An Application of Case-Based reasoning to Decision-Making in Dutch Administrative Law

Joep NOUWENS [a] and AnneMarie BORG [a] and Henry PRAKKEN [a,1]
[a] *Department of Information and Computing Sciences, Utrecht University, The Netherlands*

Abstract. This paper reports on an experiment on using case-based reasoning in Dutch administrative law. The use case is decision-support for human medical experts at the Dutch Central Office of Driving Certification who have to decide whether a citizen who applies for a driving licence is fit to drive. Case-based reasoning is investigated for this purposes because of its potential advantages over machine-learning approaches as regards transparency and explainability. Both traditional case-based reasoning, AI & Law models of precedential constraint and their combination are investigated on predictive accuracy relative to a large case base with more than 30.000 cases. A combined model is found to have the highest accuracy. The results indicate that human-in-the-loop support with a tool based on the combined model may be feasible, but whether this is indeed so requires further investigation.

Keywords. case-based reasoning, precedential constraint, administrative-law decision-making

1. Introduction

Case-based reasoning (CBR) has been recommended as a way to develop transparent and explainable data-driven AI applications, as an alternative to machine-learning approaches, which often lack transparency and explainability [11,14]. Models of case-based reasoning have been developed both in AI & law and in other areas of AI. This paper reports on an experiment on using CBR in Dutch administrative law. The Dutch Central Office of Driving Certification (in Dutch the 'Centraal Bureau Rijvaardigheidsbewijzen', below 'Bureau' for short) is interested in AI support for its decision-making about whether drivers are fit to drive. Because of government policy, the CBR is reluctant to use machine-learning methods, because of their problems with transparency and explainability. For this reason the Bureau wanted to investigate the suitability of CBR methods. Such methods explain or justify a decision in a new case by pointing at similar past cases with the same outcome.

[1]Corresponding Author: Henry Prakken, h.prakken@uu.nl.

The question arises which kind of CBR model is suitable. The Bureau initially considered general CBR models, which essentially model cases as features with a decision and then develop similarity measures between cases as a means to suggest new decisions [5]. A simple similarity measure for domains with only binary features is the proportion of features on which two cases agree. However, traditional CBR methods have limited explainability capabilities in that they cannot express whether a feature is pro or con a particular decision. Formal AI & Law models of case-based reasoning [6,7,13] can explain this, since they capture to which extent a feature value is pro or con a particular decision. This allows for explaining decisions by so-called *a fortiori reasoning*: a decision can be explained by pointing at a precedent with the same decision and where the current case is on all features at least as strong for the same outcome as the precedent, so that the case base 'forces' the decision. In this way, a difference in feature values that in general CBR would be a relevant difference between cases can in a model of precedential constraint be an even stronger reason for the same decision.

However, a fortiori reasoning also has a limitation, namely that that for unforced decisions it does not provide an unambiguous outcome but instead regards both outcomes as allowed. Accordingly, the aim of this paper is to study in the context of the Bureau application scenario how the two approaches can be combined in a way that combines their advantages while avoiding their shortcomings. Four different models will be tested, one general CBR model, two variants of a formal AI & law model of precedential constraint and a combination of the general CBR and the AI & law model. The models will be evaluated in terms of how well they agree with the decisions of the human decision makers (accuracy, precision, recall).

This evaluation approach requires justification, since it is not immediately obvious why a tool that is evaluated on how well it *reproduces* human decisions can be useful for *supporting* human decision-makers. According to Ashley [2, p. 102], Aleven [1] "argued persuasively that predictive accuracy was one measure of the reasonableness of a computational model of argument" in that "Good predictive performance would inspire confidence that the arguments made by the program (...) have some relation to the reality of legal reasoning" [1, p. 212]. In Section 5 we will discuss to which extent this can be plausibly argued.

This paper is organised as follows. We will first in Section 2 outline the Bureau domain and introduce the two formalisms used in our experiments. Then in Section 3 we will present our experimental setup after which in Section 4 we present the results. In Section 5 we will discuss the results, after which we conclude in Section 6. In this paper we only describe and summarise the essential elements of the experiment. The full details can be found in [10].

