Label Indeterminacy in AI & Law

Cor STEGING [a] Tadeusz ZBIEGIEŃ [b]
[a] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen*
[b] *Department of Legal Theory, Jagiellonian University*

Abstract. Machine learning is increasingly used in the legal domain, where it typically operates retrospectively by treating past case outcomes as ground truth. However, legal outcomes are often shaped by human interventions that are not captured in most machine learning approaches. A final decision may result from a settlement, an appeal, or other procedural actions. This creates label indeterminacy: the outcome could have been different if the intervention had or had not taken place. We argue that legal machine learning applications need to account for label indeterminacy. Methods exist that can impute these indeterminate labels, but they are all grounded in unverifiable assumptions. In the context of classifying cases from the European Court of Human Rights, we show that the way that labels are constructed during training can significantly affect model behaviour. We therefore position label indeterminacy as a relevant concern in AI & Law and demonstrate how it can shape model behaviour.

Keywords. Label indeterminacy, machine learning, legal decision-making

1. Introduction

With recent advances in machine learning, artificial intelligence is increasingly deployed in law, yet many approaches inherently contain characteristics that are incompatible with legal tasks. Examples include the intrinsic retrospective nature of machine learning, and the inability to faithfully explain their reasoning [1,2], and generative AI models suffer from issues such as hallucinations [3].

One crucial machine learning issue that has not yet been explored in the legal domain is *label indeterminacy* [4], where the labels of certain training instances are inherently unknown because an intervention by a decision maker has affected the final outcome. An example comes from medicine, where doctors must decide whether to continue or withhold life-sustaining treatment. If treatment is withheld, the patient does not survive, and it is impossible to know whether they would have survived had treatment continued. The resulting data is therefore *selectively labelled*, as the observed outcomes are contingent on the decisions of the doctor [5]. Historical cases where treatment was withheld thus yield indeterminate labels, and any machine learning model trained on such data must take this indeterminacy into account. Research from other domains has shown that the

This manuscript has been accepted for presentation as a short paper at the 38th International Conference on Legal Knowledge and Information Systems (JURIX) in Turin, December 9 to 11 of 2025.

way that we treat these indeterminate labels can have drastic effects on the behaviour and outcomes of AI models [4].

The legal domain also knows many scenarios in which the outcome has been affected by an intervention, and thus where label indeterminacy is at play. A famous example is that of granting bail: if an individual is not granted bail, we do not know how they would have acted if they were given bail [5]. If one were to train AI models on the historical cases, one would need to account for the pretrial detention cases, as those labels are indeterminate. If this is not accounted for, the model effectively learns to replicate past detention decisions rather than predicting the defendant's likelihood of returning for a court appearance, thereby potentially perpetuating historical bias, as illustrated by concerns raised as in the COMPAS system [6]. Excluding all non-bail cases is likewise problematic, since they do not constitute a random sample of bail cases.

Indeterminate labels can also arise in the case of settlements, where we do not know what the litigation outcome would have been if the case had been taken to court. Excluding settled cases from the training data can yield an unrepresentative model, as cases are not settled at random, and thus litigated cases alone may not be an accurate representative sample of all cases. Taking this one step further, in legal judgment prediction where cases could have been appealed but the judgment was accepted, we cannot know what the outcome would have been if the case had been appealed. The original judgment then may not reflect what an appellate court would have decided. Yet excluding original judgments from training data is problematic, since not all cases are or can be appealed based on the content.

In this paper, we critically examine the legal task of court case predictions. We examine when and why decisions in non-appealed cases should be considered indeterminate rather than ground truth, and what methods exist that can be used to impute the labels of these indeterminate cases. In an experiment using court case data from the European Court of Human Rights (ECtHR), we show that the commonly used methods to handle label indeterminacy can lead to different prediction behaviours of the models that we train.

2. Background

Machine learning approaches to automated legal prediction constitute an established subfield, with various models attempting to classify case outcomes across different legal domains [7,8,9,10,11]. Most of the previous research has used past case outcomes as fixed objective labels. This relies on an often implicit and rarely discussed modelling assumption that the final judgment represents an unambiguous resolution of the dispute and an objectively determined outcome. This assumption is particularly problematic in the context of label indeterminacy, where legal outcomes may depend on interventions such as appeals, settlements, or procedural decisions [12,13]. As the sub-field of automated legal prediction matures, core theoretical questions about what is meant by 'legal prediction' have emerged, along with considerations that must be addressed in practical applications [14,15]. In this context, the assumption that past case outcomes are always fixed and objective warrants revisiting, both from the computational and jurisprudential perspectives.

