# On the relevance of algorithmic decision predictors for judicial decision making

Bex, Floris; Prakken, Henry

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

# On the relevance of algorithmic decision predictors for judicial decision making

Floris Bex
Utrecht University, The Netherlands
Tilburg University, The Netherlands
f.j.bex@uu.nl

Henry Prakken
Utrecht University, The Netherlands
University of Groningen, The Netherlands
h.prakken@uu.nl

## ABSTRACT

In this article, we discuss *case decision predictors*, algorithms which, given some features of a legal case predict the outcome of the case (i.e. the decision of the judge). We discuss whether, and if so how, such prediction algorithms can be used to support judges in their decision making process. We conclude that case decision predictors can only be useful in individual cases if they can give legal justifications for their predictions, and that only these legal justifications are what should matter for a judge.

## CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → *Natural language processing*; *Machine learning*.

## KEYWORDS

legal prediction, legal decision making, application of algorithms

**ACM Reference Format:**
Floris Bex and Henry Prakken. 2021. On the relevance of algorithmic decision predictors for judicial decision making. 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/10.1145/3462757.3466069

## 1 INTRODUCTION

The prediction of the decision of legal cases by means of machine-learning algorithms has become a hot topic [1, 3, 4, 12, 16]. Such algorithmic predictors can have various uses in the law. In this paper we discuss their application to support judges in individual cases, focusing on *algorithmic decision predictors*: algorithms that predict the final decision of a legal case, a decision that would otherwise be made by the judge(s) (such as guilty/not guilty, rule for plaintiff/defendant). Algorithmic decision predictors are sometimes claimed to improve the predictability and consistency of judicial decision making, which is demanded by the principle of equality (cf. [10]). According to these claims, judges can use decision predictors in order to come to more consistent, more informed and less biased judgments [4, 8, 17]. Others, however, fear that when judges' decisions are informed by algorithmic case predictors, people will not be judged any more on the legal merits of their individual case but on the basis of general statistics [19]. This is related to O'Neill's [18] criticism of 'bucketing', the practice of basing a decision about an individual (e.g., about granting the person a loan) on the fact that the individual is member of a particular class of which a statistical frequency is known instead of on the particular situation of that individual. O'Neill [18, 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.

To be able to evaluate this debate it is necessary to have a clear picture on what information a prediction of a decision by an algorithm in a particular case gives to the judge deciding the case. One answer is given in [4]: "an AI system can be trained to accurately forecast based on past behaviour what a user's decision would be in a situation absent lapses in rationality." So if an algorithm performs well on a test set and if it predicts a particular decision in a new case, then an arbitrary rationally-thinking judge would if assigned to the case, take the predicted decision. Of course, algorithms are rarely 100% accurate, so we look at the *probability* that an arbitrary competent judge assigned to the case would take a predicted decision. We want to investigate to what extent an algorithmic case prediction can yield such a decision probability: how, and under which assumptions, does a prediction in a particular case combined with information about an algorithm's performance on a test set yield a decision probability for a new case?

This last question immediately gives rise to a new question: why would judges be interested in probabilities at all when deciding a case? After all, we expect judges not to give probabilistic reasons for their decisions (except perhaps on matters of fact) but legal reasons. Still, judges have always looked at what their colleagues decide in similar cases and there are good reasons for doing so, such as improving the consistency of intra-judicial decision making [10, par. 8]. Underlying this is the assumption that if the great majority of their colleagues would take the same decision, then it presumably is the right decision. Of course, this assumption is at best defeasible and this leads to a second idea, namely, that if an algorithmic decision predictor performs well in the test phase, then its predictions yield the 'normal' decision of the case, so that a judge could only deviate from a prediction if there are special circumstances in the case. We also want to investigate to what extent such thinking is justified.