2. The Bureau domain and its formal modelling

When citizens apply for a driving license, a medical expert of the Bureau has to determine whether they are 'fit to drive' in that they are healthy enough to safely participate in traffic based on their mental and physical health. The health information provided by the citizen is categorised in terms of a considerable number

of relevant dimensions[2], more precisely, 'nature-severity' pairs. Some dimensions are boolean, others have numerical values, some have date values and some have a free-text formal value. Boolean dimensions stand for the presence or absence of a defect, An example of a multi-valued dimension is VISUS, a nature corresponding to eye defects, which has several severities, such as whether only the left- or right eye was measured or both. For each VISUS-severity pair a numerical value between 0.0 (for complete blindness) and 3.0 (for exceptional eyesight) can be entered. Dimensions with a date value stand for the beginning of a certain disease, such as diabetes, or the latest occurrence of a certain medical deviation, such as a hypo in case of having diabetes. Finally, dimensions with free-text values are used by the medical experts to record observations about the citizen's health that cannot be expressed in one of the other formats.

Often a decision can be made automatically with the help of clear rules but other cases have to be decided by a medical expert. It is these cases for which the Bureau wanted to investigate CBR support.

2.1. Precedential constraint

As the formal model of precedential constraint we use the dimension-based result model of [7], using notation of [13]. A *dimension* is a tuple $d = (V, \leq_s, \leq_{s'})$ where V is a set (of values) and $\leq_s$ and $\leq_{s'}$ two partial orders on V such that $v \leq_s v'$ iff $v' \leq_{s'} v$. A *value assignment* is a pair (d, v). The functional notation $v(d) = x$ denotes the value x of dimension d. Then a (dimension-based) *case* is a pair $c = (F, outcome(c))$ such that D is a set of dimensions, F is a set of value assignments to all dimensions in D and $outcome(c) \in \{s, s'\}$. Then a (dimension-based) *case base* is a set of cases assumed to be relative to a set D of dimensions in that all cases assign values to a dimension d iff $d \in D$. Likewise, a (dimension-based) *fact situation* is an assignment of values to all dimensions in D. As for notation, $v(d, c)$ denotes the value of dimension d in case or fact situation c. Finally, $v \geq_s v'$ is the same as $v' \leq_s v$. Finally, a case base is *inconsistent* iff it contains two cases $c = (F, s)$ and $c' = (F', s')$ such that $F \leq_s F'$.

In Horty's result model a decision in a fact situation is forced iff there exists a precedent c for that decision such that on each dimension the fact situation is at least as favourable for that decision as the precedent. Horty formalises this idea with the help of the following preference relation between sets of value assignments.

Definition 2.1. [Preference relation on dimensional fact situations.] Let F and F' be two fact situations with the same set of dimensions. Then $F \leq_s F'$ iff for all $(d, v) \in F$ and all $(d, v') \in F'$ it holds that $v \leq_s v'$.

Then precedential constraint is defined as follows.

Definition 2.2. [Precedential constraint with dimensions: result model.] Let CS be a case base and and F a fact situation given a set D of dimensions. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (F', s)$ in CB such

[2]Since features with a direction towards a decision are in AI & law called dimensions, we will from now on use the term 'dimension' instead of 'feature'.

that $F' \leq_s F$. Moreover, deciding F for s is *allowed* iff deciding F for s' is not forced.

As an example (taken from [13]), consider the issue whether the fiscal domicile of a person who moved abroad for some time has changed. Assume there are two dimensions d_1, the duration of the stay abroad in months and d_2 the percentage of the tax-payer's income that was earned abroad during the stay. For both values, increasingly higher values increasingly favour the outcome *change* and decreasingly favour the outcome *no change*. So, for instance, $(d_1, 12m) <_{change} (d_1, 24m)$ and so $(d_1, 24m) <_{no\ change} (d_1, 12m)$. An example of a fact situation is $F = \{v(d_1) = 30m, v(d_2) = 70\%\}$ and an example of a case is $c = (F', change)$ where $F' = \{v(d_1) = 12m, v(d_2) = 60\%\}$. Deciding F for *change* is then forced because of c since $(d_1, 12m) <_{change} (d_1, 30m)$ and $(d_2, 60\%) <_{change} (d_2, 70\%)$. If instead F has $v(d_2) = 50\%$ then $(d_2, 60\%) \not<_{change} (d_2, 50\%)$, so deciding F for *change* is not forced because of c. If, moreover, c is the only precedent case, then deciding F for *change* is not forced but only allowed, as is deciding F for *no change*. Note, finally, that if the case base contains a second case $c' = (F'', no\ change)$ where $F'' = \{v(d_1) = 18m, v(d_2) = 60\%\}$, then the case base is inconsistent since c' is better for *change* than c' while yet its decision is *no change*.

