Effects of XAI on Legal Process

Aileen Nielsen
aileen.nielsen@gess.ethz.ch
ETH Zurich, Center for Law and Economics

Stavroula Skylaki
laura.skylaki@thomsonreuters.com
Thomson Reuters Labs

Milda Norkute
milda.norkute@thomsonreuters.com
Thomson Reuters Labs

Alexander Stremitzer
astremitzer@ethz.ch
ETH Zurich, Center for Law and Economics

ABSTRACT

Despite strong scholarly interest in explainable features in AI (XAI), there is little experimental work to gauge the effect of XAI on human-AI cooperation in legal tasks. We study the effect of textual highlighting as an XAI feature used in tandem with a machine learning (ML) generated summary of a legal complaint. In a randomized controlled study we find that the XAI has no effect on the proportion of time participants devote to different sections of a legal document, but we identify potential signs of XAI's influence on the reading process. XAI attention-based highlighting may change the spatio-temporal distribution of attention allocation, a result not anticipated by previous studies. Future work on the effect of XAI in legal tasks should measure process as well as outcomes to better gauge the effects of XAI in legal applications.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in visualization**.

KEYWORDS

explainable AI, legal process, user attention allocation, human-computer interaction, experimental study

ACM Reference Format:
Aileen Nielsen, Stavroula Skylaki, Milda Norkute, and Alexander Stremitzer. 2023. Effects of XAI on Legal Process. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594536.3595128

1 INTRODUCTION

Much scholarship about the future of Artificial Intelligence (AI) in law stresses the need for accountability and human supervision of AI powered tools, often through the use of explainability. Among the instrumental motives for research on eXplainable AI (XAI) in the legal domain is the need to ensure that AI does not distort outcomes in the legal process [8]. In studying legal process, we refer to the method by which legal work is carried out, including all that pertains to procedural justice concerns but also the day-to-day manner in which legal professionals go about their work.

XAI focuses on making AI results more understandable to humans by enabling AI systems to provide insights about the reason why they reach certain outputs [1]. There is a wealth of approaches that can be used to surface the underlying explanatory factors on which an AI output is based [10]. Besides the algorithmic design and performance of the methods, a crucial consideration is how explanations are presented to end users. The effectiveness of the explanations should match people's expectations and needs, and be assessed in the context of human experience, approval, and comfort [15].

A common way of visualizing machine-generated explanations in text is by highlighting the input text using heatmap color palettes [25]. With the rise of neural networks, attention weights or gradient-based explainability methods have been used to highlight salient input text spans with darker background [5]. Qualitative studies in the legal domain have shown that such approaches are perceived as useful and enhance self-reported measures of trust [19]. It is not clear, however, how adding XAI features, such as highlighting, may affect the reading process of the user in a typical real-world workflow [12].

In this study, we investigate whether introducing XAI features, such as the commonly used option of text highlighting, may alter the reading process of end users. We build upon the work established by Norkute *et al.* for the task of summarizing a legal complaint [19], where it is crucial that machine guidance not interfere with professional conduct and accountability. We asked U.S. law students to go through the text of a legal complaint and find answers to legal questions related to the complaint. We tested the use of machine-generated summaries with and without text highlighting integrated as a form of XAI. We explored whether XAI in the form of text-highlighting changes where users focus their reading. We also tested whether XAI changes the process by which a user works through a complaint, that is the ordering of where they spend time in the document and when.

The remainder of this paper is structured as follows. Section 2 discusses previous work on the field of legal XAI with the use of text highlighting. Section 3 describes the study methodology, including the behavioral data collection and visualization we use to characterize the process of legal work. Section 4 presents and discusses the results. Section 5 presents the conclusions and future directions of this research.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595128

2 RELATED WORK

2.1 AI and Explainability in the Legal Domain

Recent advances in deep learning technologies in combination with the availability of vast amounts of legal data have accelerated AI applications in legal workflows [24]. Despite the increasing demand for automation in the legal domain, ML systems might still struggle to reach desired performance levels due to specific characteristics of legal work. Such characteristics include the elaborate and specialized nature of legal language, the high level of domain expertise required, the high frequency of exceptions, the low tolerance for risk and errors, as well as the expectation that system outputs include a method to establish accountability [23]. In areas where machine-powered decisions might seriously impact human lives, such as in the law, it is commonly regarded as essential that AI systems provide explanations for their outputs [9]. The motivation for approaches that are interpretable, explainable, and trustworthy is fueling a recent upsurge of activity in the field of XAI across various disciplines, including in legal applications [18].

