Annotating Legal Cases from the European Court of Human Rights

Jack MUMFORD [a,1], Katie ATKINSON [a] and Trevor BENCH-CAPON [a]
[a] *Department of Computer Science, University of Liverpool, UK*

Abstract. In this paper we report on two user studies that have involved human annotation of European Court of Human Rights (ECtHR) datasets and discuss our plans for future enhanced studies. Legal cases from the court are annotated in accordance with an Angelic Domain Model (ADM) that is designed to capture the key elements used to determine the case outcome. These annotations are used to drive the training and testing of machine learning systems designed to effectively predict and explain legal case outcomes. We discuss our annotation pipeline, including annotator training, inter-annotator reliability evaluation, and the dissemination of the annotation outputs and associated metadata.

Keywords. legal case annotation, natural language processing, European Court of Human Rights

1. Introduction

The European Convention on Human Rights (ECHR) has proved a popular domain for AI and Law research, especially for the use of machine learning applications (e.g. [1], [2], [3]). The domain, however, has also been addressed using more traditional symbolic techniques [4]. Although, of course, effort is required to build the model, the symbolic approach is able to achieve considerably higher accuracy. While the machine learning approaches rarely achieve more that 80% accuracy, the symbolic approach gives accuracy of over 90%. Moreover, symbolic approaches can readily explain their results, whereas explanations from machine learning approaches tend to be rather unsatisfactory [5]. This has led some to suggest a hybrid approach, using machine learning to ascribe factors, and a model relating these factors to outcome to resolve the cases. In [5], the model is learnt by training a model on the ascribed factors, while in [6] a previously constructed symbolic model is used. The hybrid approach was first tried in [7], but there performance was not good: while the logical model achiever over 90% accuracy with manually ascribed factors, this fell to below 70% when factors were ascribed automatically. Since then, however, natural language techniques have improved immeasurably, and [5] claims good performance for this approach, albeit in a domain which is amenable to learning factor ascription.

The hybrid approach, however, requires annotated data with which to train the systems, and while the unannotated decisions of the ECHR are readily available[2], annotated

[1] Corresponding Author: Jack Mumford, email: jack.mumford@liverpool.ac.uk
[2] hudoc.echr.coe.int/

versions are not. To create such a dataset we undertook an exercise to annotate decisions with the elements of the symbolic model of Article 6 of the ECHR produced in [4]. We also conducted a second study in which the annotators were asked only to classify the cases into violations and non violations.

The symbolic model is described in Section 2. Section 3 describes and discusses the annotation exercises. Section 4 indicates further studies we wish to undertake, and Section 5 offers some concluding remarks.

2. Modelling the ECHR – Article 6

Although the ECHR has proved a popular domain for using machine learning approaches designed to predict the outcome of cases, such approaches have several problems.

- Explanation is, at best, rudimentary. Even when explanations are generated, as in [1], they lack a robust legal grounding and often hinge on spurious correlations such as dates and locations rather than on sound legal principles.
- Accuracy is not high: although some experiments have achieved results around 80%, the majority fall below this, often well below.
- The system may be using incorrect rationales [8], thus enforcing systematic wrong decisions.
- Performance tends to degrade over time [2]: the systems are trained on past decisions, and so cannot address the evolution of case law needed to predict future cases.

For these reasons, it can be argued that symbolic models retain an important place in such systems. Legal reasoning within the ECtHR context, particularly under Article 6 —- the right to a fair trial -— demands a nuanced approach. Article 6 stands out not only due to its prominence as the article with the greatest number of applications, but also because of its procedural and objective nature, making it a fitting testbed for foundational work in developing practical, principled machine learning systems. There is a scarcity of models that incorporate expert-informed legal knowledge, an exception being the Angelic methodology ([4], [9], [10],), which encapsulates the hierarchical reasoning structures used by legal experts to navigate complex cases in an Angelic Domain Model (ADM).