To address these questions, we first in Section 2 give a brief overview of the main types of algorithmic case predictors for legal cases. We then discuss in Section 3 the various senses in which probabilities can be derived given an algorithm and its evaluation with a test set. The heart of our paper is Section 4, in which we discuss to which extent the probabilities derived from an algorithm

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

ACM ISBN 978-1-4503-8526-8/21/06…$15.00
https://doi.org/10.1145/3462757.3466069

and its evaluation can be applied to a new case that is to be decided in court. Our main conclusion will be that in practice such an application is almost never warranted. We then in Section 5 discuss what this means for the hope that the use of algorithmic case predictors by judges in individual cases will improve the consistency and predictability of judicial decision making.

## 2 ALGORITHMIC DECISION PREDICTORS

Algorithmic decision predictors come in, roughly, three types: predictors on the basis of legally relevant factors, predictors on the basis of features unrelated to the merits of a case and predictors on the basis of the textual description of a case (for a recent overview see [3]). We focus on supervised classification algorithms that predict a categorical outcome – one of multiple possible decisions, such as affirm/reverse, guilty/not guilty – and not on algorithms that provide a continuous output – such as a regression algorithm that predicts the length of a sentence or the amount of damages to be paid. Furthermore, note that we only focus on predictors that, given some features of a case, predict the *final decision* of the case, and that we do not include, for example, algorithms for estimating recidivism risk, as these do not provide a final case decision.

*Predicting on the basis of legally relevant factors.* One approach predicts decisions on the basis of legally relevant factors in a case, by using either machine-learning techniques or a symbolic model of legal reasoning. [1] This approach describes the facts of a case at a higher level of abstraction than the concrete facts. The factors are assumed to be legally relevant for the case decision, so they can be used for generating informative explanations of a prediction.

The first studies into prediction on the basis of factors applied general machine-learning techniques to encodings of cases in terms of legally relevant factors. An early AI & law example is Mackaay & Robillard [14], who studied the prediction of a type of Canadian tax case with the nearest-neighbor rule. In AI & Law, various factor-based models for case-based reasoning have been used for generating knowledge-based case decision predictions without the use of machine learning techniques. Examples are the studies of Ashley and his PhD students on the case law concerning misuse of trade secrets in American law [2, 7]. Accuracy levels were obtained of up to 88% [2] and 92% [7]. An advantage of this approach is that the arguments generated about the predicted decision can be used as explanations of the prediction based on legal knowledge and in a form not unlike the arguments of human judges or lawyers.

*Predicting on the basis of case metadata.* Several authors have used supervised machine learning based on case features that are not related to the merits of the case. An example is the algorithm that predicts decisions of the American Supreme Court on the basis of structured metadata such as the kind of case, the date at which it was decided and which lower court decided the original case [12]. This algorithm, which correctly predicted 70% of the decisions, cannot explain the predicted decisions in a legally meaningful way, since the features on the basis of which it makes its predictions are 'extra-legal', that is, they are not related to the merits of the case.

*Predicting on the basis of the textual description of a case.* Other algorithms predict decisions based on the text of case law, where statistical correlations are identified between, for example, word combinations in the text and the case decision. Examples are algorithms that predict whether the European Court of Human Rights (ECHR) will for a specific article from the Convention with the same name decide whether that article was violated, on the basis of part of the text of the decision by the Court [1, 16] or the facts of the case as communicated to the parties [15]. The performance of these different algorithms is largely comparable, with accuracy and F-measures ranging between 75% and 80%. Although it would seem that this kind of algorithm looks at the legal aspects of the case (procedural history, facts), the identified statistical correlations do not say anything about the legally relevant reasons for the decision of a case. Therefore these algorithms can also not explain their predicted decisions in a legally meaningful way.

## 3 FROM ALGORITHM PERFORMANCE TO PROBABILITIES

Recall that we want to investigate to which extent the performance of an algorithm on a test set justifies the idea that an arbitrary competent judge assigned to a case will likely take the predicted decision. We call the probability at stake here the *decision probability*, the probability that an arbitrary competent and rational judge assigned to a particular case will take decision $X$ in that case, given that the algorithm predicts "$X$", that is, that the case will receive decision $X$. In formulas this is $Pr(X|``X")$, where "$X$" stands for the algorithm predicting decision $X$. The precise way in which this decision probability can be determined is for present purposes irrelevant, but the idea is that this probability can somehow be derived from the algorithm's performance on the test set. One candidate method is using *Precision*, the percentage of predictions "$X$" on the test set that are correct (i.e. the true positives divided by the total number of positively predicted cases). Interpreted as a frequency-type probability, the precision is $Pr(X|``X")$, which looks like the decision probability we are after. However, we do not commit to exactly this way of determining the decision probability – for present purposes, all that is relevant is that this decision probability will be defined in terms of an algorithm's application to a test set, and the crucial thing to note is that this makes the step to a probability of the same form for a new case that is not in the test set non-trivial.