2.2. Similarity measure

For defining a similarity measure the orderings $\leq_o, \leq_{o'}$ of dimension values are ignored. The measure we propose is an adaptation of known measures from general CBR [4,9]. Our adapted definition has to take into account three characteristics of the Bureau data set, namely, that dimension values can be empty (not-specified), can be of type *date* and can have free text format.

An empty value represents the absence of a health deviation, which is important information for a case. This makes it important for the formula to take into account. Additionally, the presence of a filled textual value has to be taken into account even when textual fields do not match entirely. For example, suppose two cases have a present *chronic heart failure - other* dimension, denoting a textual explanation of the presence of chronic heart failures of 'other' types. Even though the values of these cases differ, this still makes the cases more similar than two cases where only one has this chronic heart failure dimension. Therefore, textual fields are taken into account in the similarity measure, and cases containing such textual fields are not removed from the KB for the similarity measure.

Our similarity measure defines a dissimilarity value $d(C_i, C_j)$ for dimension d between two cases C_i and C_j) as a numerical value between between 0 and 1. Formula 1 denotes the final dissimilarity value of values C_i and C_j, which accounts for different types of values and empty cells.

$$d(C_{ik}, C_{jk}) = \begin{cases} 0, & \text{if } C_{ik} = C_{jk} \text{ (or both are empty);} \\ 1, & \text{if } C_{ik} \text{ is empty or } C_{jk} \text{ is empty and not both;} \\ 0.75, & \text{if } C_{ik} \text{ and } C_{jk} \text{ are text values (and not equal);} \\ \frac{|C_{ik}-C_{jk}|}{|C_{ik}+C_{jk}|}, & \text{if } C_{ik} \text{ and } C_{jk} \text{ are numbers and } |C_{ik} - C_{jk}| \leq |C_{ik} + C_{jk}|; \\ date(C_{ik}, C_{jk}), & \text{if } C_{ik} \text{ and } C_{jk} \text{ are date values and } date(C_{ik}, C_{jk}) \leq 1; \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

where

$$date(C_{ik}, C_{jk}) = datedifference\_in\_days(C_{ik}, C_{jk})/10000 \tag{2}$$

First, if two values are equal or if they are both empty, the dissimilarity will obviously be equal to 0, as both dimensions are equal.

If only one of the dimensions is empty, the dissimilarity will be equal to 1, since the nature-severity combination is only present in one of the two cases. If both values are text values (and not equal), the dissimilarity value will be 0.75, as the presence of dimension values shows more similarity between the two cases than the absence of one of the two values.

If C_i and C_j are numerical values, the dissimilarity between the dimensions is calculated based on the formula of 3, taken from [9], but only if both values are negative or both values are positive (equivalent to the condition $|C_i - C_j| \leq |C_i + C_j|$). Otherwise, the value of $\frac{|C_i-C_j|}{|C_i+C_j|}$ would be higher than 0.

$$Canberra: \quad DISS(C_i, C_j) = \sum \frac{|C_{ik} - C_{jk}|}{|C_{ik} + C_{jk}|} \tag{3}$$

If C_{ik} and C_{jk} are date values, the date difference in days is calculated according to Formula 2. The output of this value is divided by 10.000, and the minimum of this output value and 1 is returned. As most date differences are smaller than 10.000 days, this number was chosen as maximum difference between two dates. A value higher than this number hardly ever occurs, and thus the similarity between dates closer to each other must have a bigger influence on the similarity.

Finally, if none of the conditions above are met, the output for the dimension will be 1. Entering the values for all dimensions $d \in D$ for both cases, the total dissimilarity value can be calculated, a value between 0 and 1.

