The rationale improves when training on a type B dataset of the same size. Figure 2b supports this observation: all of the 12 relevant features are deemed to be highly impactful in the classification process by SHAP. The exception here is the gender feature, which is given a relatively low SHAP value. This finding fits the fact that in our domain the gender feature is only important in a very small subset of cases (males between the age of 60 and 65). From Table 1 we know that the conditions were not learned perfectly, and the low impact assigned to the age variable reveals a possible reason for this observation.

When training on a larger type A dataset, accuracies improve though the rationale is still not sound, as evident from the low accuracy scores on the Age-Gender and Patient-Distance datasets (see Table 1). The SHAP plot of this network, as shown in Figure 2c, displays lower impact values for the noise variables compared to the plot in Figure 2a. Moreover, SHAP now assigns high impact values to the patient-type and distance features, which it did not do when training on a smaller type A dataset. The high impact scores on patient-type and distance shows that the network has discovered the relevance of these features, which could explain the increase in performance on the Patient-Distance dataset when training on a larger type A dataset when compared to a smaller type A dataset (64.44% versus 50.05% as seen in Table 1). The system seemingly makes use of the patient-type and distance features, hence the high impact value returned by SHAP for these features, but it does not use them correctly as in the domain representation. In this case, therefore, SHAP was not able to detect the unsound rationale.

The impact scores of the networks trained on a larger type B dataset are similar to those of the networks trained on a smaller type A dataset (Figure 2d). The impact of the noise features is smaller than on a smaller type B dataset, however, and the gender feature now has a relatively higher impact than the noise features. Based on the four graphs

in Figure 2, the networks trained on a larger type B dataset provide the most desirable impact distribution, with the highest impact for relevant features and the lowest impact for the noise features. This is supported by previous results, as networks trained on larger type B datasets provided the highest accuracies on all datasets (Table 1), suggesting the most sound rationale.

Summarizing these results, we find that the unsound rationale we discoverd previously using the rationale evaluation method, is clearly exposed in the SHAP values for networks trained on less, and perhaps insufficient, data points (Figure 2a). This suggests that explainable AI methods can be used to evaluate the rationale of trained systems to some extent. When training on more data points, however, the unsoundness of the rationale is not as clearly exposed. Figure 2c assigns high impact values to all of the relevant features (except for gender, though its small impact can be accounted for). This makes it seem as if the rationale is sound, whereas the method for rationale evaluation has shown that it is not (see Table 1). Based solely on the SHAP and LIME explanations, we would not be able to know that the rationale is unsound for the networks trained on a large type A dataset. Therefore, even though the XAI methods can expose a faulty rationale, they cannot guarantee a sound rationale. Previously research claimed that it is possible to make the right decisions without knowing why [19,20]. These results expand upon that notion, and suggest that it is also possible to know what is important without knowing why. In other words, we find that systems that have both a high accuracy and assign high importance to the correct features are still not guaranteed to use a sound rationale.

6. Conclusion

Previous research [20,21] has shown through a method for rationale evaluation that learning systems can achieve high accuracies without having a sound rationale. The current study set out to investigate the rationale actually used by a trained system using explainable AI (XAI) methods. Earlier research [19] used a preliminary XAI method with the aim of discovering the rules of a neural network with limited success, as rules with non-straightforward combinations of factors are difficult to discover. In the present study, we used modern XAI techniques (SHAP [15] and LIME [14]) to identify the features of the dataset with the largest impact on the classification process. Though the exact relationship between the features is still unclear, the explanation methods exposed the unsound rationales of some of the systems, reaffirming earlier findings [19,20] using a state-of-the-art explanation method. This suggests that XAI techniques can be used to evaluate the rationale of a system. However, for some conditions, the XAI methods did *not* detect an unsound rationale when the rationale was known to be unsound. Neither high accuracies, nor acceptable explanations from XAI techniques therefore can guarantee a sound rationale. This finding further supports the need for a rationale-evaluation method in order to obtain responsible AI.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

References

[1] Verheij B. Artificial intelligence as law. Artificial Intelligence and Law. 2020;28(2):181-206.
[2] Atkinson K, Bench-Capon T, Bollegala D. Explanation in AI and law: Past, present and future. Artificial Intelligence. 2020;289.
[3] Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman C, O'Brien D, et al. Accountability of AI under the law: The role of explanation. SSRN Electronic Journal. 2017 November.
[4] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019;267:1-38.
[5] Rissland EL, Ashley KD. A case-based system for trade secrets law. In: Proceedings of the 1st International Conference on Artificial Intelligence and Law. ICAIL '87. New York, NY, USA: ACM; 1987. p. 60-6.
[6] Ashley KD. Modeling Legal Arguments: Reasoning with Cases and Hypotheticals. Cambridge (Massachusetts): The MIT Press; 1990.
[7] Verheij B. Artificial Argument Assistants for Defeasible Argumentation. Artificial Intelligence. 2003;150(1–2):291-324.
[8] Atkinson K, Bench-Capon T, Bex F, Gordon TF, Prakken H, Sartor G, et al. In memoriam Douglas N. Walton: the influence of Doug Walton on AI and law. Artificial Intelligence and Law. 2020:1-46.
[9] Verheij B. Dialectical argumentation with argumentation schemes: An approach to legal logic. Artificial intelligence and Law. 2003;11(2-3):167-95.
[10] Vlek CS, Prakken H, Renooij S, Verheij B. A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law. 2016;24(3):285-324.
[11] Hage JC, Leenes R, Lodder AR. Hard cases: a procedural approach. Artificial Intelligence and Law. 1993;2(2):113-67.
[12] Gordon TF. The Pleadings Game: An Artificial Intelligence Model of Procedural Justice. Dordrecht: Kluwer; 1995.
[13] Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable artificial intelligence. Science Robotics. 2019;4(37).
[14] 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; 2016. p. 1135-44.
[15] Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765-74.
[16] Górski Ł, Ramakrishna S. Explainable artificial intelligence, lawyer's perspective. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law. ICAIL '21; 2021. p. 60-8.
[17] Mumford J, Atkinson K, Bench-Capon T. Machine learning and legal argument. In: Proceedings of the 21st Workshop on Computational Models of Natural Argument; 2021. p. 47-56.
[18] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations; 2015. .
[19] Bench-Capon T. Neural networks and open texture. In: Proceedings of the 4th International Conference on Artificial Intelligence and Law. ICAIL 1993. ACM, New York; 1993. p. 292-7.
[20] Steging C, Renooij S, Verheij B. Discovering the rationale of decisions: Towards a method for aligning Learning and reasoning. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law. ICAIL '21. ACM, New York; 2021. p. 235-9.
[21] Steging C, Renooij S, Verheij B. Discovering the rationale of decisions: Experiments on aligning Learning and reasoning. In: The Explainable & Responsible AI in Law (XAILA) Workshop; 2021. .
[22] Možina M, Žabkar J, Bench-Capon T, Bratko I. Argument based machine learning applied to law. Artificial Intelligence and Law. 2005;13(1):53-73.
[23] Wardeh M, Bench-Capon T, Coenen F. Padua: a protocol for argumentation dialogue using association rules. Artificial Intelligence and Law. 2009;17(3):183-215.