An ADM is crafted with the help of legal experts to reflect a factor-based legal model [11], where legal factors are abstract patterns of fact that are used to establish prioritised sufficient conditions for the resolution of relevant legal issues for the case, which in turn provide necessary and sufficient conditions for the case outcome. The development of the ADM for Article 6 focuses on this structured reasoning, providing an explainable tool for determining case outcomes.

Recent research has established a hybrid architecture ([6], [12]) that combines BERT-based NLP [13], specifically a hierarchical BERT structure (H-BERT) [14], with the ADM. The H-BERT NLP layer is used to interpret the case texts and ascribe findings to the leaf-factor nodes of the ADM, after which the internal logic of the ADM determines the case outcome. This hybrid structure not only allows for explainable predictions at a macro-level of legal abstraction via the ADM, but also affords micro-level insight through the attention mechanisms, which link case facts to specific ADM factors.

	A	B	C	D	E
1	ID	001-177388	001	001-	001-
2	DATE	10\/10\/2017	09\	27\/	11\/
3	CASE	CASE OF TAŞÇI v. TURKEY	CA	CASE	CASE
4	URL	https://hudoc.echr.coe.int/	htt	https	https
5					
6	CRIMINAL LIMB?	N	Y	N	Y
7	FLAGGED?				Y
8					
9	FAIR	N	N	N	Y
10	Right Not to Incriminate Oneself				
11	Independent and Impartial	N			
12	Equality of Arms and Adversarial Hearing				
13	Integrity of Evidence				
14	Access to Court		N	N	
15	Legal Certainty is Upheld				

(a) User study 1: Excerpt from participant excel file for ADM annotations.

	A	B	C	D	E
1		Circumstances	Domestic	rcumstanc	Domestic
2	Case file	001-154595_circumstances	001-154595_domestic	001-16467	001-16467
3	Flagged?				
4	Your verdict	V	v	N	n
5	Comment				
6					

(b) User study 2: Excerpt from participant excel file for outcome predictions file.

Figure 1. Example excerpts from participant excel files for both user studies.

Preliminary findings suggest that this hybrid architecture enhances performance, especially in contexts with sparse data. Nonetheless, the success of the ADM factor ascription is highly dependent on the availability of robustly annotated datasets. Therefore, the need for meticulously annotated corpora cannot be overstated, as it provides the foundational knowledge enabling ML applications to bridge the gap between technical prediction and principled legal reasoning.

3. Annotating ECtHR Legal Cases

In this section we begin by describing the design of our annotation studies on the ECtHR Article 6 legal case corpus[3]. After which, we report on the inter-annotator agreement results. This is followed by a discussion on the effect of different training regimes on annotator performance. Finally, we provide details of our approach to dissemination of the output annotated datasets and associated metadata.

3.1. Study Design

Our research efforts are focused on the capabilities of AI systems to predict and explain the outcomes of cases pertaining to Article 6 of the ECHR. We undertook two user studies to inform and guide this investigation.

User Study 1: ADM-Based Annotations
The first study aimed to generate a dataset rich in annotations aligned with the legal reasoning structure of Article 6 cases, as encapsulated by the ADM. Our participants, composed of 27 final-year law undergraduates -— 16 with specific ECHR university-level study and 11 without -— who were tasked with annotating a legal case corpus. They ascribed from natural language descriptions from the LAW section of cases to corresponding nodes of the ADM, a process detailed in our recent publication [6].

This study provided two-fold benefits: it facilitated the understanding of the distribution of legal factors and issues relevant to Article 6 resolutions; and enriched AI system training for enhanced outcome prediction and explanatory capabilities. Participants un-

[3]Ethical approval was applied for and granted by the University of Liverpool for both user studies undertaken that are reported in this paper.