3. Experimental setup

We first removed cases with obvious errors, such as type errors in the dimension values. This left us with 30.584 cases. We then had to deal with the fact that the medical specialists often leave dimension values empty. Since this reflects the absence of a health problem for the considered nature-severity pair, we completed these dimensions with their most favourable value for the outcome 'fit to drive' (although for applying the above similarity measure the dimension values are regards as 'empty').

We then used the model and algorithms of Van Woerkom et al. [16] to compute the preference relations on dimensions by calculating Pearson's correlation for every dimension value. In doing so, we had to ignore all cases with features with textual values, since these cannot be naturally ordered. This left us with 15.843 cases, modelled in terms of a set of 123 dimensions, for which a preference relation could be determined for all their dimensions. The orders of every preference relation were then checked manually by an expert of the Bureau and adjusted if the expert regarded the preference ordering as incorrect. The remaining 14.741 cases had a textual dimension, for which the preference relation was empty.

We used four different models for predicting driver fitness. **Model 1 (traditional CBR)** applies the above similarity measure to suggest the decision of the case in the case base that has the highest similarity with the considered fact situation. If multiple cases have the highest similarity, then arbitrarily one of them will be returned. **Model 2 (negative AF)** applies precedential constraint to a fact situation as follows: if a single decision is forced, then predict that decision, if both decisions are allowed or both decisions forced or a fortiori reasoning cannot be applied because of dimensions with textual values, then predict 'not fit to drive'. **Model 3 (positive AF)** is the same as model 2 except that in the latter case it predicts 'fit to drive'. Model 2 and 3 can be seem as, respectively, favouring safety and giving drivers the benefit of the doubt. Finally, **Model 4 (combined CBR)** combines traditional CBR and a fortiori reasoning as follows: if one decision is forced, then predict that decision, otherwise predict the decision with the similarity measure of traditional CBR.

We then implemented and tested all four CBR approaches independently, using [16]'s implemented algorithms for precedential constraint extended with our own algorithm for CBR similarity. As the Bureau case base contained 30.584 cases, the test set contained 7646 cases, which thus applied the traditional 80/20 split for training and testing models in machine-learning (in our case 80% for the creation of the KB, 20% for testing). Each approach was applied to the same pair of sets in order to obtain comparable results. As the performance measures we adopted precision, recall and accuracy.

4. Results

This section presents the comparative analysis of four CBR models, designated here as the Positive AF CBR model, the Negative AF CBR model, the traditional CBR model, and the combined CBR model, across three key performance metrics: precision, recall, and accuracy.

The test set consisted of 7646 randomly chosen cases from the entire Bureau dataset. From these 7646 cases, 16 cases contained nature-severity combinations that required manual decision making or came with a severity that represented a shorter validity period of the license, which returned the automatic decision of 0. This remains us with 7630 test cases in total. 3730 of the 7630 cases had an actual decision of '1' (fit to drive) and 3900 had decision '0' (unfit to drive). Tables 1, 2, 3 and 4 show the results of every model, where 'Pred' represents the

model's decision and 'Actual' the actual decision of a test case.

Pred \ Actual	1	0	
1	3598	2186	**5784**
0	132	1714	**1846**
	3730	**3900**	**7630**

Table 1. Results of positive AFCBR Model

Pred \ Actual	1	0	
1	1028	52	**1080**
0	2702	3848	**6550**
	3730	**3900**	**7630**

Table 2. Results of negative AFCBR Model

Pred \ Actual	1	0	
1	3491	387	**3878**
0	239	3513	**3752**
	3730	**3900**	**7630**

Table 3. Results of traditional CBR model

Pred \ Actual	1	0	
1	3421	360	**3781**
0	309	3540	**3849**
	3730	**3900**	**7630**

Table 4. Results of combined CBR model

Model	Precision	Recall	Accuracy
Positive AF CBR	0.622	0.965	0.696
Negative AF CBR	0.952	0.276	0.639
Traditional CBR	0.900	0.936	0.918
Combined CBR	0.905	0.917	0.912

Table 5. Precision, Recall and Accuracy scores per CBR model

The precision metric calculated the percentage of accurately recognised positive cases among all cases classified as positive by the models. With a precision of 95.2%, the negative AF CBR model performed best in this comparison, demonstrating its superior capacity to find appropriate scenarios with few false positives. With a precision of 90.5% and 90.0% respectively, the combined CBR and the traditional CBR model trailed closely behind, while the Positive AF CBR model showed lower precision at 62.2%.