To clarify the scope of this paper, it is first useful to distinguish label indeterminacy from *selection bias*, in which datasets capture only a non-representative slice of all dis-

putes [16]. Both phenomena affect legal prediction, but in different ways. For example, excluding settled cases from a dataset introduces selection bias, as the remaining cases may not be representative of all disputes. In contrast, label indeterminacy arises when the outcome itself is unknowable: when settled cases are included in the dataset, the true court outcome of those cases cannot be determined. Consequently, a 'final judgment' in a dataset may therefore not reflect the whole story, depending on the qualities of a particular process.

Distinct from these data-centric issues is the discussion on *legal indeterminacy*, a jurisprudential debate that should not be conflated with label indeterminacy. The indeterminacy debate developed from the realist critique of legal formalism [17,18]. It argues that legal reasoning often fails to yield a single inevitable answer because legal rules and principles are underdetermined [19,20] . Critical legal studies scholars have further emphasized that the law functions as a political practice with outcomes shaped by existing power dynamics rather than by mechanistic rule-following [20,21].

One of the most influential rebuttals to strong forms of legal indeterminacy comes from Dworkin [22], who argued that even in hard cases there is, in principle, a 'right answer' discoverable through proper interpretation (see also [23]). This debate was of historical significance for the development of modern thinking about law and continues to influence it today.

More broadly, and apart from the debate on legal indeterminacy, scholars have noted that the notion of 'truth' in law is often constituted within institutional and argumentative practices. For example, in the legal context Patterson challenges the correspondence view of truth [24,25], and instead, locates truth in the accepted forms of legal argument. This makes 'truth' a matter of institutional endorsement rather than strict fact [26]. This linguistic-practice approach emphasizes that what counts as 'true' in law can be contingent on shared argumentative norms, rather than on an external, fixed ground truth [26]. The argumentative perspective has also gained traction within the AI & Law domain [27,28,29] .

While our aim is not to settle theoretical disputes about the indeterminacy of law or define what the ground truth in the legal domain actually means, there are grounds to argue that outcomes of legal disputes are not purely mechanical outputs of legal rules but are contingent and, at times, reversible products of procedural frameworks, substantive norms, and other difficult-to-capture contextual factors.

Treating all labels as static ground truths therefore risks misrepresenting the very phenomena machine learning models seek to capture. In our upcoming experiments, we show that the way we treat indeterminate labels can lead to major differences in model behaviour. Moreover, it has the potential to reinforce historical biases and ultimately undermine the reliability of AI applications in law.

3. European Court of Human Rights

We investigate label indeterminacy from a legal perspective in the domain of the European Court of Human Rights (ECtHR), where we develop a machine learning model to predict case outcomes. Predicting ECtHR case outcomes is a challenging and well-studied academic task [14], making it a suitable context for our study.

The ECtHR has several judicial formations, with most cases initially heard by a Chamber, a seven-judge panel that decides cases by majority vote [30]. In certain cir-

cumstances, either by referral or after relinquishment of jurisdiction by a Chamber, a case can be sent to the Grand Chamber, a 17-judge body that serves as the Court's final authority. Referral is not automatic, and many Chamber judgments remain final without Grand Chamber review. This procedural structure raises a key question for modeling: if a case is not referred, should the Chamber judgment be treated as definitive? Legally, the answer is nuanced: Chamber decisions are binding unless overturned or amended, yet they lack the authority of Grand Chamber judgments. From a machine learning perspective, we cannot model both simultaneously, as their decisions can conflict, for example, when a case is appealed and receives a different ruling from the Grand Chamber.

In this study, we decide to model the decisions of the Grand Chamber, the Court's highest authority, and treat these decisions as the ground truth. Consequently, Chamber decisions are considered indeterminate, as they do not necessarily align with the Grand Chamber's rulings. This framework allows us to investigate how different strategies for handling label indeterminacy, such as treating Chamber decisions as provisional or excluding them entirely, affect classifier behaviour.

