Discovering the Rationale of Decisions

Steging, Cor; Renooij, Silja; Verheij, Bart

Published in:
Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law

DOI:
10.1145/3462757.3466059

Publication date:
2021

Discovering the Rationale of Decisions: Towards a Method for Aligning Learning and Reasoning

Cor Steging
c.c.steging@rug.nl
Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen

Silja Renooij
s.renooij@uu.nl
Department of Information and Computing Sciences, Utrecht University

Bart Verheij
bart.verheij@rug.nl
Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen

ABSTRACT

In AI and law, systems that are designed for decision support should be explainable when pursuing justice. In order for these systems to be fair and responsible, they should make correct decisions and make them using a sound and transparent rationale. In this paper, we introduce a knowledge-driven method for model-agnostic rationale evaluation using dedicated test cases, similar to unit-testing in professional software development. We apply this new quantitative human-in-the-loop method in a machine learning experiment aimed at extracting known knowledge structures from artificial datasets from a real-life legal setting. We show that our method allows us to analyze the rationale of black box machine learning systems by assessing which rationale elements are learned or not. Furthermore, we show that the rationale can be adjusted using tailor-made training data based on the results of the rationale evaluation.

CCS CONCEPTS

• **Applied computing** → **Law**.

KEYWORDS

Learning knowledge from data, Explainable AI, Responsible AI, Machine Learning

ACM Reference Format:
Cor Steging, Silja Renooij, and Bart Verheij. 2021. Discovering the Rationale of Decisions: Towards a Method for Aligning Learning and Reasoning. In *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21), June 21–25, 2021, São Paulo, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3462757.3466059

1 INTRODUCTION

In AI and Law, explainability is a key requirement in system design, due to the need for the justification of decisions. For machine-supported decisions, this is encoded in the GDPR's right to explanation. Four types of explanations can be distinguished, all of which have been applied to AI and Law [Atkinson et al. 2020a]: contrastive explanations [Ashley 1990; Rissland and Ashley 1987; Verheij 2003a], selective explanations [Atkinson et al. 2020b; Verheij 2003b], probabilistic explanations [Vlek et al. 2016] and social explanations [Atkinson et al. 2020b; Gordon 1995; Hage et al. 1993].

This requirement of explainability is problematic for the application of central machine learning techniques in law. Neural networks, for example, are known to perform well, but behave like a black box algorithm. Hence, explanation techniques have been developed to 'open the black box' (cf. LIME [Ribeiro et al. 2016], SHAP [Lundberg and Lee 2017]). Even in the domain of vision (where the successes of deep learning are especially significant), the necessity of such methods is underpinned by studies regarding adversarial attacks that show that slight perturbations of images, invisible to the human observer, can radically change the outcome of a classifier [Goodfellow et al. 2015].

In this paper, we therefore evaluate black box machine learning methods with a focus on proper explainability, and not only in terms of accuracy as in the standard machine learning protocol. We are in particular interested in evaluating the discovered rationale underlying decisions, where the rationale is the knowledge structure that can justify a decision, such as the rule applied. We aim to measure the quality of rationale discovery, with an eye on the possibility of improving rationale discovery.

Our work builds on a study investigating whether neural networks are able to tackle open texture problems [Bench-Capon 1993] (also investigated in [Možina et al. 2005; Wardeh et al. 2009]). To measure and possibly improve rationale discovery, we create dedicated test datasets, on which a machine learning system can only perform well if it has learned a particular component of the knowledge structure that defined the data. The idea is similar to how unit testing works in professional software development: we define a set of cases, targeting a specific component, in which we know what the answer should be, and compare that to the output that the system gives.

In order to focus on what is methodologically feasible, we do not use natural language corpora (such as conceptual retrieval [Grabmair et al. 2015], argument mining [Mochales Palau and Moens 2009; Wyner et al. 2010] or case prediction [Ashley 2019; Brüninghaus and Ashley 2003; Medvedeva et al. 2019]). Instead we work with datasets of artificial decisions with known underlying generating rationale. Other earlier discussions of neural networks in law are [Hunter 1999; Philipps and Sartor 1999; Stranieri et al. 1999].