## 4 APPLYING GENERAL FREQUENCY-TYPE PROBABILITIES TO NEW CASES

For the answer to the question how the step from a probability derived from performance on a test to a probability for a new case can be made, we turn to the philosophy of probability theory. Philosophers distinguish *frequency-type* and *belief-type probability.* Probabilities are of the frequency type if they are based on relative frequencies. Usually, the frequencies are relative to outcomes of experiments that can be repeated indefinitely, such as tossing a coin or rolling a dice, but we consider the special case where they are derived from a given finite set of test cases. Dawid [9] calls such probabilities 'statistical' probabilities. In contrast, probabilities are of the belief type if they are about the degree to which a proposition is believable. Such probabilities can also be attached to propositions

[1] 'Factors' are here not just CATO-style boolean factors but any abstract fact pattern that can have two or more values.

that a single event occurs. The probabilities that can be defined in terms of an algorithm's performance on the test set are all of the frequency type, since they are based on the relative number of true/false positives/negatives. However, what we want is a belief-type probability, namely, the probability that a given new case will be decided as predicted by the algorithm.

So what we are interested in is what information a prediction of a decision gives to a judge *in a particular case* that the judge has to decide. The italicised words are crucial, since when a probability is interpreted as a frequency (or in Dawid's [9] terms as a statistical probability), it does not by itself say anything about a particular case. As is well known (e.g. [11, p. 137]), there is a logical gap between frequencies and an individual probability: turning a frequency-type probability into a probability about a particular case is a decision, which has to be justified. Now how can this decision be justified? It turns out that this requires a number of assumptions.

### 4.1 From the test set to the set of future cases

Clearly, the move from the past to the future is only justified if the set of future cases has the same proportions as the test set. However, this is not guaranteed (see also e.g. [5, 6]). First, the decisions of judges can change in that they start deciding on different grounds or weighing reasons in different ways than they used to do. This can happen, for instance, when moral or political opinions in society change, or because different judges with different legal opinions are assigned to the same type of case. Also, the distribution of types of cases can change because of changes in the world. Moreover, the algorithm could be overfitted on inessential features of the training data (a well-known problem in statistics and machine learning). So (as is well known in the literature on machine learning) in order to accept a probability based on the test set as a probability for a future set of cases, we have to make at least the following assumptions: judges continue to decide cases on the same grounds; the frequency of the various types of cases remains the same; and the algorithm made its predictions on the test set for the right reasons.

### 4.2 Yielding a decision probability for an individual case

This is not yet all. If the assumptions listed in the previous section are justified, then all we know is that the frequency-type probability derived from the test set can also be applied to a future set of cases (which can be open-ended). However, we are not after a probability of a *kind* of event (decisions predicted by this algorithm) but after the probability of a *single* event (this decision predicted by this algorithm). The former can be frequency-type but the latter must be belief-type. We could apply the so-called frequency principle [11, p. 137] and let the latter equate the former. However, if we do so, that is, if we base our probabilities concerning individual case decision predictions on frequencies, then we in fact make a crucial assumption. This assumption is that *the only ways in which cases can relevantly differ is in the properties on which the relative frequencies are defined*, that is, on their real and predicted decision, just as in familiar text book examples about urns with coloured balls the only relevant way in which the balls can differ is in their colour. While in the textbook examples this assumption is justified, for legal cases it is not. Judges who have to decide a case know much more about it than its predicted decision. And the point is that if a judge has more information than just membership of the 'reference class' of the relative frequency (for instance, '80% of the cases with predicted decision $X$ have decision $X$'), then it is irrational to rely on the frequency-based probability concerning that class. Instead, one should look at the probability of the decision conditional on the more specific reference class that corresponds to one's knowledge about the case. And this, of course, amounts to thinking about the particulars of case as judges are used to do.