Recall assesses the model's ability to identify all actual positive cases within the dataset. The Positive AF CBR model outperformed the others in this metric

with a recall rate of 96.5%. The traditional CBR model had a recall of 93.6%, slightly exceeding the combined CBR model, which recorded a recall rate of 91.7%. The Negative AF CBR model scored an extremely low recall score compared to the other three models, only scoring 27.6% in this metric.

Finally, the accuracy represents the proportion of all correct decisions (both positive and negative) made by the models over the total number of cases. The traditional CBR achieved the highest overall accuracy at 91.8%, followed closely by the combined CBR model with a 91.2% accuracy score. The Positive and Negative AF CBR models showed somewhat lower accuracies at 69.6% and 63.9% respectively. Since accuracy is seen as the most important and widely used performance metric in testing AI-models, the traditional CBR has shown to be the optimal model for this situation compared to the other three models, followed closely by the combined CBR model. Table 5 summarises the performance scores per model.

In total, 2.926 of the 7.630 test cases were decided by a forced decision of the a fortiori algorithm in the models that used AFR. The remaining 4704 cases were determined by using the similarity measure in the combined CBR model.

5. Discussion

Looking at the results, it can be observed that the traditional CBR model and the combined CBR model outperformed the Positive and Negative AF CBR models. This can be easily explained by the fact that the Positive and Negative AF CBR models always make an automatic decision when a fortiori reasoning does not generate a uniquely forced decision. Since the unforced cases were quite spread out over both decisions, this means that both the Positive and the Negative AF CBR model performed poorly. A more unexpected outcome is the fact that the traditional CBR model outperformed the combined CBR model, even though the combined CBR model should have only found stronger evidence for certain decisions by using a fortiori reasoning.

A closer analysis of the results reveals that a main source of incorrect predictions of the models using a fortiori reasoning is the inconsistency of the case base. Van Woerkom et al. [16], propose a measure to calculate the degree of consistency of a case base as "the relative frequency of cases in the case base that have their outcome forced for the outcome they did not receive". Results of this measure show that the subset of the Bureau case base to which a fortiori reasoning can be applied (recall that this subset consists of 15.843 cases) only comes with a consistency percentage of 54.5%, which means that only 8.642 of the 15.843 cases were consistent. In order to increase the accuracy of the combined CBR model, the Bureau will have to determine which cases must be removed from the current dataset to prevent incorrect forced decisions.

Another possible explanation of the poor performance of models 2 and 3 (the 'default' a fortiori models) is that 2657 of the 4704 cases with no unique forced outcome were cases to which a fortiori reasoning could not be applied because of the presence of textual dimensions. Admittedly, our current handling of textual dimensions is rather coarse, which is a limitation of our study. Having said so,

the combined model (model 4) achieved 86,3% accuracy for these cases, which is still quite high.

Finally, in some cases it could be determined by Bureau experts that an incorrect prediction by a fortiori was caused by an incorrect preference ordering on dimension values.

As said in the introduction, Ashley [2], citing [1], argued that predictive accuracy was one measure of the reasonableness of a computational model of argument. We believe that this only holds under the following assumptions:

1. The system contains the same knowledge as the human decision-maker, which is all and only the relevant knowledge.
2. The system reasons with the knowledge in the same way as the human decision-maker, which is rationally sound reasoning.
3. Different human decision-makers decide in the same way.

Under these assumptions a tool with high predictive accuracy can be useful for decision support since it contains the same knowledge as used by the human decision-makers and reasons with it in the same way as the human decision-makers and thus produces acceptable decisions, while in doing so it is not subject to the cognitive limitations of the human decision-maker, such as flawed memory or making reasoning errors. Note that this is the same motivation as the motivation given for the introduction of rule-based expert systems in Dutch public administration [15]. The aim was not to build systems with superior knowledge or reasoning models but instead to mitigate the problems observed in practice that human 'street-level bureaucrats' often overlooked relevant regulations and often made logical reasoning errors or calculation errors.