2 REPLICATION EXPERIMENT

The first step towards developing our method for rationale evaluation was replicating the study by Bench-Capon [1993]. This was done using modern, widely-used neural network methods and with

ICAIL'21, June 21–25, 2021, São Paulo, Brazil

ACM ISBN 978-1-4503-8526-8/21/06.
https://doi.org/10.1145/3462757.3466059

significantly larger datasets, in order to reaffirm that the claims made in 1993 still hold today. The study introduces a fictional legal domain, where the eligibility for a welfare benefit for elderly citizens is determined by the conjunction of six independent conditions. Artificial datasets were generated specifying personal information of elderly citizens with their eligibility for the welfare benefit. Multilayer perceptrons were trained and tested on these datasets, and managed to achieve high accuracy scores (above 98%).

Using special test datasets, it was shown that the neural networks were unable to properly learn the first and the last of the six conditions. Furthermore, the networks performed significantly worse when ineligibility was caused by the failure of only a single condition. The training data was therefore altered such that ineligible people only failed on a single condition, rather than on multiple conditions as in the original training dataset. By making these adjustments to the training dataset, the neural networks were able to learn conditions more adequately, while maintaining similar accuracy scores. However, even after adjustment, the conditions that defined the data were not learned perfectly.

In our replication, we discovered that even with more data and modern, commonly used neural networks, the nets are still unable to learn all six conditions that define eligibility, despite high accuracies (99%). Using two dedicated datasets (as also defined in [Bench-Capon 1993]), it was shown that the nets still did not learn condition the first and last condition. The rationale of the nets is therefore not sound, despite high accuracies. Just as in the original study, adjusting the training data based on expert knowledge of the domain significantly improves the rationale of the net without seriously impacting the accuracy.

Additionally, we created a simplified version of the domain, containing only the first and last condition, to see how well the networks are able to extract a simplified rationale. In this domain, the networks *were* able to learn both conditions.

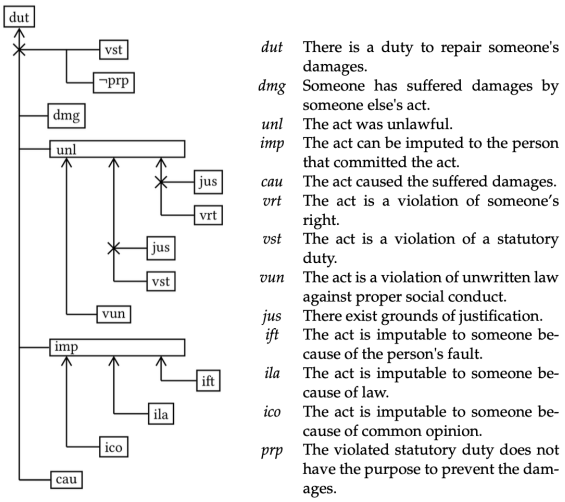
The methods and results of the replication experiment as well as its variations, can be found in detail in [Steging et al. 2021].

3 TORT LAW: DOMAIN AND DATASETS

Following up on the fictional welfare benefit domain, we study a non-fictional legal setting, namely Dutch tort law. This domain uses only Boolean variables, but allows for exceptions to underlying rules. This section describes the underlying knowledge structure of the Tort Law domain using logic, from which we will generate datasets to train a series of neural networks. These networks will subsequently be analysed using a method we propose for assessing the quality of their rational discovery. To this end we need two types of datasets for the purpose of testing. The first are standard test sets sampled from the complete domain to evaluate the accuracy of the networks. The second type is a dedicated test set designed to target a specific aspect of the domain knowledge. This section describes all datasets we use.

3.1 Domain

Our domain concerns Dutch tort law: articles 6:162 and 6:163 of the Dutch civil code that describe when a wrongful act is committed and resulting damages must be repaired. This 'duty to repair' (*dut*) can be formalised as follows:



(A) Arguments and their attacks in the domain of Dutch tort law. (B) Elementary propositions in the domain of Dutch tort law.