In the legal domain, several studies have explored the value or the effects of explainability methods, for example in document review [6, 17, 26], legal decision prediction [4], case feature prediction [16], and legal document summarization [19]. It remains, however, unclear whether the introduction of XAI in a legal task affects user perception and behavior in additional ways [12].

2.2 Text-highlighting as means of AI explainability

A common way of assigning importance to the features that have most influenced a model's decision for a single instance is in the form of highlights of the original text input [25]. This is true for both ante-hoc explainability approaches, where the model is interpretable by design, and post-hoc explainability, which refers to methods that are designed to interpret black-box models, including deep learning models [22]. The most salient text spans stand out visually by means of modifying the intensity of the background color based on a heatmap color scheme defined by the saliency of each text span. The intensity values of the background color are determined by the explainability method used, for example, attention weights [2], integrated gradients, SHAP, LIME, etc. This type of visualization is increasingly common, despite a lack of behavioral information about how legal professionals respond to such approaches.

2.3 The effects of text-highlighting on reading strategies of legal text

Initial evidence on the effect of explainability via text background highlighting has not been consistent. Norkute *et al.* showed that introducing attention-weight text highlighting in a legal complaint accelerates processing speed for the task of evaluating an ML-generated summary of the facts of the complaint while it increases the trust of legal professionals for the correctness of the generated summary [19]. In a different setting, Branting *et al.* reported that text highlighting, based on the attention of an ML model that classified legal prediction, *did not demonstrably improve human decision speed or accuracy* [4]. Another study on legal classification compared the user's perception of the quality of highlighting generated by different explainability methods but did not examine how including the highlighting in the first place affected the user experience or work outcomes [26].

These previous studies examine the presence of text highlighting through the lens of evaluating a model's decisions, but they do not interrogate how XAI affects end users' reading process or comprehension. Manual text highlighting is a common strategy to enhance reading comprehension [3], and text highlighting, when used effectively, can aid reader's retention [7].

There are several important differences, however, between machine generated highlights for explainability versus user manual highlighting for content retention, the most prominent of which are 1) the potentially continuous nature of the machine-generated highlighting to indicate various degrees of saliency for different text spans, and 2) the higher probability of erroneous highlighted text spans in the case of false model predictions.

3 METHODOLOGY

3.1 Population of interest

We recruited participants from a total of nine United States (U.S.) law schools through snowball sampling. Data was collected from January to April of 2022. Ultimately 206 participants joined the experiment. The law schools were located in a variety of geographic regions across the U.S. and from both top and mid-ranked law schools. 93% of participants were pursuing a J.D. degree. Of the J.D. degree participants, 49% were 1Ls, 26% were 2Ls, and 21% were 3Ls.

The degree to which law students are good proxies for practicing legal professionals remains an open question [11]. The task we study, reviewing a complaint, is a task often left to law clerks and junior legal firm associates. We posit that law students should produce results with good external validity for understanding the behaviors of law clerks or junior law associates given that these positions are typically filled by recent graduates.

3.2 Online interface

We built an interface modeled after the design of Norkute *et al.* [19], see Figure 1. The interface displayed the text of a legal complaint (Figure 1A) and a few sentences long ML summary of the complaint (Figure 1B). Participants were asked to read the complaint text to answer questions about the complaint (Figure 1C).

Participants were randomly assigned to an experimental treatment. The two possible experimental conditions were as follows: 1) no machine-generated legal complaint text highlights added on the complaint text 2) machine-generated legal complaint text highlights added on the complaint text (Figure 1). The highlights and summary were generated using the same methods and data as previously described in [19]. The summary was always provided alongside the complaint text.