We focus on cases brought under Article 6 of the European Convention on Human Rights, which concerns the right to a fair trial and constitutes by far the largest category of applications to the Court. We frame the task as a binary classification problem: determining whether a violation of Article 6 occurred (yes or no) based on the facts as documented by the ECtHR post-ruling. Technically, this falls into the larger category of outcome categorization tasks in legal NLP [14].

4. Method

This section details the setup of our experiment, including the dataset, the models employed, and the various existing strategies for handling the indeterminate labels of the Chamber cases.

4.1. Dataset

The dataset for our study was obtained from the ECHR Open Data project [31] and comprises ECtHR cases from 1968 to 2023. It includes 5,902 Chamber cases and 191 Grand Chamber cases related to Article 6. We format the dataset such that each case contains the facts section in natural language, the outcome label (violation or non-violation), and an indication of whether it was judged by the Chamber or the Grand Chamber. Our model is trained on the text of the facts section to predict case outcomes. To account for temporal effects in the law [15], we split the dataset by year of judgment: cases before 2015 form the training set, while cases from 2015 onward constitute the test set. The resulting dataset distributions are summarized in Table 1.

4.2. Model and preprocessing

Because ECtHR case files are long, we use an architecture capable of handling large documents. We employ a Longformer [1], which supports inputs of up to 4,096 tokens [32]. Analysis of our training data shows that Chamber cases average approximately 2,315 to-

[1] `https://huggingface.co/allenai/longformer-base-4096`

Table 1. Dataset distribution, where cases from 2015 and later are used as test set. Seven balanced train sets are generated from the train set.

Dataset	Total	Test Set		Train Set		Balanced Train Sets	
		Size	% Violation	Size	% Violation	Size	% Violation
Grand Chamber	191	34	47.06%	157	63.06%	116	50.00%
Chamber	5902	871	69.80%	5031	87.48%	1376	50.00%

kens, while Grand Chamber cases average 5,638 tokens (Table 2). Only 85.3% of Chamber cases and 45.9% of Grand Chamber cases fit within the 4,096-token limit. We first apply light preprocessing to remove structural noise and unnecessary tokens, followed by head-tail truncation for longer cases to make them fit within the token limit (Table 2).

Table 2. Token Statistics for Chamber and Grand Chamber Cases

Dataset	Metric	Cases	Max	Mean	Median	% $\leq$ 4096
Chamber	Raw	5031	75456	2317.9	1303.0	85.3
	Preprocessed	5031	4096	1756.0	1303.0	100
Grand Chamber	Raw	157	22029	5638.6	4466.0	45.9
	Preprocessed	157	4096	3275.6	4096.0	100
Total	Raw	5188	75456	2442.5	1354.0	84.1
	Preprocessed	5188	4096	1802.0	1353.0	100

As shown in Table 1, the label distribution is heavily skewed toward violations. To mitigate bias toward the majority class, we balance the datasets to a 50%-50% distribution. For both Chamber and Grand Chamber cases, we create seven balanced training datasets. Each dataset includes all non-violation cases and an equal number of violation cases, with minimal overlap in violation cases across datasets. This results in seven balanced Chamber training datasets containing 1,376 cases each, and seven balanced Grand Chamber datasets containing 116 cases each (Table 1). We retain the natural imbalance of the test set, accounting for it appropriately during evaluation.

4.3. Imputing indeterminate labels

In our experiment, we model Grand Chamber decisions and therefore treat Chamber decisions as indeterminate, since the outcomes could have differed if these cases had been appealed to the Grand Chamber. We apply and compare nine methods for imputing these indeterminate labels to investigate their effects on model behaviour. The label imputation methods largely follow previous work by Schoeffer et al. [4], and an overview is provided in Table 3. For each method, we briefly describe its procedure and the unverifiable assumptions underlying its application.

Correct Chamber (corr). In this approach, the default in the legal judgment prediction literature, the labels of Chamber cases are left unchanged, under the assumption that they reflect the decisions the Grand Chamber would have made. However, this assumption is not always valid, as the Grand Chamber may reach different conclusions than the Chamber, as evident in cases that were appealed and subsequently overturned.