Non-perfect accuracy can then be explained as follows that these assumptions do not hold categorically:

1. Humans may use other or more knowledge. For instance, they may assume different sets of dimensions, different similarity relations or different preference relations on dimensions). This can also be since the knowledge engineering was flawed.
2. Humans remember or reason imperfectly. For instance, they may ignore relevant past cases or apply the a fortiori reasoning model incorrectly.
3. Different human decision-makers decide differently.

As indicated above, the last point (inter-expert inconsistency) was confirmed by our results, while we also found examples in which a preference ordering on dimension values was with hindsight judged to be incorrect by the experts.

As regards possible usefulness in practice, the results indicate that fully automatic decision-making is infeasible but that a human-in-the loop application might be feasible. Possible advantages of such an application are: no overlooking of relevant cases, insight in the consistency of decision-making by different experts, adherence to the logic of a fortiori reasoning and time-efficiency. In particular, the human decision-maker can investigate whether discrepancies are due to flawed system modelling, the similarity function or to human errors or inconsistencies between human decision-makers. In informal discussions with the Bureau

employees we found that they in particular valued that the tool makes them aware of possible inconsistencies and possible changes in human decision behaviour over time. We note that the first benefit, awareness of inconsistent decision-making, exists because of our modelling of the problem in terms of dimensions instead of just unordered features, as in traditional CBR. Whether a system based on the present ideas will have these benefits requires further experimentation, since it is well known that good performance on a reasoning task in artificial settings does not imply usefulness in practice [8].

6. Conclusion

In this paper we investigate whether 'traditional' case-based reasoning models and formal AI & Law models of precedential constraint can be applied in a realistic setting of administrative-law decision-making. Our main positive finding was the high accuracy of a model that combines the two approaches (model 4). To the best of our knowledge, such a combination of 'traditional' and formal AI & law models of case-based reasoning is novel. We believe that even if the combined model is used, then in ambiguous cases (that is, cases in which the traditional model decides) the preference orderings on the dimensions can still be useful for explaining the prediction. In our case study we informally found that experts at the Bureauin particular valued that the tool makes them aware of possible inconsistencies and possible changes in human decision behaviour over time.

Also novel is the large scale of our experiments, with thousands of precedents, compared to the just 147 precedents in the US trade-secret law case base used by [1] and much follow-up work. The only other work of this scale that we know of is [16], who experiment with a case base with almost 6000 cases. Arguably, the larger the case base, the higher the degree of its inconsistency will be, which raises issues not encountered with case bases of small size. See [3,12] for two recent studies of CBR with inconsistent case bases. The issue of inconsistent case bases is an important topic for future research. For instance, it could be investigated to what extent different case modelling methods influence the consistency degree of a case base. Another topic for future research is a more sophisticated treatment of textual dimensions, possibly using natural-language processing methods.

All in all, our experiments, combined with informal feedback by Bureau experts, suggests that a tool with humans in the loop based on our combined model might be useful in practice, but whether this is indeed so requires evaluation studies of a different kind in realistic decision-making settings.

References

[1] V. Aleven. Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. *Artificial Intelligence*, 150:183–237, 2003.

[2] K.D. Ashley. A brief history of the changing roles of case prediction in AI and law. *Law in Context*, 36:93–112, 2019.

[3] I. Canavotto. Reasoning with inconsistent precedents. *Artificial Intelligence and Law*, 2023. https://doi.org/10.1007/s10506-023-09382-7.