The legal documents we study are those filed by plaintiffs to initiate a civil lawsuit in various courts in the U.S.. The documents are usually between 10-100 double spaced pages of highly structured text. The case text presented to the participants in the study was processed by Optical Character Recognition (OCR), applied to the original PDF files.

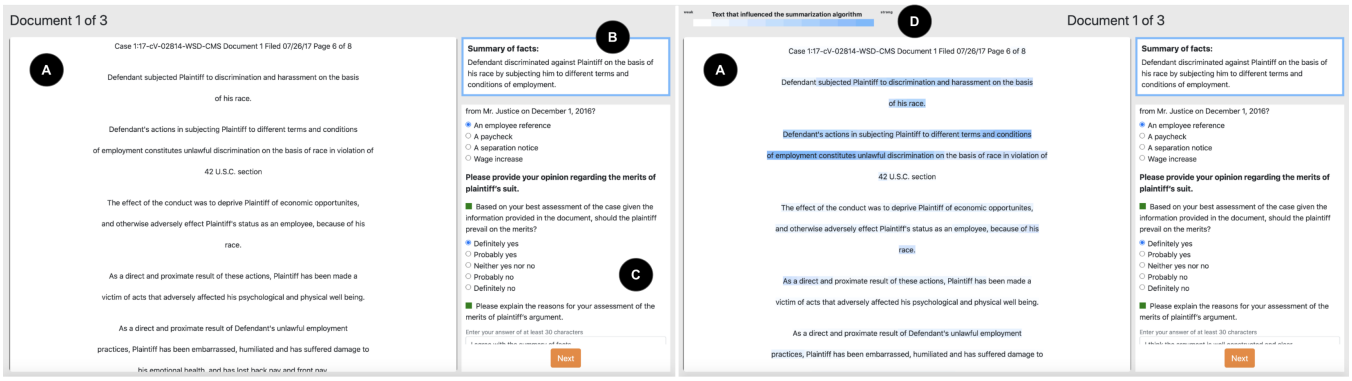

Figure 1: The experiment interface On the left: the Summary condition. On the right: Summary + Highlighting condition. A - Text display. B - ML-generated Summary C - textual questions D - highlighting legend, visible only in S + H treatment.

3.3 ML generated summary and highlights

The ML summaries were summaries of allegations made by the plaintiffs. They were automatically generated using a Pointer Generator network [20] as implemented in the OpenNMT-Py toolkit [13]. The model was trained from scratch on 800'000 court cases and associated editor written summaries. The ML generated summaries of all the complaints used in the present study were judged to be acceptable by editors who write these summaries at the partner organization that created the summarization model. The ML highlights were created by making use of the metadata produced by the Pointer Generator model. For more details on summary and highlights generation, see [19].

3.4 Procedure

Participants received an email invitation to participate in the study. They were asked to use their laptop or computer to do the task. They were randomly assigned on a between-subjects basis to one of the two conditions. The first screen of the interface obtained informed consent to take part in the study. Next, each participant was shown a tutorial that explained the interface elements and the task in more detail. Then, each participant read three legal complaints and answered questions about each complaint. The participants could not move on from one complaint to the next until all questions were answered. There was no time limit for the task. At the end of the task participants were asked demographic information questions and upon completion of the survey they received an Amazon voucher at a value of $20 USD.

The three legal complaints covered (in order presented) employment discrimination on the basis of race, fraud by one partner in a limited liability company, and recovery of attorney's fees for a previously litigated violation of the Fair Credit Reporting Act. To reduce the variation in the experimental results that would otherwise be attributable to document variation rather than experimental manipulation, the order of documents was always the same; the only difference between conditions was the presence or absence of ML highlights. We limited the number of documents to three due to the difficulty of securing law students in experimental studies and the consequent high costs of recruitment - asking students to read more documents would have resulted in a longer experiment time and more difficulty and expense in recruiting law students.