ID	Name	Description	Assumption / Risk
corr	Correct Chamber	Includes Chamber cases without altering their labels	Assumes past Chamber decisions reflect Grand Chamber decisions
obs	Observed only	Trains only on Grand Chamber cases	Assumes missing at random; fails if appeal is outcome-related
$obs + ip$	Observed + IP	Trains only on Grand Chamber cases and applies inverse propensity weights to correct for sample bias	Assumes positivity and no unobserved confounders
nn	Nearest neighbor	Imputes Chamber labels using the most similar Grand Chamber case	Assumes a valid similarity metric
exp_{all}	All experts	Uses each expert label individually with equal weighting	Assumes all expert inputs are equally valid, including conflicting information
exp_{avg}	Average expert	Averages multiple expert assessments into one label	Assumes the mean is meaningful; removes disagreement information
exp_{max}	Max expert	Labels violation if at least one judge voted for it	Assumes a judge voting for violation is correct; dissenting votes disregarded
exp_{min}	Min expert	Labels non-violation if at least one judge voted against it	Assumes a judge voting against violation is correct; other votes disregarded
exp_{agr}	Expert agree	Keeps only cases with full expert consensus	Assumes unanimous agreement implies correctness; ambiguous cases excluded

Table 3. Overview of methods used to impute the labels of Chamber cases.

Observed only (obs). Here, all Chamber cases are excluded from the training dataset, leaving only Grand Chamber cases as ground truth. This method implicitly assumes that Chamber cases are missing at random. However, because Chamber cases are not a random subset and appeals may be outcome-dependent, excluding them can introduce bias into the training data.

Observed only with inverse propensity weighting ($obs + ip$). In this method, chamber cases are excluded, but inverse propensity weighting is applied to Grand Chamber cases during training to reduce sampling bias. As a result, Grand Chamber cases that are similar to Chamber cases are assigned higher weights. This method relies on two key assumptions. First, that every case has a non-zero probability of being appealed to the Grand Chamber (positivity), which does not hold in practice. Second, that there are no unobserved confounders, meaning the decision to appeal depends only on observable factors in the dataset, an assumption that cannot be verified.

Nearest neighbor (nn). In this method, we do include Chamber cases, and their labels are imputed based on similar Grand Chamber cases. Each case is first converted to embeddings using the Longformer. The embeddings are normalized, and the similarity between each Chamber case and all Grand Chamber cases is computed using the dot product. Each Chamber case is then assigned the label of the Grand Chamber case with the highest similarity. This method assumes that the similarity metric captures meaning-

ful legal similarity and that cases deemed similar according to this metric would result in the same decision.

All experts (exp_{all}). This and the following approaches require expert labels for Chamber cases. Each case is duplicated n times, once for each expert, and assigned that expert's label. Each instance receives a sample weight of $\frac{1}{n}$. In our setting, the seven judges on a Chamber case are treated as experts, with their votes for or against a violation serving as labels weighted by $\frac{1}{7}$. This method assumes all expert assessments are equally valid, even if they conflict, so the model is trained on potentially inconsistent information.

Average experts (exp_{avg}). Labels are represented as continuous values, calculated as the number of judges voting for a violation divided by the total number of judges n (seven in our case). A limitation of this approach is that it obscures individual disagreements, and the notion of an 'average expert' may lack validity in real-world applications.

Max experts (exp_{max}). A case is labeled as a violation if at least one judge votes for violation, and as non-violation otherwise. This corresponds to an 'optimistic' perspective in the literature, assuming that any indication supporting a positive label is correct. It presumes that judges voting for violation are correct and that dissenting judges are incorrect.

Min experts (exp_{min}). This method adopts a 'pessimistic' perspective, labeling a case as non-violation if at least one judge votes against violation. It assumes that judges voting against violation are correct while the others are not.

Experts agree (exp_{agr}). Only Chamber cases with unanimous votes, either for or against violation, are retained. The underlying assumption is that unanimous decisions are correct, which may not always hold in practice.