Figure 1: Arguments and attacks (A) and their elementary propositions (B) in Dutch tort law [Verheij 2017].

$$
\begin{aligned}
dut(x) &\iff c_1(x) \wedge c_2(x) \wedge c_3(x) \wedge c_4(x) \wedge c_5(x)\\
c_1(x) &\iff cau(x)\\
c_2(x) &\iff ico(x) \vee ila(x) \vee ift(x)\\
c_3(x) &\iff vun(x) \vee (vst(x) \wedge \neg jus(x)) \vee (vrt(x) \wedge \neg jus(x))\\
c_4(x) &\iff dmg(x)\\
c_5(x) &\iff \neg(vst(x) \wedge \neg prp(x))
\end{aligned}
$$

where the elementary propositions are provided alongside an argumentative model of the law in Figure 1 [Verheij 2017], and conditions c_2 and c_3 capture the legal notions of unlawfulness (*unl*) and imputability (*imp*) respectively.

Compared to the fictional welfare domain in [Bench-Capon 1993] and our replication variations [Steging et al. 2021], the Dutch tort law domain is captured in 5 conditions for duty to repair (*dut*), based upon 10 Boolean features. Each condition is a disjunction of one or more features, possibly with exceptions. The feature capturing a violation of a statutory duty (*vst*) is present in both condition c_3 and c_5, rendering these dependent.

3.2 Datasets

We generate four different types of datasets, each for different purposes.[1] For most types of datasets, the generating process is at least partly stochastic and repeated for every repetition of an experiment. Using the same *type* of dataset, for example in training and testing a neural network, does therefore not mean that the exact *same* dataset was used in both training and testing. Table 1 shows an overview of the datasets of the tort law domain.

With 10 Boolean features there are $2^{10} = 1024$ possible unique cases that can be generated from the argumentation structure of the tort law domain in Figure 1. Each case has a corresponding outcome for *dut*, indicating whether or not there is a duty to repair someone's damages.

[1]The Jupyter notebooks used for generating the data can be found in the following Github repository: *https://github.com/CorSteging/DiscoveringTheRationaleOfDecisions*

Table 1: An overview of the tort law datasets. Datasets marked with an asterisk are used for testing purposes only. For each type of dataset, the size and label distribution is given.

Dataset	Size	T/F label distribution
Regular	5,000/500	50%/50%
Unique*	1024	10.94%/89.06%
Unlawfulness*	168	66.67%/33.33%
Imputability*	128	87.5%/12.5%

The *unique dataset* contains these 1024 unique instances for the 10 features plus the label. In this dataset, there are 912 instances where *dut* is false and 112 instances where *dut* is true (10.94%).

The *regular type datasets* are generated such that *dut* is true in exactly half of the instances. The sets are regular in the sense that balanced label distributions are common in machine learning problems. These regular datasets are generated by sampling uniformly from the subset of cases from the unique dataset, such that each possible case is represented equally within the 50/50 label distribution. In a typical machine learning experiment, only a subset of the possible cases is typically available and presented to a network, upon which the network will have to learn to generalize to all possible cases. In addition to generating regular type datasets with 5,000 cases, we therefore also generate smaller regular type datasets with only 500 instances; the latter contains 35.35% of the unique instances.

In the tort law domain we focus on the notions of unlawfulness (c_2) and imputability (c_3) to assess whether the networks are able to discover conditions in the data. For each of the two conditions, we create a dedicated dataset.

The *Unlawfulness dataset* is the subset of the unique dataset in which the features for the unlawfulness condition c_2 can take on any of their values, while the other features have values that are guaranteed to satisfy the remaining conditions. Whether or not there is a duty to repair is therefore solely determined by whether or not condition c_2 is satisfied. All combinations of values of the other features are considered. The Unlawfulness dataset therefore consists of 168 unique instances, of which 66.66% have a positive *dut* value.

The *Imputability dataset* is a similar subset of the unique dataset, but now the features for the imputability condition (c_3) can take on any value, except that the value of *vst* must be such that condition c_5 is satisfied. The value of $dut(x)$ is now completely dependent on whether or not condition c_3 evaluates to true. Due to the interdependency of conditions c_3 and c_5, the Imputability dataset only has 128 unique instances, 87.5% of which have a positive *dut* value.

4 EXPERIMENTAL SETUP AND RESULTS

In this section we describe and motivate the experiments we performed for the tort law domain and report on their results.

4.1 Experiments

We decide to use neural networks like in [Bench-Capon 1993]. The method is model-agnostic, however, meaning that it can be applied to any other machine learning model as well. We assume that assessing and improving rationale discovery is relevant only for models that perform well on their respective task. Our first step, after training the above mentioned neural networks, is therefore to evaluate their performance on typical test sets in terms of the standard accuracy measure. Subsequently we will evaluate the performance of the networks on the dedicated, knowledge-driven test sets that were specifically designed for assessing the networks' quality of rationale discovery.