Table 1. Inter-annotator agreement proportion and Fleiss' kappa score across violation, non-violation, and all cases (results reported to 3 d.p.). Results are reported to compare the 'Domain' and 'Non-domain' knowledge groups, as well as performance by all participants.

		agreement	kappa score
Violation	Domain	**0.923**	**0.633**
	Non-domain	0.893	0.544
	All participants	0.898	0.551
Non-violation	Domain	0.901	0.571
	Non-domain	**0.905**	**0.585**
	All participants	0.901	0.570
All cases	Domain	**0.908**	**0.597**
	Non-domain	0.901	0.583
	All participants	0.899	0.574

derwent two hours of training on the first day and subsequently dedicated two hours per day over four days to the annotation task. Annotations involved ascription indicators -— 'y' for positive ascription, 'n' for negative ascription, or a blank for non-ascription -— within provided Excel spreadsheets, which also contained basic case metadata and links to the case summaries on HUDOC. Regular discussions after every hour of work allowed for reflective conversations and collaborative problem-solving. The study concluded with a survey capturing the participants' experiences and feedback.

User Study 2: Outcome Prediction

Our second user study was designed to establish a human performance benchmark in outcome categorisation – a prevalent AI task in legal predictions [15]. This study, described in our recent publication [16], provided insight into effective prediction mechanisms based solely on the facts of a case, as described in the FACTS section of its summary document available on HUDOC.

We enlisted 41 final-year undergraduates, drawing from computer science, general law, and those with ECHR specialisation, designated respectively as **Weak**, **Moderate**, and **Strong** domain knowledge groups. The students were further divided, with 19 receiving training on the use of the ADM and 22 serving as a control group without ADM access. The study entailed two hours of initial training and eight hours of annotation work split into zero-shot and few-shot conditions. Participants annotated predictions of 'v' for violation and 'n' for no-violation based on the case facts presented in text files, making a first prediction based on the *circumstances of the case* and then a second prediction after reading the supplementary *relevant legal framework*. No further assistance was provided beyond the initial training and technical support.

Post-annotation, participants completed a survey to reflect on their experience and confidence in their categorisation performance. This setup offered a nuanced understanding of how varying degrees of domain knowledge impacted the accuracy and reliability of outcome predictions.

3.2. Inter-annotator Agreement Results

Interpretation of Agreement Measures

In assessing the reliability of our annotations, we employed Fleiss' kappa [17] – a statistical measure that is well-suited for evaluating agreement between multiple raters. The

Table 2. Inter-annotator agreement proportion and Fleiss' kappa score across different domain knowledge setups (results reported to 3 d.p.). Results are reported to compare: groups provided the ADM and those without, across all reviewed cases regardless of zero-shot or few-shot conditions; and zero-shot output and few-shot output, regardless of model provision.

		agreement	kappa score
ADM model	Weak	0.636	0.156
	Moderate	0.693	**0.303**
	Strong	0.669	0.191
	Overall	**0.709**	0.258
No model	Weak	0.558	0.047
	Moderate	0.682	**0.256**
	Strong	0.606	0.130
	Overall	**0.686**	0.236
Zero-shot	Weak	0.648	0.127
	Moderate	0.720	**0.296**
	Strong	0.634	0.181
	Overall	**0.742**	0.274
Few-shot	Weak	0.582	0.076
	Moderate	0.688	0.178
	Strong	0.595	0.073
	Overall	**0.715**	**0.207**

choice of Fleiss' kappa acknowledges the complexity of the annotation task at hand, involving multiple individuals, each bringing a unique interpretation of the legal cases based on varying levels of domain knowledge. Fleiss' kappa provides a more balanced and robust assessment than simpler percent agreement calculations, particularly in contexts where more than two raters are involved.

The interpretation of kappa scores, however, presents a challenge. The conventional benchmarks for kappa [18,19] have been critiqued for their derivation from studies with only two or three raters and binary categories. These benchmarks, presented in Table 3, are often improperly generalised to studies with larger cohorts of raters and more diverse categories, leading to potential misinterpretations of the kappa scores. We can, however, use the interpretations prescribed in Table 3 as pessimistic lower bounds for the quality of annotator agreement present in our studies. Moreover, the results of our two studies, with their substantial rater groups, will provide more valid benchmarks for direct comparisons to similarly sized studies.