4.4. Experimental set-up

We apply each of the nine label imputation methods to each of our seven balanced training sets, resulting in a total of 63 distinct datasets for training the Longformer model. Following light empirical parameter tuning, all models are trained with a batch size of 2 and a learning rate of 2×10^{-5} for 3 epochs. Further details regarding model implementation and training procedures are available in our GitHub repository. [2]

5. Results

For each label imputation method, we average the results across the seven balanced train sets to investigate how it affects the behaviour of the model. Note again that we only consider Chamber Case decisions as ground truth. Table 4 reports the mean performance of the models on both test sets. In the top rows, we show the test set containing only Grand Chamber cases (34 cases). These serve as the ground truth and provide, to some degree, an objective measure of performance. The bottom rows of Table 4 show results for 871 Chamber cases. Although performance is higher for this set, the outcomes of Chamber cases are considered indeterminate, so these figures cannot be interpreted as true performance measures. Figure 1 illustrates the distribution of predictions for models trained using different methods of imputing indeterminate labels on both test sets, show-

[2] `https://github.com/CorSteging/Label-Indeterminacy-in-AI-Law`

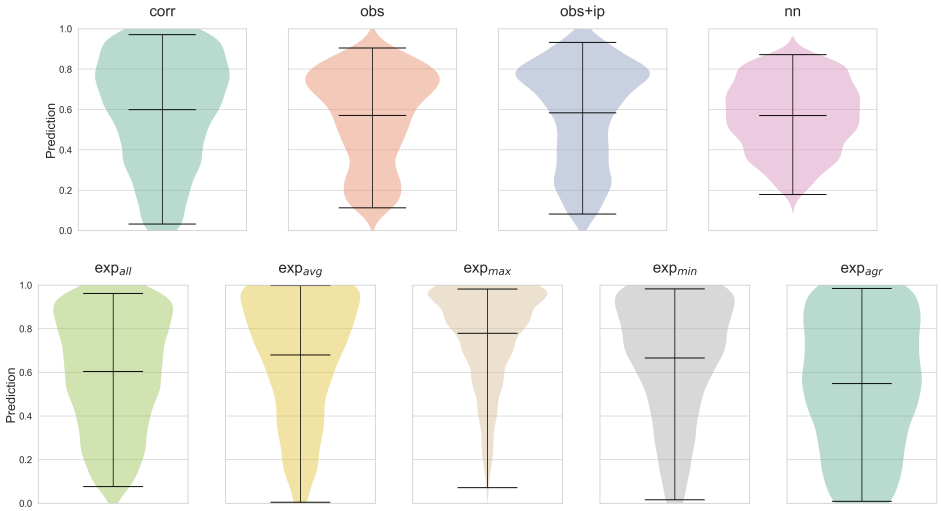
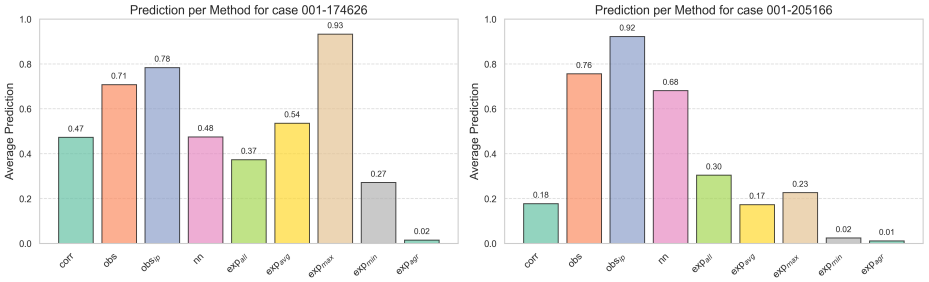


Figure 1. Violin plots displaying the distribution of the predictions for models trained using different label imputation methods.

Table 4. Matthews Correlation Coefficient (MCC) by Method for Grand Chamber cases (ground truth labels) and Chamber cases (indeterminate labels)

	Grand chamber test set								
	corr	obs	obs+ip	nn	expall	expavg	expmax	expmin	expagr
Mean MCC	-4.67	18.74	23.11	5.49	-5.86	1.89	1.42	3.86	1.29
Std Dev	6.10	19.96	17.36	10.97	14.71	17.31	19.20	16.55	13.40
	Chamber test set								
	corr	obs	obs+ip	nn	expall	expavg	expmax	expmin	expagr
Mean MCC	14.43	0.14	1.23	9.10	16.17	14.47	11.58	14.37	15.15
Std Dev	2.61	2.95	6.78	6.60	4.35	3.93	5.52	3.06	3.71



(a) (b)

Figure 2. The mean prediction for two specific cases using different label imputation methods.

ing that the choice of method influences the predictions of the model. In Figure 2, we show the different predictions for two specific cases, further demonstrating the effects of the various label imputation methods.