4.1.1 Neural network architectures. Similar to the original experiments, three multilayer perceptrons were used with one, two and three hidden layers, respectively [Bench-Capon 1993]. The nets all have 10 input nodes, corresponding to the number of features and a single output node, representing duty to repair. The node configuration (i.e. number of nodes per layer) of each network is as follows:

- One hidden layer network: 10-12-1
- Two hidden layer network: 10-24-6-1
- Three hidden layer network: 10-24-10-3-1

We use the MLPClassifier of the scikit-learn package [Pedregosa et al. 2011], the sigmoid function as the activation function, the Adam stochastic gradient-based optimizer [Kingma and Ba 2015], with a constant learning rate of 0.001. A total of 50,000 training iterations are used with a batch size of 50. Recall that the focus of this study is not on creating the best possible classifier, but to assess rationale discovery.

4.1.2 Training and performance testing. The three types of neural networks are trained and tested on all combinations of different datasets from Table 1. Every combination of training dataset and testing dataset is evaluated in terms of the accuracy of the resulting network on the test data. Because some of the datasets are stochastic (each generated dataset is slightly different), the whole process of data generation, training and testing is repeated 50 times. The mean classification accuracies along with their standard deviations are reported. To assess the rationale discovery capabilities of all the trained networks, we study their performance on the dedicated test sets for unlawfulness and imputability conditions. Performance is measured both quantitatively, using standard accuracy, and qualitatively by a more detailed comparison of actual and expected outcomes.

4.2 Results

Table 2 shows the mean classification accuracies over 50 runs, together with their standard deviations, for the different combinations of training and testing sets in the tort law domain. The table includes the quantitatively measured performance on the two dedicated test sets.

We can evaluate how well conditions c_2 (unlawfullness) and c_3 (imputability) are learned. For these conditions, the network should output 1 in cases from the Unlawfulness dataset where the case is unlawful (c_2), or in the Imputability dataset where the case can be imputated to a person (c_3); otherwise the output should be 0. The mean output of the 3 layer network over 50 runs for the two training sets on the Unlawfulness and Imputability datasets is presented in Table 3.

Table 2: The accuracies obtained by the neural networks in the tort law domain.

	Trained on all instances				Trained on smaller dataset			
	General	Unique	Unlawfulness	Imputability	General	Unique	Unlawfulness	Imputability
1 hidden layer	100±0	100±0	100±0	100±0	98.45±0.5	97.24±0.89	92.8±3.47	91.22±4.04
2 hidden layers	100±0	100±0	100±0	100±0	99.03±0.44	98.27±0.78	95.71±3.1	94.38±3.84
3 hidden layers	99.86±0.37	99.76±0.66	99.67±1.83	99.5±1.56	98.23±0.72	96.83±1.28	92.96±5.33	91.45±3.51

Table 3: Mean network output on the Unlawfulness and Imputability datasets versus the logical evaluation of the unlawfulness resp. imputability conditions.

Trained on all instances		Trained on smaller dataset	
Unlawfulness	**Output**	**Unlawfulness**	**Output**
False	0	False	0.018
True	1	True	1
Imputability	**Output**	**Imputability**	**Output**
False	0	False	0.875
True	1	True	1

5 DISCUSSION

5.1 Standard Accuracy

Standard accuracy is measured to see whether the learned models are able to solve the classification problem, regardless of whether or not they discovered the rationale underlying the data. We find accuracies of 100% or near 100% for networks trained on all instances (see Table 2). When presented with all unique instances, the networks with one and two hidden layers are able to perfectly predict the outcome from Dutch tort law, and the network with three hidden layers can create a very close approximation.

Presenting a neural network with all available cases is in practice often infeasible. If it is possible, then a simple lookup table rather than a neural network would most likely suffice. For this reason, we also trained the networks on a subset of only around 35% of the unique instances (see Table 2). As expected, the accuracies of the networks on the general test sets drop, but only slightly (to 98-99%). Even on the unique test set, accuracies remain around 96%. This suggests that it is possible for the models to approximate tort law with a small subset of the unique cases.

