Can Predictive Justice Improve the Predictability and Consistency of Judicial Decision-Making?

Floris BEX [a] and Henry PRAKKEN [b]
[a] *Utrecht University & Tilburg University, The Netherlands*
[b] *Utrecht University & University of Groningen, The Netherlands; European University Institute, Italy*

Abstract. There has recently been talk of algorithms that predict decisions in legal cases being used by the judiciary to improve the predictability and consistency of judicial decision making. We argue that their use may minimise the error rate of decisions in the long run, but that this would require not only major technical advances but also major changes in legal thinking about what is the most important objective of judicial decision-making: optimising individual justice in a particular case or reducing errors in the long run. We further argue that if algorithmic decision predictors give any useful information in individual cases to judges at all, this is not in its predictions but in its explanations.

Keywords., Predictive justice, decision prediction, predictability, consistency

1. Introduction

Using machine-learning algorithms to predict decisions in legal cases has become a hot topic [3,13,10,1]. One use of these *algorithmic decision predictors* [5] is to help litigants estimate, for example, their chances of winning a case (e.g., commercial products like Lex Machina and Premonition.ai). Another possible use of predictors is to use them to analyse human biases or the influence of extraneous, non-legal factors on legal decision making [4,14]. Finally, a more contentious use of algorithmic predictors is their use by the judiciary (courts, judges): perhaps not as fully-automated 'robo-judges', but possibly for supporting judges in individual cases – it is this latter use of algorithmic decision predictors that is the main subject of this paper.

The use of algorithmic decision predictors by the judiciary is claimed to improve the predictability and consistency of judicial decision making, which is demanded by the principle of equality [8]: judges can use decision predictors as a support tool when drafting judgements to identify cases and patterns which lead to certain decisions [1], in order to come to more consistent, more informed and less biased judgments [4,14]. Some even argue that AI can be used as a 'monitor' [4] that shows the judge what the rational decision would be in a new case given the history of similar cases.

To be able to evaluate whether algorithmic decision predictors can have these claimed benefits, it is first necessary to have a clear picture of what information such al-

gorithms provide. This we discussed in [5], concluding that even if we have a prediction by an algorithmic predictor that performs well on a test set, we still cannot say that a rationally-thinking judge would probably take the predicted decision. In this paper, we aim to discuss exactly what is meant by the predictability and consistency of judicial decision making, and whether the use of algorithmic decision predictors by judges can improve such predictability and consistency.

2. Preliminaries

We first summarise our answer to the main question of [5]: if we have a prediction by an algorithmic decision predictor and information about the algorithm's performance, can we determine the so-called *decision probability* that an arbitrary rational judge assigned to the case would take the predicted decision?

In [5], we assumed that given an algorithm's performance measures (e.g. *precision*, the percentage of positive predictions that are correct) a statistical conditional probability can be derived that an arbitrary case C will receive decision D given that the algorithm predicts D for C. This probability is statistical in that it is not about an individual case c to be decided but about the class of all cases C for which the algorithm can give a prediction. By contrast, a decision probability is a conditional probability for an individual case c and a particular decision d that case c will receive decision d. Here we have the reference class problem, namely that a probability for an individual case c is not logically implied by a statistical probability for the class of all cases C to which c belongs. Instead, equating the individual probability of some event to the statistical probability for all events of the same class expresses a relevance judgement that the only thing that is relevant as regards the event is what is stated in the statistical probability. For example, if we know that 80% of the people with an Italian first are Italian (a statistical probability) and all we know of a particular person that he is called Giovanni, then we may rationally conclude that the individual probability that this Giovanni is Italian is 80%. However, if we also know that this Giovanni's surname is not Italian but Dutch, then this is clearly additional relevant information, so the statistical probability cannot be applied to him any more. So in the case of our decision predictor, equating the decision probability for a case c to the statistical probability for the class C of all c expresses that all that is relevant as regards c is the predicted outcome of a case. However, this relevance assumption is unjustified, since judges always know more about the case at hand than just its predicted outcome. We therefore concluded that an algorithmic decision predictor cannot be said to give the 'normal' or rational decision of a case given the history of similar cases.