Comparative Insights from the Studies

Our comparative analysis of the two studies indicates a higher level of agreement amongst raters in the first study compared to the second. The first study's results, as reported in Table 1, showcase that domain knowledge positively influences the level of agreement, especially in cases of violations. Non-domain participants still maintained a high level of agreement, which may reflect the structured training provided and the objective nature of Article 6. Our raters achieved a very high proportion of agreement in their annotation, with the kappa scores indicating strong reliability of agreement given the significant number of subjects – 23 nodes from the ADM across hundreds of cases.

Table 3. Interpreting κ values for a 2-annotator 2-category example [18]. Often inappropriately applied to annotated data involving larger cohorts of annotators and categories

κ	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

The second study's findings, reported in Table 2, suggest low agreement scores regardless of the various types of domain knowledge represented. However, at different granularities of investigation, patterns do emerge: the 'Moderate' domain knowledge group attained slightly higher agreement scores compared to the 'Weak' and 'Strong' groups, as did those participants provided with the ADM model compared to those without, and interestingly the same was true zero-shot agreement compared with few-shot outputs. These annotator agreement observations broadly align with the outcome classification performance reported in [16], such that greater annotator agreement corresponds to more accurate outcome classification, with the exception that participants were able to correctly classify outcomes with slightly greater reliability under few-shot conditions compared with zero-shot.

Addressing Study Limitations

While the first study yielded rich annotations, it is imperative to note the absence of a 'gold standard' set of annotations established by legal experts. The reliance solely on student annotations, even those with domain expertise, constitutes a limitation as it might overlook the nuanced judgments a seasoned practitioner would provide. However, the high levels of agreement, including the near perfect majority agreement scores reported in [6], in conjunction with the alignment of the ADM content and reviewed text, suggest that the outputs were of a reliably good quality.

The second study also presents its limitations, particularly the indistinct nature of zero-shot and few-shot datasets. The lack of a clear demarcation complicates comparison, since some cases will have been reviewed under zero-shot conditions that other participants reviewed under few-shot conditions.

Conclusion

Overall, the inter-annotator agreement results provide valuable insights into the complexities of annotating legal texts and the capabilities of varying levels of domain knowledge in contributing to AI prediction and explanation in the realm of ECHR Article 6 cases. These findings underscore the importance of considering rater cohort size in kappa interpretations and the imperative for rigorous standards in annotation practices to support the development of AI systems in legal domains.

3.3. The Importance of Annotator Training

Our two studies present contrasting training regimes that provide valuable insights into the importance of comprehensive training for complex tasks such as legal annotation.

Study Training Regimes

In the first study, the training was extensive and interactive. Participants received a thorough introduction to the study's aims, supplemented by the actual decision-making model (ADM)[4] documents for Article 6 (note that a more up-to-date version of the ADM was made available for the second study). This was followed by practical annotation sessions, both individually and in groups, with subsequent whole cohort discussions with the study organisers (including an academic with legal expertise). This comprehensive approach was designed to maximise the annotators' comprehension and, consequently, their agreement rates.

The second study adopted a different approach, tailored more towards a real-world learning experience akin to that of current data-driven NLP systems. Training was less extensive, with an initial presentation and a brief phase of familiarisation with the annotation tools. Practice cases did not come with predefined outcomes to encourage a zero-shot learning approach. Those annotators provided the ADM[5] were given less time with the practice cases, which was replaced with time to examine the ADM and ask questions. An open-book test on the ADM was then followed by more time to read and question. A final closed-book test on the ADM was administered with these test results serving as an indicator of each annotator's understanding of the ADM. This was intended to measure the ability to learn and apply legal reasoning without the benefit of extensive training or expert guidance. All annotators completed four hours of outcome prediction annotation under zero-shot conditions, before being allocated one hour additional training with eight practice cases with pre-defined outcomes. This ensured annotators could conclude their annotations for the remaining four hours of work under few-shot conditions.