5.2 Rationale Discovery

Looking at the performance of the networks on the dedicated test sets partially exposes how well the rationale is captured by the network. We designed these test sets such that each one targets a single condition from the domain. In addition to considering the accuracy on these dedicated test sets, we qualitatively evaluate the rational discovery capabilities of the networks by comparing their outputs with the actual outputs we would ideally expect for the different domains.

Recall that on the Imputability dataset, networks should output 1 if the act is imputable to the person, and 0 otherwise; on the Unlawfulness dataset, the networks should output 1 if the case is unlawful, and 0 otherwise. Table 3 shows how well the networks were able to internalize the notions of unlawfulness and imputability. When trained on all instances, the mean output of the networks is 0 if the logical evaluation of unlawfulness is false, and 1 if it is true, which is exactly what it should do. Networks trained on all instances attain a perfect score on the Imputability dataset as well. This can also be seen in Table 2, where the networks score 100% accuracy on the Unlawfulness and Imputability datasets after training on all instances.

With less data, however, accuracies drop to around 92-95% for the Unlawfulness dataset and 91-94% for the Imputability dataset. This accuracy may still seem high, but we should take into account the label distributions (66.67-33.33% and 87.5-12.5%, respectively). Table 3 shows that networks still perform perfectly on cases in which the unlawfulness and imputability conditions evaluate to true. When the conditions are false, however, mistakes are made. The average output of networks on the Unlawfulness dataset increases to 0.018, which should be 0, meaning that networks classify some lawful cases as unlawful. In the Imputability dataset, the mean output increased more drastically to 0.875 when imputability is false, meaning that in 87.5% of the instances in which the act is not imputable to a person, the network incorrectly decided that it should be. This means that despite high accuracy on the general test set, the networks largely ignored the concept of imputability.

5.3 A Method for Rationale Evaluation

Although our experiments and discussion focused on specific example domains and neural networks, our approach for rationale evaluation can be interpreted as a general method independent of the machine learning algorithm applied. Building on the results of this paper, we therefore proposes a knowledge-driven method for model-agnostic rationale evaluation, consisting 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.

The first step is based on the assumption that efforts for assessing and possibly improving the rationale discovery capabilities of a learned model are only taken if the general performance of the model is already considered good enough. Here we assume performance is measured using accuracy, but other measures can be employed as well and the threshold of what is considered good enough may vary per domain and application.

The second step in our method depends on domain knowledge. Hence the method effectively is a quantitative human-in-the-loop solution for rationale evaluation.

In the third step, performance is again evaluated, by now not only considering accuracy but also examining model output and expected output in terms of the dedicated test sets.

The method does not currently specify how the dedicated test sets are constructed. We aim to further operationalize the rationale evaluation method by using information about the knowledge in the domain, and the distribution of examples, for instance building on Bayesian networks. Subsequently, the information gained by using this rationale evaluation method can be used to improve the rationale of the system by adjusting the training data accordingly, such as in [Bench-Capon 1993] and our replication variants [Steging et al. 2021], effectively allowing us to impose sound rationale discovery.

6 CONCLUSION

The work in this paper was inspired by Bench-Capon's 1993 paper that investigated whether neural networks are able to tackle open texture problems. The conclusions were that trained networks can perform very well in terms of accuracy, even though some conditions from the domain are not learned [Bench-Capon 1993]. Similar results were found when we repeated the experiments with larger training datasets, in order to ensure that the original conclusions about conditions that were not learned are not due to a lack of data.

The idea of constructing test cases to test specific conditions inspired us to propose a method for assessing rationale discovery capabilities by designing dedicated test datasets and to evaluate performance on these knowledge-driven test sets, combining quantitative and qualitative evaluation elements in a hybrid way. Adjusting the training dataset based on this evaluation methods demonstrates that the rationale can be improved using knowledge-driven tailor made training sets [Bench-Capon 1993; Steging et al. 2021].

In the real life tort law domain, with a non-fictional knowledge structure and different characteristics, a similar pattern can be observed as before: the networks failed to learn the independent condition that defines imputability, despite high accuracies on the general test set.