3.5 Data Collection

We recorded data about how participants behaved during the experiment as they read the legal complaint. Every second we recorded the uppermost and lowermost line of text visible to a participant. This method accounted for the possibility of different screen sizes or browser settings. With this data, we could then determine when and for how long any given text was visible on the screen. A schematic of the data recording is shown in Figure S1 in the Supplementary information.[1]

3.6 Visualization

Temporal allocation of user attention

We first consider the temporal allocation of attention using one-dimensional visualization (1d) to address whether the XAI feature changes how participants allocated their time within the document. For each token location in the document, we calculate for what portion of a participant's time that token location was visible. Previous work correlating scrolling to eye tracking [40] has shown that scrolling location gives an unbiased estimate of attention allocation, and so we take scrolling position to be an unbiased indicator of attention. Within each experimental treatment, we compile the total amount of time that a token location was visible across all users and then normalize this distribution,[2] creating an empirical probability distribution function (epdf) representing the attention probability of each token location. This visualization allows us to study which parts of the document received relatively more or less attention.

Spatio-temporal allocation of user attention

We visualize spatio-temporal allocation of user attention with a two-dimensional visualization (2d), plotting time versus current token position as users scrolled through the document. This visualization allows us to study the process by which participants interacted with the document. A 2d visualization of reading allows

[1] https://osf.io/t4s8p/

[2] The results reported below were robust to normalizing across an experiment group or to normalizing on a per-user basis.

us to distinguish participants who may have allocated the same overall proportion of time to the same section of a document but who did so in different orders.

We create a curve for each participant, in which the data points pairs of coordinates such that the first dimension gives the time that has passed and the second dimension gives the token location within the document, defined as the last visible token location. After the curve is computed, we normalize the temporal dimension so that the temporal maximum is equal to 1.[3] The spatio-temporal visualization gives insights into the reading process. Imagine a participant reads the document at a constant rate from the beginning to the end. That curve would be a straight line running from (0, 0) to (1, max token location). We finally create a 2d histogram by aggregating all data points from all user curves.

4 RESULTS AND DISCUSSION

4.1 Temporal allocation of user attention

We present the 1d results for Document 2 in Figure 2, with plots for Documents 1 and 3 available in the Supplementary Information Figure S3. In addition to the two epdfs, Figure 2 plots the ML attention scores used to generate the highlighting in the experimental interface (the ML Vals curve)[4]. For all Documents, the epdfs for the Summary and the Summary + Highlighting condition are visually quite similar; further, the epdfs are not strongly correlated with the ML Highlight Values.

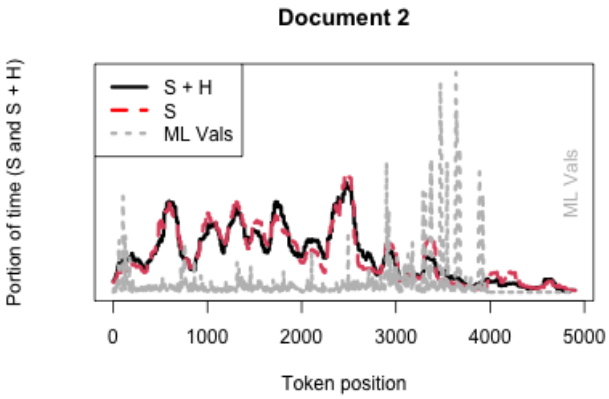

Figure 2: Epdfs for Summary (S) and Summary + Highlighting (S + H) are similar; both are dissimilar to the ML Vals.

To test our intuition that the Summary condition produced epdfs indistinguishable from those of the Summary + Highlighting condition, we apply a Kolmogorov-Smirnov distribution test, a nonparametric goodness-of-fit test [14, 21], to bootstrapped samples from each of the epdfs. There is no statistically significant difference between the epdfs for the Summary condition and the Summary + Highlighting condition (Document 1: D = .04, p = .3; Document 2: D = .04, p = .5; Document 3: D = .04, p = .4).

The highlighting did not influence where participants devoted more or less attention within the document. If the allocation of attention had been changed in the presence of highlighting, we would need to understand whether and how such a change posed challenges to using the ML tool for the responsible practice of law. The fact that temporal attention allocation does not change in this experiment suggests that machine assistance need not create behavioral concerns as to the possibility of legal professionals abdicating professional judgment, even when those professionals receive access to XAI features. This result is surprising given [19]'s prior results with the same tool, showing that those with highlighting worked faster than those without. The equivalence of the process-oriented 1d epdfs suggests that the speedup previously identified likely does not come from a qualitatively different pattern of attention allocation.