Training Insights and Annotator Performance

No significant differences were observed among the various participant groups, which may point to limitations in university education for preparing students for the task of annotation of legal cases, especially outcome prediction. However, standout performances by certain individuals, particularly in the second study, hint at the potential significance of careful annotator selection beyond training efforts.

The first study's high agreement rates suggest that thorough training can serve as a strong levelling force among annotators. In the second study, while the ADM-provided groups did not show a significant correlation between ADM knowledge test performance and outcome classification performance, they did demonstrate higher overall agreement and less variance, as indicated in Table 2.

Recent research [20] suggests that GPT-4 may be able to match the performance of competent human annotators in legal text annotation. Nonetheless, the complexity of different legal domains and annotation tasks may pose varying levels of difficulty, as evidenced by our studies, where the first yielded much higher agreement than the second. We propose that the first study's training level was adequate for the complexity of the task, while the second study's training may have been insufficient without expert-led worked examples.

Legal Interpretation and Machine Learning

[4] Available at `https://github.com/jamumford/ECHR_Article6_ADM_Ascribe`

[5] Available at `https://github.com/jamumford/Human_Legal_Verdict_Prediction`

Interestingly, annotators from the second study did not show significant performance improvement after reading additional legal context, potentially due to insufficient training on how to integrate this information into case analysis. The conjecture that participants could benefit from expert guidance underpins the importance of including robust legal interpretation as a crucial aspect of reliable prediction.

The observations from the second study do not represent a training failure but rather an intentional design choice reflecting the learning curve of contemporary data-driven NLP systems and their potential limits in mapping to human legal reasoning [21]. These insights should guide the training of machine learning systems, highlighting the necessity to integrate advanced techniques like information retrieval and temporal embedding in order to connect cases to relevant precedents. Such incorporation is vital not only for improving prediction performance but also for enhancing the models' ability to provide human-friendly explanations and justifications.

The lessons from these studies emphasise the need for a multidisciplinary approach in training annotators and developing AI systems, where legal expertise, cognitive insights, and technical knowledge converge to produce reliable and explainable outcomes in legal AI applications.

3.4. Annotated Dataset and Metadata Dissemination

We have made our datasets from both studies available to the public on an open-access basis, providing detailed annotation outputs and supporting metadata.

Accessibility of Data
The datasets, inclusive of anonymised individual annotator outputs, are readily accessible: for the first study[6], we have additionally provided summary annotation files; for the second study[7], the annotated outputs are presented in the same file format used by the annotators. The choice of data format reflects the intended use and audience. The first study's dataset is provided in JSON format, acknowledging its richness and potential utility for academic researchers who require structured and machine-readable data. Conversely, the second study's annotations are in XLSX format, as this was the format used by annotators for its simplicity and ease of use during the annotation process. Though not primarily intended for extensive academic application, the XLSX files are shared to maintain transparency and to allow for the reproduction of results.

Data Annotation Process and Quality
The use of XLSX files for annotations was a pragmatic decision that favored ease of setup. However, this choice necessitated substantial post-processing to rectify errors and inconsistencies in the data. This experience underpins the argument for developing bespoke annotation platforms that can enforce quality standards and ensure the integrity and uniformity of the data collected.

Metadata and Case Information
The metadata associated with the legal cases is primarily composed of case identifiers, which can be used to retrieve full case details from the HUDOC database. This approach

[6]Available at `https://github.com/jamumford/ECHR_Article6_ADM_Ascribe`
[7]Available at `https://github.com/jamumford/Human_Legal_Verdict_Prediction`

balances the need for data richness with privacy concerns and the practicality of data sharing.

Evaluation Tools and Scripts
To assist in the replication of our reliability and productivity evaluations, we have provided scripts at the corresponding GitHub repositories. These tools are instrumental for researchers who wish to analyse the data further or apply the methodology to their own datasets. In making these resources available, we aim to foster a collaborative and open research environment. The datasets and tools can serve as a foundation for future studies, encouraging methodological rigor and innovation in the ongoing exploration of AI and legal judgment prediction and justifiable explanation.