3. Predictability and consistency

In the introduction we noted that some think that if judges take predictions of algorithmic decision predictors into account when deciding a case, this will improve the predictability and consistency of judicial decision-making. Two questions arise here (i) What do the terms predictability and consistency mean in this context?; and (ii) How can an algorithm be used to improve predictability and consistency?

We initially assume that in the context of judicial decision making predictability and consistency mean the same (although this is debatable). One interpretation of predictabil-

ity and consistency is then that the *same* case would be decided the same by different judges if assigned to the case. Another interpretation is that *similar* cases are decided in the same way (or a similar way) by the same or different judges. The second interpretation implies the first but not vice versa. We think that in both interpretations 'consistency' and 'predictability' indeed mean the same. The latter is not true for a third interpretation of predictability, corresponding to the gambler's wish to maximize expected utility in the long run. For instance, many cases might be substantially different from each other, so that even if like cases are treated alike, the predictability of the decision is low. A gambler who wants to bet on legal case decisions might indeed be advised to take an algorithm's prediction into account, since the gambler will often have no more information about the case than the algorithm's prediction plus statistical information about its performance. We next discuss for each of these interpretations how an algorithmic decision predictor can be used in order to improve predictability and consistency.

Deciding the same case in the same way. If predictability and consistency of judicial decision-making means that the *same* case is decided the same by different judges, then a sure way to guarantee this is to give all judges the same algorithmic decision predictor and to require that they all have to follow its predictions in all cases. Then different judges would, when assigned to the same case, be guaranteed to take the same decision. However, this does not make sense, for one since we do not know whether all decisions in the training and test set were correct. If all judges blindly follow the algorithm's prediction, then its accuracy will increase to 100%, and this would further lead to a tendency to make the predicted decision the legally correct one even if this cannot be justified. A possible counterargument here is that judges should not automatically follow the algorithm but just be willing to be informed by it. But then the main problem discussed Section 2 arises again: the judge cannot know from a prediction alone (combined with a statistical probability on the algorithm's performance) whether the predicted decision is what other judges assigned to the same case would likely decide. At the very least the algorithm should be able to explain the grounds for its prediction in legally meaningful terms. We will discuss this issue in more detail in Section 4, noting for now that there is a serious danger that judges who are told that they should consult the algorithm before taking their decision feel an unjustified pressure to accept the predicted decision as the legally correct one. This may in turn make that judges will think less hard about a case and become intellectually 'lazy'. So letting judges be informed by the predictions of algorithmic decision predictors has no clear advantages while there are real dangers. We cannot know whether predictability and consistency of judicial decision making (in the first sense) can be improved by letting judges be informed by the predictions of algorithmic decision predictors without combining them with an explanation in legally meaningful terms.

Deciding similar cases in the same way. How can predictability and consistency be improved if that means that *similar* cases should be decided the same? Is this improved if we require judges to consult decision predictors as a source of information? Again, if all we have is the prediction by an algorithm and some (statistical) probability, we cannot know. And even if we have a decision probability for an individual case, the prediction in itself would still not give any information about similar cases. In fact, it might well be that an algorithm treats cases that judges would regard as similar as different or vice versa. For example, text-based decision predictors, which identify statistical correlations are identified between certain words or combinations in the text and the case decision

(such as the algorithms of [1,13] that predict outcomes of the European Court of Human Rights), could fail to recognise that linguistically small differences are legally very relevant. The converse may also happen, i.e., that the algorithm treats cases as different that judges would treat as similar, since the algorithm recognises differences that are legally irrelevant. Recall in this respect that with such predictors we cannot even know to which extent their predictions are based on aspects related to the merits of the case. At best, knowledge-based predictors that generate their predictions in a case-based way could give such information. We will further discuss this issue below in Section 4.

So also if predictability and consistency of judicial decision making means that like cases are decided alike, we can conclude that we cannot know whether it will be improved by letting judges be informed by the predictions of algorithmic decision predictors without combining them with an explanation in legally meaningful terms. Incidentally, in both interpretations of predictability and consistency there is a further reason for this, since for knowing whether using the algorithm will improve predictability and consistency, we will have to compare the situation with use of the algorithm to the current situation; and there is no a priori reason to assume that judges without algorithmic support will be less predictable and consistent.