6. Discussion

When interpreting the results, it is important to note that Chamber case outcomes are considered indeterminate, so we cannot determine whether the model's predictions on these cases are correct. Moreover, all nine label imputation methods rely on unverifiable assumptions (Table 3), and none can be regarded as absolutely 'correct'. We chose these methods from the existing literature primarily to illustrate the problem of label indeterminacy in law, though some of their underlying assumptions may be especially implausible in legal contexts. This paper does not aim to provide a best-performing method or model, but rather to show why ignoring label indeterminacy is misguided.

Table 4 shows low MCC scores, indicating limited predictive performance. Prior work mostly reports accuracy and F1-scores [16], which do not fully capture all confusion matrix quadrants, a crucial consideration for test sets with heavily skewed label distributions (see Table 1). We also account for temporal effects by splitting the data into training and test sets based on judgment year, a setup shown to be more challenging [15]. That study reported mean MCCs between 8.16 and 26.30 depending on the test year. Using the last nine years as our test set, we observe an average MCC of 14.43 on Chamber cases when training the model 'as usual' (*corr*), which aligns with these expectations. Performance on Grand Chamber cases is generally lower and more variable, except for the *obs* and $obs + ip$ methods, which use only Grand Chamber cases for training and achieve higher results on the Grand Chamber test set.

Figure 1 shows that the model's predictive behaviour varies depending on the label imputation method used during training. For example, the exp_{max} method generally produces high predictions, approaching 1.0, while the *nn* method is more conservative, yielding lower predictive values.

To further highlight the impact of label imputation on predictive behaviour, Figure 2 shows the average predictions of models trained with each of the nine methods for two example cases. In Figure 2a, predictions range from 0.93 using the exp_{max} method to 0.02 with exp_{agr}, a difference of 0.91, with the other methods falling in between. Similarly, Figure 2b also shows a 0.91 difference, but between $obs + ip$ and exp_{agr}. Examining the two cases revealed no clear reason for the divergent predictions. Both are Chamber cases, one with a single dissent and the other decided by majority, suggesting that ordinary cases such as these can be strongly affected by how label indeterminacy is handled.

Overall, our results highlight the importance of accounting for label indeterminacy, as the choice of method can substantially affect model behaviour. Predicting ECtHR case outcomes is challenging even for humans, as evidenced by dissenting opinions, and serves mainly as an academic illustration. We therefore view the experiment as an illustrative case study that shows how different label imputation strategies influence model predictions. Similar issues arise in other legal contexts, such as settlements, withdrawals, procedural defaults, bail, or appeals. We therefore argue that both future research and developers in the legal domain should explicitly consider label indeterminacy when designing machine learning models.

7. Conclusion

In this paper, we introduce the concept of label indeterminacy to the legal domain and highlight its critical implications for machine learning. Certain legal outcomes are inherently unknowable because they depend on interventions whose absence or alteration could have produced different results. This creates indeterminate labels that cannot be ignored in predictive modeling. Using the ECtHR as a case study, we show that commonly applied label imputation strategies can yield different prediction behaviours, emphasizing the need to explicitly address indeterminacy when designing and evaluating AI systems in law.

Acknowledgements

This research was partially funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. This research was also supported by Future Law Lab under the Strategic Programme Excellence Initiative at Jagiellonian University. The authors wish to express their gratitude to Jakob Schoeffer, Piotr Bystranowski, Maciej Próchnicki, Bartosz Janik, Michał Rachalski, and Alexander Stachurski for their valuable feedback and support during the preparation of this article.

References

[1] Bench-Capon T. The need for good old fashioned AI and Law. International Trends in Legal Informatics: A Festschrift for E Schweighofer. 2020:22-36.

[2] Steging C, Renooij S, Verheij B. Rationale discovery and explainable AI. In: Schweighofer E, editor. Legal Knowledge and Information Systems - JURIX 2021: The Thirty-Fourth Annual Conference. vol. 346 of Frontiers in Artificial Intelligence and Applications. Vilnius, Lithuania: IOS Press; 2021. p. 225-34.

[3] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024;16(1):64-93.