Our argument is an instance of what philosophers call the problem of finding the right reference class when performing 'direct inference'. It is this reference-class argument that gives a philosophical justification for O'Neill's [18] criticism of 'bucketing' and more generally for the fear of trial by statistics. In essence it means that if nothing more is known of an algorithmic decision predictor than its performance on the test set, then its predicted decisions cannot be regarded as the decision that an arbitrary judge assigned to the case would likely take. So a judge who wants to know what his or her colleagues would likely decide in an individual case, should not consult the algorithm since it does not provide the correct decision probability for the case. This in turn means that there is no meaningful sense in which an algorithmically predicted decision is the 'normal' decision for the case, from which a judge could only deviate if he or she can point at special circumstances that make this case different than a normal case of this kind.

To explain this further, imagine that cases are distributed in such a way that many cases are 'clear', for which a decision predictor would always be correct, but many other cases are 'hard', for which a decision predictor would often be incorrect, but the algorithm cannot explain to which type a new case belongs. Then only in the clear cases can the predicted decision be said to be the 'normal' one. But how can the judge know which case is easy and which case is hard? To know this, the judge has to think about the particulars of the case as judges always do. But then the judge can just as well ignore the algorithmic prediction.

### 4.3 Objections to the reference class argument

In the previous subsection we concluded that in practice it will be impossible to rationally derive a case-specific decision probability from frequency-type probabilities based on experiments with a test set, so that judges who want to know what their colleagues would likely decide in the case cannot obtain an answer to their question by consulting a case decision predictor. We now discuss possible objections to our reference-class argument for this conclusion.

First, it might be argued that it is still rational to stick to a statistical decision probability for a new case, since there often are no statistics on which a more specific frequency-type probability can be based. Yet this is a reasoning fallacy: if one wants to express a decision probability in such cases, one should take the additional information into account. If this cannot be done on the basis of known frequencies, one should form a probability based about one's information about the specific case, on the penalty of making the unfounded assumption that this additional information is irrelevant for the decision of the case (cf. [11, p. 137]).

A variant of this argument is the argument that a belief-type probability is always less well founded than a frequency-based probability, so that a judge who wants to know what his or her colleagues would likely decide can still look at what an algorithmic decision predictor with a high precision predicts. However, this argument fails, since if one knows more about the case, then sticking to the frequency is even less well-founded. Consider the analogy of an urn with 80% red balls and 20% blue balls. If this is all one knows and one draws a ball from the top of the urn, then it is rational to assume that there is an 80% probability that it will be red. But suppose now that the person who filled the urn tells you that he first put all the red balls in and then all the blue balls, and that he did not shake or stir, and that you take the ball from the top of the urn. It is now irrational to stick to the 80% probability that the ball will be red. In fact, the inverse probability (just 20% chance that the ball will be red) seems more rational.

One may also consider technical solutions to the reference class problem. The first is to inspect the test set to check the algorithm's performance on subsets of test cases of particular types, as an attempt to make it more likely that the class memberships considered for the algorithm's performance coincide with the knowledge the judge has about a particular case. This is a good idea in theory, but note that this approach in fact amounts to building a legal-knowledge model of the reasons relevant for a decision. Moreover, the created subclasses may be too small to yield reliable probabilities, since in the law the collections of cases usually are not very big [5]. Furthermore, a too fine-grained feature set may lead to an overfitted model that does not easily generalise [6, p. 9].