3.1. Reducing error rates in the long run

We finally consider the gambler's interpretation of predictability. It might be argued that there is still some rationality in relying on statistical probabilities concerning decision predictions in individual cases. Assume, for instance, that it can be established through empirical research that judges on average make fewer mistakes when they always follow an algorithmic prediction than when they look at all particulars of a case as recommended by us in [5]. (It may be hard to conduct such research but let us for the sake of argument assume that it can be done.) Would it then not be more rational for a judge to reply on the outcome predictions?[1] We believe that the answer depends on which values are to be optimised in judicial decision making: should a judge, when faced with a case, be primarily interested in minimising the rate of erroneous decisions in the long run or should the judge primarily aim to optimise individual justice in the case at hand?

Consider an analogy from consumer banking. A bank that has to decide whether to grant a loan to a customer is not interested in optimising the quality of the decision for an individual customer but in minimising losses in the long run. Given this objective, it is rational for the bank to rely on statistical frequencies concerning classes of customers, even if the individual customer in the case at hand may have additional relevant characteristics not considered in the statistical probability. By contrast, it is in the customer's interest that these additional characteristics are considered by the bank, since the customer wants to be treated fairly. As we earlier observed in [5], this is related to O'Neill's [15] criticism of 'bucketing', the practice of basing a decision about an individual on the fact that the individual is a member of a particular class of which a statistical frequency is known instead of on the particular situation of that individual. O'Neill [15, pp. 145–6] argues that, although this strategy might optimise the decision maker's profit in the long run, it may lead to unjust decisions in individual cases.

Applying the same thinking to our problem, the same question should be answered by designers of procedures of judicial decision-making: is the main objective of judicial

[1] This question was brought to our attention by Giovanni Sartor in personal communication.

decision-making to minimise the rate of erroneous decisions in the long run or to optimise individual justice in the case at hand? Ultimately, this is a matter of legal policy. If the objective is chosen to be optimising individual justice, then algorithmic decision predictors have no relevance for judges deciding individual cases [5]. But if the objective is chosen to be minimisation of errors in the long run, we do not see any principled rational reason not to rely on algorithmic decision predictors, provided it can be established that their use indeed leads to a lower rate of incorrect decisions. However, there are serious practical obstacles. First, creating algorithms that provably reduce error rates is far from trivial and may require major technological advances, which makes the remainder of this discussion largely hypothetical. Second, it seems to us that most legal procedures are mainly meant to optimise individual justice so that benefiting from algorithmic decision predictors in the long run would require major changes in legal-procedural thinking. For instance, an obvious way to ensure error-rate reduction would be to always follow the prediction but then the judge would ignore the particulars of a case, which would very likely violate current procedural rules. If, for these reasons, the prediction is used as just one of the inputs for the judge besides the particulars of the case, then, as noted above, the problem arises how exactly the prediction should be combined with these particulars. For one thing, the advantage of reducing error rates might be lost. Moreover, the danger is that the predicted outcome is incorrectly assumed to be the normal outcome of the case, from which a rational judge could only deviate if the particulars of the case contain exceptional circumstances; as we explained at length in [5], this assumption is unjustified. Finally, relying on the predictions of a non-transparent algorithm would create an explainability problem, especially given current procedural justification requirements on judicial decisions.

4. Providing explanations

We can conclude from Section 3 that a decision prediction on its own, even when combined with quantitative performance information, cannot help judges making their decision-making more predictable and consistent in legally desirable ways. But is this different if the prediction is combined with an explanation for it? The answer is negative if the explanation cannot be given in terms of reasons related to the merit of the case. So it is not a good idea to use algorithms like the one of [10], which make their predictions based on extraneous factors, such as information about the judges, the litigants, the solicitors, the type of case or the jurisdiction. But this implies that a text-based predictor like the ones of [1,13] is also not useful, since it cannot extract any legally relevant information from the texts to which it is applied and use it for explaining a prediction in legal terms. In consequence, there is no way to identify whether its prediction is based on legal grounds or on extraneous factors. So the only kinds of decision predictor that could possibly yield legally relevant information to a judge are those that base their predictions on legally relevant factors.