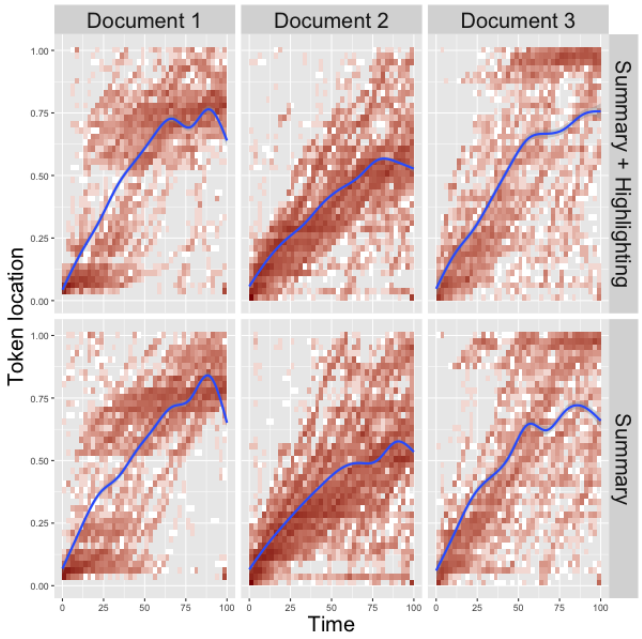

Figure 3: Spatio-temporal attention distributions. The x-axis is time (normalized to 100% per participant); the y-axis is token location. The solid lines represent loess fits. For Document 2, Highlighting appears to reduce the variation of the distribution relative to a loess or linear fit.

4.2 Spatio-temporal allocation of user attention

We next consider the spatio-temporal distribution of the reading process. As shown in Figure 3, the spatio-temporal distribution looks different for each document; the reading process was distinct for each document. The differences, if any, between Summary and Summary + Highlighting within each document are more difficult to characterize.

[3]Each user produced a time series of uncertain and typically different length; the results reported here are robust to including all data points with this normalization or, in the alternative, downsampling the data to equalize data point contribution as between participants.

[4]We used the square root transformation of the attention score to generate the highlighting coloring for the text. ML Highlight Values are not directly comparable to the epdfs.

The spatio-temporal distribution in Document 2 looks distinctly linear, and the distribution looks less variable in the case of Summary + Highlighting than in the case of Summary. We fit a least squares linear regression to each spatio-temporal distribution for Document 2 and find that the mean absolute value of the residuals is significantly lower in the case of Summary + Highlighting than in the case of Summary (Wilcoxon rank sum test W = 3e7, p < .0001). This result holds in leave one out cross-validation, so the difference in the magnitude of the residuals is not the outcome of errant scrolling by a lone participant.

We do not propose a linear regression fit for the distributions for Document 1 and Document 3, as a visual inspection makes clear that a linear fit is a poor choice for those cases. Figure 3 suggests that the contents of the document are likely more influential in shaping the spatio-temporal distribution than the absence or presence of highlights. Nonetheless, we observe an opportunity to distinguish the effects of XAI in Document 2, where the presence of the highlighting likely made end users read in a more uniform fashion than with just a summary. Future work with larger samples may enable more definitive conclusions about the procedural impact of XAI, but this current result already shows the importance of assessing the procedural impact of XAI in legal use cases.

5 CONCLUSIONS

To date, most experimental studies of XAI have identified subjective changes in confidence or trust when users work with a ML tool that includes XAI features. The few studies where outcomes in legal tasks were studied have yielded mixed results as to the performance impact of XAI. We propose a different approach: measuring the process of work. We study 1d and a 2d visualizations of attention allocation, each approach yielding different insights. The 1d epdfs reveal that XAI does not change the allocation of overall proportions of time. The 2d spatio-temporal distributions show that document-specific effects are likely stronger than XAI effects on reading process; there is, however, some possibility that highlighting changes the spatio-temporal process distribution.