A second technical solution is to obtain the probability for a single case directly from the algorithm, that is, the probability of a certain decision $X$ given the set of features $F$ that represents the case, or $Pr(X|F)$. Simpler predictive algorithms directly output such a prediction probability for a single case, and for e.g. neural networks or support vector machines (SVMs) it is possible to estimate this probability based on the output of the model (cf. [20]). It can be argued that it is exactly this probability the judge needs: the algorithm captures the behaviour of the judges in the training set cases and then directly outputs the probability that these judges would rule $X$ in a case like the current one with features $F$. However, this still does not yield the probability that an arbitrary judge would rule in that way given the case, because there need not be a relation between the correctness of predictions and the prediction probability. For example, the algorithm can predict the wrong decision with a high probability, or the algorithm may over- or underestimate individual probabilities simply because this leads to better classification performance. Furthermore, using such advanced techniques brings along even more assumptions and makes it even harder to determine what exactly the given probability means, particularly for a judge with no background in statistics or machine learning. So instead of relying on this probability, the judge would be better off thinking about the particulars of case as normal.

A final objection is that an algorithm does not have to be perfect, as long as it performs better than human decision makers. Here sometimes the medical domain is mentioned, in which it is widely accepted that, for instance, a human oncologist has to consult a data-driven predictive algorithm for recognising skin cancer if this algorithm has been proven to perform better than humans [21]. However, this analogy beaks down, since unlike in the medical example, a legal predictive algorithm and a judge perform different tasks. In the medical example human and algorithm perform the same task, namely, recognising cancer in images of, for instance, birthmarks. Moreover, the estimates of human and algorithm are compared to the same (objective) truth: by examining the cells under a microscope it can be determined with certainty whether there is cancer. Thus a human expert and an algorithmic expert are compared in terms of the same standard. In such a case a comparison between how humans and algorithms perform is meaningful and the algorithm can be said to perform better than the human doctor, namely, by recognising malign spots missed by the human doctor. However, an algorithmic decision predictor performs a different task than the judge. A decision predictor predicts which decision a judge would take, which is a different task than the task the judge performs, which is *deciding* the case. Then it is meaningless to compare the performance of the algorithm and the human judge. What is more, even a correct prediction of a legally incorrect decision would count as a success for the predictive algorithm. Such situations may arise, for instance, since the test set contains legally incorrect decisions [5]. Correctly predicting a decision is not the same as predicting a correct decision.

## 5 CAN A DECISION PREDICTOR IMPROVE PREDICTABILITY AND CONSISTENCY?

In Section 4 we concluded that a judge who has to decide a case and who wants to know what an arbitrary rational judge assigned to the case would probably decide, cannot rely on the statistics provided by (the evaluation of) an algorithmic decision predictor. However, this leaves the question what other benefits consulting such an algorithm can have for a judge in an individual case. This we discuss below, focusing on the alleged benefit of improving the predictability and consistency of judicial decision making.

First, we have to determine what the terms predictability and consistency mean in this context. Assuming that they mean the same, there are two interpretations. One interpretation is that the *same* case is decided the same by different judges. 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 can now ask how an algorithmic decision predictor can be used in order to improve predictability and consistency. If these terms mean that the *same* case is decided the same by different judges, then a sure way to guarantee predictability and consistency is to give all judges the same algorithmic decision predictor and to require that they all 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, since as we argued in Section 4.3 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 both its accuracy and precision 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.

What if predictability and consistency mean 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, for mere decision predictors we cannot know. Suppose an algorithm with 90% precision predicts decision X for case C. Does the judge then treat like cases alike if s/he follows the prediction? We cannot know, since the prediction in itself would not give any information about similarity with other cases. In fact, it might well be that an algorithm treats cases that judges would regard as similar as different or vice versa (likewise [6, p. 6]). For example, text-based decision predictors like the ECHR predictor could fail to recognise that linguistically small differences are legally very relevant.

However, 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 a SCOTUS-like predictor is ruled out. But this implies that an ECHR-type predictor is also ruled out, since it cannot extract any legally relevant information from the texts to which it is applied, so there is no way to identify whether its prediction is based on legal grounds or on extraneous factors. Only decision predictors that base their predictions on legally relevant factors could possibly yield legally relevant information about similar cases to a judge.