4. Enhanced Annotation Studies

In the pursuit of advancing legal technology and its application in real-world judicial processes, we are embarking on a project that fosters an interdisciplinary collaboration between academic Law and Computer Science departments. The objective is to develop datasets and computational techniques that will augment the determination of legal case outcomes and remedies, injecting a new level of efficiency and clarity into the justice system. In this section we discuss two projects planned for continued research on legal decision support tools for ECtHR cases.

4.1. Legal Case Difficulty and Outcome Prediction and Explanation

This ambitious project is structured around a two-stage pipeline that marries the analytic strengths of artificial intelligence with the nuanced understanding of legal experts.

Stage One: Predicting Case Difficulty
The first stage is dedicated to assessing how to gauge the difficulty of a case through an analysis of its factual narrative. By harnessing explainable AI, we aim to create models that can not only assess the complexity of cases brought before the European Court of Human Rights (ECtHR) but also elucidate the reasons behind their difficulty ratings.

Stage Two: Predicting and Justifying Outcomes
The second stage addresses the challenge of forecasting case outcomes and furnishing legally sound justifications for them. This step is critical for validating the AI's decisions and ensuring that they align with the judicial reasoning processes, thus reinforcing transparency and trust.

Objectives and Collaborative Endeavors
Our approach is defined by several key objectives:

1. Prediction of Case Difficulty: Develop explainable AI models that can reliably identify the complexity of ECtHR cases from their factual descriptions.
2. Outcome Prediction and Justification: Develop explainable AI models that can reliably predict case outcomes and provide reasoned justifications, ensuring transparency and understanding of judicial decisions.

3. ECtHR Collaboration: Explore the integration of our AI tools into the workings of the ECtHR for transformative operational change and enhanced access to justice.

Co-Creation and Annotation Studies
Our methodology is grounded in early and active collaboration with stakeholders at the ECtHR. This engagement will enable production of a specification capturing features important for reliable identification of case difficulty and outcome predictions.

In alignment with this, our annotation studies will engage law students to produce and refine a dataset that will be instrumental in training AI models for both identifying case complexities and predicting outcomes. This dataset will leverage 'case importance' metadata from HUDOC, combined with student evaluations of pre- and post-decision case documents, to discern if tangible content differences exist that affect interpretation of case difficulty. The output annotated dataset will serve to attribute case difficulty to meaningful rationales informed by the specification co-created with the ECtHR.

Technical Development and Evaluation
The annotated dataset forms the backbone of our technical development, enabling us to integrate advanced NLP techniques with established ADM-based explainable AI frameworks. This integration will not only streamline the prediction of case difficulties and outcomes but also enhance the analytical capabilities of AI with respect to legal reasoning.

The culmination of this project will be an exhaustive series of user evaluations, involving law students and professionals from the ECtHR. These evaluations will not only test the efficacy of our AI tools but will also provide insights into their practicality, potentially shaping their future adoption in the legal field.

In summary, this enhanced annotation project aims to build a bridge between technological innovation and legal expertise, ultimately contributing to a more transparent, accessible, and efficient legal system.

4.2. Legal Case Summary Generation

A second project is also planned to investigate the integration of generative models into the prior work to determine the level of support that can be given to the task of case summary generation.

Objectives and Collaborative Endeavors This project will focus on the following objectives:

1. Making use of the recent research produced on explainable AI [6] and couple this with generative AI applications to produce tools for drafting ECtHR case summaries;
2. Evaluating the effectiveness of the new tools produced in providing well-justified, legally-grounded content for case summaries;
3. Exploring the possibility of deployment of the prototype tools in an applied policy scenario.