Our findings reflect one of the first reported studies of procedural change due to XAI. We identify a pattern of reduced variance in the spatio-temporal distribution in one document. We do not take a position as to whether the reduction in variance is normatively problematic; however, this possibility of behavioral changes requires further attention in experimental work to assess potential prevalence. The possibility also requires further attention from legal practice ethicists to better understand whether variance-reducing XAI features are acceptable or even desirable. Future empirical and theoretical work on legal XAI should account for the importance of legal process.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations* (9 2014).

[3] Dennis Baron. 2009. *A Better Pencil: Readers, Writers, and the Digital Revolution.* Oxford University Press.

[4] K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. 2019. Semi-Supervised Methods for Explainable Legal Prediction. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 22–31. https://doi.org/10.1145/3322640.3326723

[5] Adrian M. P. Brasoveanu and Razvan Andonie. 2020. Visualizing Transformers for NLP: A Brief Survey. *2020 24th International Conference Information Visualisation (IV)*, 270–279. https://doi.org/10.1109/IV51561.2020.00051

[6] Rishi Chhatwal, Peter Gronvall, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2018. Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding. *2018 IEEE International Conference on Big Data (Big Data)*, 1905–1911. https://doi.org/10.1109/BigData.2018.8622073

[7] Nidhi Roy Choudhury and Nirali Bhansali. 2022. Highlighting And the Effects of The Color of The Highlighter on Retention: A Review of Literature. *The International Journal of Indian Psychlogy* 10 (3 2022). Issue 1.

[8] Mariano-Florentino Cuéllar. 2019. A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness. *Columbia Law Review* 119 (11 2019), 1773–1792. Issue 7.

[9] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38 (10 2017), 50–57. Issue 3. https://doi.org/10.1609/aimag.v38i3.2741

[10] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. 2022. Explainable AI Methods - A Brief Overview. , 13-38 pages. https://doi.org/10.1007/978-3-031-04083-2_2

[11] Dan M Kahan, David Hoffman, Danieli Evans, Neal Devins, Eugene Lucci, and Katherine Cheng. 2016. "Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment. *William and Mary Law School Scholarship Repository* (2016). https://scholarship.law.wm.edu/facpubs/1801

[12] Serhiy Kandul, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, François Fleuret, and Markus Christen. 2023. Explainable AI: A Review of the Empirical Literature. *SSRN Electronic Journal* (2023). https://doi.org/10.2139/ssrn.4325219

[13] Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. (5 2018). http://arxiv.org/abs/1805.11462

[14] A Kolmogorov. 1933. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.* 4 (1933), 83–91.

[15] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. (10 2021). arXiv:2110.10790

[16] Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2022. Evaluating Explanation Correctness in Legal Decision Making. *Proceedings of the Canadian Conference on Artificial Intelligence* (5 2022). https://doi.org/10.21428/594757db.8718dc8b

[17] Christian J. Mahoney, Jianping Zhang, Nathaniel Huber-Fliflet, Peter Gronvall, and Haozhen Zhao. 2019. A Framework for Explainable Text Classification in Legal Document Review. (12 2019).

[18] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. (2 2019).

[19] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3411763.3443441

[20] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* 1 (4 2017), 1073–1083. http://arxiv.org/abs/1704.04368

[21] N Smirnov. 1948. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* 19 (1948), 279–281. Issue 2. https://doi.org/10.1214/aoms/1177730256

[22] Ajay Thampi. 2022. *Interpretable AI: Building explainable machine learning systems* (1 ed.). Manning.

[23] Rob van der Meulen. 2022. Gartner Predicts That "Human-in-the-Loop" Solutions Will Comprise 30% of New Legal Tech Automation Offerings by 2025. *Gartner* (2 2022).

[24] Serena Villata, Michal Araszkiewicz, Kevin Ashley, Trevor Bench-Capon, L. Karl Branting, Jack G. Conrad, and Adam Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law* 30 (12 2022), 561–591. Issue 4. https://doi.org/10.1007/s10506-022-09327-6

[25] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. https://doi.org/10.18653/v1/N16-1174

[26] Łukasz Górski and Shashishekar Ramakrishna. 2021. Explainable artificial intelligence, lawyer's perspective. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 60–68. https://doi.org/10.1145/3462757.3466145