[4] Schoeffer J, De-Arteaga M, Elmer J. Perils of Label Indeterminacy: A Case Study on Prediction of Neurological Recovery After Cardiac Arrest. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25. New York, NY, USA: Association for Computing Machinery; 2025. p. 1080–1094.

[5] Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 275–284.

[6] Flores AW, Bechtel K, Lowenkamp CT. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed Probation. 2016;80:38.

[7] Katz DM, Bommarito MJ, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States. PLOS ONE. 2017;12(4):e0174698.

[8] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Computer Science. 2016;2:e93.

[9] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317-23.

[10] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.
[11] Cui J, Shen X, Nie F, Wang Z, Wang J, Chen Y. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges; 2022.
[12] Galanter M, Cahill M. "Most Cases Settle": Judicial Promotion and Regulation of Settlements. Stanford Law Review. 1994;46(6):1339-91.
[13] Lever J. Why Procedure Is More Important than Substantive Law. The International and Comparative Law Quarterly. 1999;48(2):285-301.
[14] Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. Artificial Intelligence and Law. 2022:1-18.
[15] Steging C, Renooij S, Verheij B. Taking the law more seriously by investigating design choices in machine learning prediction research. In: ASAIL 2023. Automated Semantic Analysis of Information in Legal Text. Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023). Braga, Portugal: CEUR-WS; 2023. p. 49-59.
[16] Medvedeva M, Mcbride P. Legal Judgment Prediction: If You Are Going to Do It, Do It Right. In: Preoțiuc-Pietro D, Goanta C, Chalkidis I, Barrett L, Spanakis G, Aletras N, editors. Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore: Association for Computational Linguistics; 2023. p. 73-84.
[17] Hasnas J. Back to the Future: From Critical Legal Studies Forward to Legal Realism, Or How Not to Miss the Point of the Indeterminacy Argument. Duke Law Journal. 1995 Oct;45(1):84-132.
[18] Spaić B. Formalism. In: Sellers M, Kirste S, editors. Encyclopedia of the Philosophy of Law and Social Philosophy. Dordrecht: Springer; 2023. .
[19] Singer JW. The Player and the Cards: Nihilism and Legal Theory. Yale Law Journal. 1984;94(1):1.
[20] Unger RM. The Critical Legal Studies Movement. Harvard Law Review. 1983;96(3):561-675.
[21] Tushnet M. Critical Legal Studies: A Political History. The Yale Law Journal. 1991;100(5):1515-44.
[22] Dworkin R. No Right Answer? In: Hacker PMS, Raz J, editors. Law, Morality, and Society: Essays in Honour of HLA Hart. Oxford: Clarendon Press; 1977. p. 58-84.
[23] Whitman JQ. No Right Answer. In: Jackson J, Langer M, Tillers P, editors. Crime, Procedure and Evidence in a Comparative and International Context: Essays in Honour of Professor Mirjan Damaska. Oxford: Hart; 2008. p. 371-92.
[24] Aquinas T. Summa Theologiae. Cambridge, UK: Cambridge University Press; 2007. Original work 1265–1273, Prima Pars, Q.16.
[25] Coleman JL. Truth and objectivity in law. Legal Theory. 1995;1(1):33-68.
[26] Patterson DM. Law and Truth. New York: Oxford University Press USA; 1996.
[27] Governatori G, Bench-Capon T, Verheij B, Sartor G, Wyner A, Grabmair M, et al. Thirty years of Artificial Intelligence and Law: the first decade. Artificial Intelligence and Law. 2022;30(4):481-519.
[28] Sartor G, Araszkiewicz M, Atkinson K, Wyner A, Bench-Capon T, Governatori G, et al. Thirty years of Artificial Intelligence and Law: the second decade. Artificial Intelligence and Law. 2022;30(4):521-57.
[29] Villata S, Araszkiewicz M, Ashley K, Atkinson K, Bench-Capon T, Bex F, et al. Thirty years of Artificial Intelligence and Law: the third decade. Artificial Intelligence and Law. 2022;30(4):561-91.
[30] European Court of Human Rights. Rules of Court; 2025. Available at: `https://www.echr.coe.int/documents/d/echr/rules_court_eng`.
[31] Quemy A, Wrembel R. ECHR-OD: On building an integrated open repository of legal documents for machine learning applications. Information Systems. 2021:101822.
[32] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. arXiv:200405150. 2020.