Technical Development and Evaluation

To meet the aforementioned objectives, the tasks to be carried out are: i) early engagement with policy stakeholders for co-creation of a specification capturing features needed in case summaries; ii) development of a technical pipeline to enable Large Language Models to interface with the existing explainable AI tools for producing the case summaries; iii) conducting annotation studies with law students to produce and evaluate a dataset to train the AI models on the task of legal factor ascription from source documents; iv) conduct a user evaluation exercise with policy stakeholders to determine the effectiveness of the tools and consider their further implementation in practice.

This project will demonstrate the viability of combining the latest AI techniques for tackling another key legal task, and use domain knowledge captured in the annotation exercise to give focused guidance to the tools being developed and trained to produce case summaries.

5. Concluding Remarks

We have described user studies that have been undertaken by human participants to annotate cases from the European Court of Human Rights (ECtHR) for the purpose of informing the development of AI tools for legal case-based reasoning. Running such studies takes considerable preparation and planning time and we hope that by reporting upon our experiences, we are sharing valuable reflections on the annotation task and the outcomes achieved. Emboldened by the success of the studies completed so far, we have detailed further studies planned to tackle open research challenges that make use of the latest AI techniques, but which benefit from annotated datasets for the training of the machine learning aspects of the pipeline.

Acknowledgements

Our annotation studies were supported by funding from Research England under the Policy Support Funding stream.

References

[1] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the ECHR: A natural language processing perspective. PeerJ Computer Science. 2016;2:e93.

[2] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. AI and Law. 2019:1-30.

[3] Chalkidis I, Androutsopoulos I, Aletras N. Neural legal judgment prediction in English. arXiv preprint arXiv:190602059. 2019.

[4] Collenette J, Atkinson K, Bench-Capon T. Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. Artificial Intelligence. 2023;317:103861.

[5] Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, et al. Scalable and explainable legal prediction. AI and Law. 2021;29(2):213-38.

[6] Mumford J, Atkinson K, Bench-Capon T. Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases. In: Proceedings of the 19th ICAIL; 2023. p. 167-76.

[7] Ashley KD, Brüninghaus S. Automatically classifying case texts and predicting outcomes. AI and Law. 2009;17(2):125-65.
[8] Steging C, Renooij S, Verheij B. Discovering the rationale of decisions: towards a method for aligning learning and reasoning. In: ICAIL 2021. ACM; 2021. p. 235-9.
[9] Al-Abdulkarim L, Atkinson K, Bench-Capon T. A methodology for designing systems to reason with legal cases using ADFs. AI and Law. 2016;24(1):1-49.
[10] Atkinson K, Bench-Capon T. ANGELIC II: An Improved Methodology for Representing Legal Domain Knowledge. In: Proceedings of the 19th ICAIL; 2023. p. 12-21.
[11] Aleven V. Teaching case-based argumentation through a model and examples [Ph.D. thesis]. University of Pittsburgh; 1997.
[12] Mumford J, Atkinson K, Bench-Capon T. Reasoning with Legal Cases: A Hybrid ADF-ML Approach. In: Proceedings of JURIX 2022; 2022. p. 93-102.
[13] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
[14] Lu J, Henchion M, Bacher I, Namee BM. A sentence-level hierarchical bert model for document classification with limited labelled data. In: Proceedings of DS 2021. Springer; 2021. p. 231-41.
[15] Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. AI and Law. 2023;31(1):195-212.
[16] Mumford J, Atkinson K, Bench-Capon T. Human Performance on the AI Legal Case Verdict Classification Task. In: Proceedings of JURIX 2023; 2023. .
[17] Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971;76(5):378.
[18] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.
[19] Viera AJ, Garrett JM, et al. Understanding interobserver agreement: the kappa statistic. Family Medicine. 2005;37(5):360-3.
[20] Savelka J, Ashley KD, Gray MA, Westermann H, Xu H. Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In: Proceedings of the 6th ASAIL. vol. 3441. CEUR-WS.org; 2023. p. 1-12.
[21] Bex F, Prakken H. On the relevance of algorithmic decision predictors for judicial decision making. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law; 2021. p. 175-9.