[4] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Loss and gain functions for CBR retrieval. *Information Sciences*, 179(11):1738–1750, 2009.
[5] P. Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, 21:1532–1543, 2005.
[6] J. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33, 2011.
[7] J. Horty. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27:309–345, 2019.
[8] R..M. O' Keefe. Issues in the verification and validation of knowledge-based systems. In V. Ambriola and G. Tortora, editors, *Advances in Software Engineering and Knowledge Engineering*, volume 2 of *Series on Software Engineering and Knowledge Engineering*, pages 173–189. World Scientific Publishing Co, 1993.
[9] G.N. Lance and W.T. Williams. Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9(1):60–64, 1966.
[10] J. Nouwens. Combining a fortiori reasoning and a similarity measure in case-based reasoning. Master's thesis, AI Programme, Utrecht University, Utrecht, 2024.
[11] C. Nugent and P. Cunningham. A case-based explanation system for black-box systems. *Artificial Intelligence Review*, 24:163–178, 2005.
[12] J. Peters, F.J. Bex, and H. Prakken. Justifications derived from inconsistent case bases using authoritativeness. In *Proceedings of the First International Workshop on Argumentation for eXplainable AI (ArgXAI)*, volume 3209 of *CEUR Workshop Proceedings*, 2022.
[13] H. Prakken. A formal analysis of some factor- and precedent-based accounts of precedential constraint. *Artificial Intelligence and Law*, 29:559–585, 2021.
[14] J.M. Schoenborn, R.O. Weber, D.W. Aha, J. Cassens, and K.-D. Althoff. Explainable case-based reasoning: A survey. In *AAAI-21 Workshop Proceedings*, 2021.
[15] J.S. Svensson. The use of legal expert systems in administrative decision making. In A. Grönlund, editor, *Electronic Government: Design, Applications and Management*, pages 151–169. Idea Group Publishing, London etc, 2002.
[16] W. van Woerkom, D. Grossi, H. Prakken, and B. Verheij. A fortiori case-based reasoning: from theory to data. *Journal of Artificial Intelligence Research*, 2024. To appear.

References

[1] V. Aleven. Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment. *Artificial Intelligence*, 150:183–237, 2003.
[2] K.D. Ashley. A brief history of the changing roles of case prediction in AI and law. *Law in Context*, 36:93–112, 2019.
[3] I. Canavotto. Reasoning with inconsistent precedents. *Artificial Intelligence and Law*, 2023. https://doi.org/10.1007/s10506-023-09382-7.
[4] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Loss and gain functions for CBR retrieval. *Information Sciences*, 179(11):1738–1750, 2009.
[5] P. Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, 21:1532–1543, 2005.
[6] J. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33, 2011.
[7] J. Horty. Reasoning with dimensions and magnitudes. *Artificial Intelligence and Law*, 27:309–345, 2019.
[8] R..M. O' Keefe. Issues in the verification and validation of knowledge-based systems. In V. Ambriola and G. Tortora, editors, *Advances in Software Engineering and Knowledge Engineering*, volume 2 of *Series on Software Engineering and Knowledge Engineering*, pages 173–189. World Scientific Publishing Co, 1993.
[9] G.N. Lance and W.T. Williams. Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9(1):60–64, 1966.
[10] J. Nouwens. Combining a fortiori reasoning and a similarity measure in case-based reasoning. Master's thesis, AI Programme, Utrecht University, Utrecht, 2024.

[11] C. Nugent and P. Cunningham. A case-based explanation system for black-box systems. *Artificial Intelligence Review*, 24:163–178, 2005.

[12] J. Peters, F.J. Bex, and H. Prakken. Justifications derived from inconsistent case bases using authoritativeness. In *Proceedings of the First International Workshop on Argumentation for eXplainable AI (ArgXAI)*, volume 3209 of *CEUR Workshop Proceedings*, 2022.

[13] H. Prakken. A formal analysis of some factor- and precedent-based accounts of precedential constraint. *Artificial Intelligence and Law*, 29:559–585, 2021.

[14] J.M. Schoenborn, R.O. Weber, D.W. Aha, J. Cassens, and K.-D. Althoff. Explainable case-based reasoning: A survey. In *AAAI-21 Workshop Proceedings*, 2021.

[15] J.S. Svensson. The use of legal expert systems in administrative decision making. In A. Grönlund, editor, *Electronic Government: Design, Applications and Management*, pages 151–169. Idea Group Publishing, London etc, 2002.

[16] W. van Woerkom, D. Grossi, H. Prakken, and B. Verheij. A fortiori case-based reasoning: from theory to data. *Journal of Artificial Intelligence Research*, 2024. To appear.