In Section 2 of [5], we discussed two kinds of decision predictors. One kind is still based on statistics or machine learning but its cases are encoded in terms of legally meaningful features instead of as raw text or with extraneous data (e.g. [12]). The other kind is knowledge-based (e.g. [6,9]). For the first kind of system its performance could be measured for various subsets of factors, and if a case matches a particular subset, then

a probability derived from the system's performance for this subset could be reported. However, this would still only yield a statistical probability for a decision and no decision probability, so, as explained in [5] and summarised in Section 2, the judge would still have to think about the particulars of the case as usual; there is no sense in which the prediction gives the 'normal' decision of cases with this constellation of factors. Alternatively, the system could show similar cases to the judge according to some suitable notion of similarity. However, just showing similar cases is not yet a genuine legal explanation. It remains to be investigated to which extent predictors of this kind can generate legally acceptable and useful explanations in terms of their input factors.

A knowledge-based predictor can by definition yield a genuine legal explanation, since it determines the decision to be predicted by way of applying a model of legal reasoning and problem solving. So (if well designed) such a system can in principle explain its predictions in ways that judges would appreciate and understand. However, there are still some issues here. First, how do we know that the explanation given by the system is a legally acceptable one? Can we assume that a knowledge-based predictor with good test-set performance will also in a high number of cases give a legally acceptable explanation for the prediction? Perhaps, but the assumption is highly defeasible, while again the step from a statistical to a decision probability must be justified, which is far from trivial. For these reasons we believe that additional experiments of a different kind are needed to assess the legal quality of the generated explanations. Since there is no gold standard for this issue, such experiments will have to involve legal experts rating the quality of the explanations, similar to, for instance, the famous experiments in which the quality of the diagnoses and treatment advice given by the MYCIN medical expert system was evaluated [7]. Evaluating systems in this way is far from trivial [11], unlike determining numerical scores like accuracy, precision and recall, which can be automatically extracted from an experiment's confusion matrix.

Incidentally, it might be argued that if a predictor's explanation can generally be shown to be legally acceptable, then this also justifies interpreting the statistical probability based on an algorithm's performance on a test set as a decision probability for a specific case. This argument fails. First, note again that the statistical probability is not based on the specific reasons mentioned in the explanation but on the performance on the test set. Things might be better if statistical probabilities are known for specific classes of test cases, but as explained in [5], obtaining such more specific statistical probabilities is not trivial. Moreover, we would still have to justify all other assumptions needed to make the jump from the past to the future (see the end of Section 2). Instead of attempting to do all this, it is simpler to inspect the given explanation alone and ignore the fact that the decision was predicted; only the content of the explanation can give the judge an indication whether the predicted decision is a good one.

Assuming that the explanations shown by the system are generally legally acceptable, then a second question arises: how do we know that showing such an explanation is useful for judges, for instance, that the quality and consistency of their decision making increases and that bias is reduced? Here the quantitative test-set performance information is completely irrelevant. Instead, this question should arguably be answered in controlled and/or fielded experiments with actual legal decision makers, to check whether the legal quality of their decision improves when they are supported by algorithmic decision predictors. Like with the experiments for testing the legal quality of explanations, setting up such user studies in a correct way is far from trivial [11].

Validation studies of the kinds we have just suggested are, to the best of our knowledge, currently rare. This was different in the early days of AI & law research. For example, in the Netherlands, in the late 1980s and 1990s several user studies were done on the effect of knowledge-based support for civil servants deciding on social benefit applications; see [16, Section 3] for an overview. And Aleven [2] studied the effect of using CATO in teaching legal argumentation skills to law students on these skills. We believe that the current focus on data-driven approaches, with its associated quantitative performance criteria that can automatically be extracted from the experimental data, may be in part responsible for the current neglect of these other important kinds of validation studies. These studies are important if we want to convince the professional legal world that our AI & law systems can contribute to improving the quality of legal decision making.