However, we believe that only these legal explanations are what should matter for a judge, and that the judge should ignore the fact that a decision was predicted by an algorithm with good statistical performance on a test set. This use of such algorithms is not much different from how judges currently use other information sources, such as books, journals and peer consultation. Numerical performance indicators like accuracy, precision and recall can justify a degree of trust in algorithms in this general sense, but cannot indicate the quality of individual predictions or explanations. Moreover, evaluating the quality of algorithmic explanations for individual predictions requires validation studies of a kind that goes far beyond the current trend to focus on numerical performance measures like accuracy, precision and recall and is more akin to an older AI tradition of carrying out empirical validation studies with potential or actual users of the algorithm [13].

## 6 CONCLUSION

In this paper we argued that a judge who has to decide a case and who wants to know what an arbitrary rational judge assigned to the case would probably decide, cannot rely on the statistics provided by (the evaluation of) an algorithmic decision predictor. The idea that an algorithmic prediction that performed well on a test set yields the 'normal' decision of the case, from which a judge could only deviate if there are special circumstances in the case, is unfounded. Moreover, we argued that relying on the predictions of such algorithms cannot improve the predictability and consistency of judicial decision making in desirable ways. We believe that mere decision predictors, that is, predictors that cannot explain their predictions in legally meaningful terms, should not be used at all by judges as decision-support tools for individual cases. Such algorithms do not give any useful information to judges and may in fact be misleading and cause intellectual laziness.

If an algorithmic decision predictor gives any useful information to judges at all, it is not in its predictions but in its explanations for these predictions. However, we noted that whether algorithmic explanations 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, and instead involves potential or actual users of the algorithms. More generally, we believe that it is important to inform the legal world in transparent language about not only the potential benefits but also the limitations of algorithmic outcome predictors.

Finally, 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, but this requires another paper.

## REFERENCES

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science* 2 (2016), e93.

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

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

[4] B. Babic, D. Chen, T. Evgeniou, and A.-L. Fayard. 2021. The better way to onboard AI. *Harvard Business Review* (2021). http://nber.org/~dlchen/papers/The_Better_Way_to_Onboard_AI.pdf *To appear.*

[5] T. Bench-Capon. 2020. The need for good-old fashioned AI and law. In *International Trends in Legal Informatics: Festschrift for Erich Schweighofer*, W. Hötzendorfer, C. Tschol, and F. Kummer (Eds.). Editions Weblaw, Bern, 23–36.

[6] R. Binns. 2020. Analogies and disanalogies between machine-driven and human-driven legal judgement. *Journal of Cross-disciplinary Research in Computational Law* 1, 1 (2020).

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

[8] I. Chalkidis, I. Androutsopoulos, and N. Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4317–4323.

[9] P. Dawid. 2005. Probaility and Proof. (2005). http://tinyurl.com/tz85o Appendix to *Analysis of Evidence*, by T. J. Anderson, D. A. Schum and W. L. Twining.

[10] European Commission for the Efficiency of Justice (CEPEJ). 2018. European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment.

[11] I. Hacking. 2001. *An Introduction to Probability and Inductive Logic.* Cambridge University Press, Cambridge.

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

[13] R. O. Keefe. 1993. Issues in the verification and validation of knowledge-based systems. In *Advances in Software Engineering and Knowledge Engineering*, V. Ambriola and G. Tortora (Eds.). Series on Software Engineering and Knowledge Engineering, Vol. 2. World Scientific Publishing Co, 173–189.

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

[15] M. Medvedeva, , X. Xu, M. Vols, and M. Wieling. 2020. JURI SAYS: an automatic judgement prediction system for the European Court of Human Rights. In *Legal Knowledge and Information Systems. JURIX 2020: The Thirty-Third Annual Conference*, S. Villata, J. Harašta, and P. Křemen (Eds.). IOS Press, Amsterdam etc., 277–280.

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

[17] F. Muhlenbach and I. Sayn. 2019. 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.* ACM Press, New York, 224–228.

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

[19] F. Pasquale and G. Cashwell. 2018. Prediction, persuasion, and the jurisprudence of behaviourism. *University of Toronto Law Journal* 68, supplement 1 (2018), 63–81.

[20] J. Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[21] J. Susskind. 2018. *Future Politics: Living Together in a World Transformed by Tech.* Oxford University Press, Oxford.