This study therefore reaffirms the conclusions from previous work, while simultaneously introducing a model-agnostic method for assessing rationale discovery capabilities of machine-learned black box models, using dedicated test datasets designed with expert knowledge of the domain. In future research, we aim to further detail and extend our method such that by employing it, the soundness of the rationale underlying system decisions becomes tangible, and its quality can be asserted. Based on this evaluation, the training data of the black-box systems can be altered to improve their rationale. Further expanding upon this design method will bring us closer to AI that is both explainable and responsible.

ACKNOWLEDGMENTS

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

K. D. Ashley. 1990. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals.* The MIT Press, Cambridge (Massachusetts).

K. D. Ashley. 2019. A brief history of the changing roles of case prediction in AI and law. *Law in Context* 36, 1 (2019), 93–112.

K. Atkinson, T. Bench-Capon, F. Bex, T. F. Gordon, H. Prakken, G. Sartor, and B. Verheij. 2020b. In memoriam Douglas N. Walton: the influence of Doug Walton on AI and law. *Artificial Intelligence and Law* (2020), 1–46.

K. Atkinson, T. Bench-Capon, and D. Bollegala. 2020a. Explanation in AI and law: Past, present and future. *Artificial Intelligence* 289 (2020), 103387.

T. Bench-Capon. 1993. Neural networks and open texture. In *Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93)*. ACM, New York, 292–297.

S. Brüninghaus and K. D. Ashley. 2003. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL 2003)*. ACM, New York (New York), 233–242.

I. J. Goodfellow, J. Shlens, and C. Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of International Conference on Learning Representations*.

T. F. Gordon. 1995. *The Pleadings Game: An Artificial Intelligence Model of Procedural Justice.* Kluwer, Dordrecht.

M. Grabmair, K. D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg, and V. R. Walker. 2015. Introducing LUIMA: an experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, New York (New York), 69–78.

J. C. Hage, R. Leenes, and A. R. Lodder. 1993. Hard cases: a procedural approach. *Artificial intelligence and law* 2, 2 (1993), 113–167.

D. Hunter. 1999. Out of their minds: Legal theory in neural networks. *Artificial Intelligence and Law* 7, 2 (1999), 129–151.

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations*.

S. M. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4765–4774.

M. Medvedeva, M. Vols, and M. Wieling. 2019. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law* (2019), 1–30.

R. Mochales Palau and M. F. Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009)*. ACM Press, New York (New York), 98–107.

M. Možina, J. Žabkar, T. Bench-Capon, and I. Bratko. 2005. Argument based machine learning applied to law. *Artificial Intelligence and Law* 13, 1 (2005), 53–73.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

L. Philipps and G. Sartor. 1999. Introduction: from legal theories to neural networks and fuzzy reasoning. *Artificial Intelligence and law* 7, 2 (1999), 115–128.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "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*. 1135–1144.

E. L. Rissland and K. D. Ashley. 1987. A case-based system for trade secrets law. In *Proceedings of the 1st International Conference on Artificial Intelligence and Law (ICAIL '87)*. ACM, New York, NY, USA, 60–66.

C. Steging, S. Renooij, and B. Verheij. 2021. Discovering the Rationale of Decisions: Experiments on Aligning Learning and Reasoning. arXiv:2105.06758 [cs.AI]

A. Stranieri, J. Zeleznikow, M. Gawler, and B. Lewis. 1999. A hybrid rule–neural approach for the automation of legal reasoning in the discretionary domain of family law in Australia. *Artificial Intelligence and Law* 7, 2-3 (1999), 153–183.

B. Verheij. 2003a. Artificial Argument Assistants for Defeasible Argumentation. *Artificial Intelligence* 150, 1–2 (2003), 291–324.

B. Verheij. 2003b. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial intelligence and Law* 11, 2-3 (2003), 167–195.

B. Verheij. 2017. Formalizing arguments, rules and cases. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL '17)*. ACM, New York, 199–208.

C. S. Vlek, H. Prakken, S. Renooij, and B. Verheij. 2016. A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law* 24, 3 (2016), 285–324.

M. Wardeh, T. Bench-Capon, and F. Coenen. 2009. Padua: a protocol for argumentation dialogue using association rules. *Artificial Intelligence and Law* 17, 3 (2009), 183–215.

A. Wyner, R. Mochales-Palau, M. F. Moens, and D. Milward. 2010. Approaches to text mining arguments from legal cases. In *Semantic Processing of Legal Texts*. Springer, Berlin, 60–79.