Such validation studies still say little about the quality of an individual explanation in a new case, since the step from the test results to an individual new explanation is still nontrivial for all the reasons explained in [5]. However, the studies can be used by courts in their decision whether to let their judges be supported by the system. This is the same as courts deciding which law journals or other information sources it will make available for judges. Just as with, for instance, law journals, a general evaluation about its quality has to be made, as a criterion for deciding whether the judge will consult this information source at all. But just as with, for instance, law journals, judges should not automatically copy or accept what is said but only look at the content of what is being said or written.

5. Conclusion

We discussed to what extent algorithmic decision predictors can improve the predictability and consistency of judicial decision-making, given our earlier conclusion in [5] that such algorithms cannot rationally inform individual decisions of judges in a particular case. We discussed three senses of such predictability and consistency: (1) that the same case will be decided the same by different judges; (2) that a similar case will be decided the same by the same or different judges; (3) that always following the prediction will optimise the quality of a series of decisions in the long run. We argued that in the first two senses the use of such algorithms cannot improve the predictability and consistency of judicial decision-making in legally desirable ways. We also argued that this is possibly different in the third sense in that judges might minimise their error rate in the long run. However, this would require not only major technical advances but also major changes in legal thinking about what is the most important objective of judicial decision-making: optimising individual justice in a particular case or reducing errors in the long run.

We also argued that if algorithmic decision predictors give any useful information in individual cases to judges at all, this is not in its predictions but in its explanations. In particular, decision predictors are needed that can explain their predictions in legally relevant terms. However, we noted that whether support by such systems can indeed improve the quality of judicial decision making requires validation studies of a kind that goes far beyond the current trend to focus on numerical performance measures like accuracy, precision and recall. We made a plea for returning to an older AI tradition of carrying out empirical validation studies with potential or actual users of the algorithm.

We like to emphasise that our conclusions are confined to the use of algorithmic decision predictors for informing judges on what they could decide in particular cases.

Other uses of such algorithms may well have benefits, for instance, with respect to informing judges and academics about possible bias in a series of cases (cf. e.g. [4,14]). Moreover, algorithms for making different types of predictions can also be useful. For example, if the aim is to help courts in making their case management more efficient, then algorithms could be trained on features of cases that influence such efficiency, such as their duration. (By contrast, the use of case decision predictors for efficiency purposes, as suggested by Aletras et al. [1], does not make sense, since predictions of decisions do not give any information about efficiency-related aspects of the case.) Note, however, that many of the reservations we expressed in Section 2 and [5] also hold for such other predictive algorithms.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

[2] V. Aleven. *Teaching Case-Based Argumentation Through a Model and Examples*. PhD Dissertation University of Pittsburgh, 1997.

[3] K.D. Ashley. A brief history of the changing roles of case prediction in AI and law. *Law in Context. A Socio-legal Journal*, 36(1):93–112, 2019.

[4] B. Babic, D.L. Chen, T. Evgeniou, and A.-L. Fayard. A better way to onboard AI. *Harvard Business Review*, July-August, 2020.

[5] F.J. Bex and H. Prakken. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pages 175–179, New York, 2021. ACM Press.

[6] S. Brueninghaus and K.D. Ashley. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17:125–165, 2009.

[7] B.G. Buchanan and E.H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[8] European Commission for the Efficiency of Justice (CEPEJ). European ethical charter on the use of artificial intelligence in judicial systems and their environment, 2018.

[9] M. Grabmair. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law*, pages 89–98, New York, 2017. ACM Press.

[10] D.M. Katz, M.J. Bommarito, and J. Blackman. A general approach for predicting the behavior of the Supreme Court of the United States. *PloS one*, 12(4):e0174698, 2017.

[11] 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.

[12] E. Mackaay and P. Robillard. Predicting judicial decisions: The nearest neighbor rule and visual representation of case patterns. *Datenverarbeitung im Recht*, 3:302–331, 1974.

[13] M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28(2):237–266, 2020.

[14] F. Muhlenbach and I. Sayn. Artificial Intelligence and law: What do people really want?: Example of a French multidisciplinary working group. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*, pages 224–228, New York, 2019. ACM Press.

[15] C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[16] 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.