Canadian AI 2022
30 May–3 June 2022, Toronto

Evaluating Explanation Correctness in Legal Decision Making

Chu Fei Luo[†,‡,◇,*], Rohan Bhambhoria[†,‡,◇,], Samuel Dahan[†,◇,§], Xiaodan Zhu[†,‡]
[†] Department of Electrical and Computer Engineering & [‡] Ingenuity Labs, Queen's University
[◇] Conflict Analytics Lab, Queen's University
[§] Cornell Law School

Abstract

As machine learning models are being extensively deployed across many applications, concerns are rising with regard to their trustability. Explainable models have become an important topic of interest for high-stakes decision making, but their evaluation in the legal domain still remains seriously understudied; existing work does not have thorough feedback from subject matter experts to inform their evaluation. Our work here aims to quantify the faithfulness and plausibility of explainable AI methods over several legal tasks, using computational evaluation and user studies directly involving lawyers. The computational evaluation is for measuring *faithfulness*, how close the explanation is to the model's true reasoning, while the user studies are measuring *plausibility*, how reasonable is the explanation to a subject matter expert. The general goal of this evaluation is to find a more accurate indication of whether or not machine learning methods are able to adequately satisfy legal requirements.

Keywords: Machine Learning, Explainable AI (XAI), High-stakes decision making, Legal AI

1. Introduction

Explainable AI (XAI) is an important area of study that seeks to explain or interpret machine learning systems [1]. The ability to explain a decision is paramount to building trust in an AI system, especially in high-stakes domains. For example, decision making and evaluation tasks in law often have significant economic and social implications [2]. Along with ensuring an XAI method is faithful to the model itself, it is important to involve the end users to verify that their requirements are being met. Lawyers expect any tool they use in their profession to show "fairness, lack of bias, transparency in the decision process, and consequence awareness" [3]. These principles applied to AI can help detect biased legal decisions and ensure machine learning models are providing more help to society than harm [2]. However, the latest machine learning models are generally trading increased performance for higher complexity, thereby losing their inherent interpretability. XAI methods have been proposed to bridge this gap, but their evaluations are geared towards general machine learning tasks, and are not thoroughly evaluated by subject matter experts. This raises serious concerns on their reliability in high stakes scenarios.

It is therefore important to validate the feasibility of general XAI methods on specific domains of interest. There are several industries that emphasize the need for explainability, including medicine, law, and finance [4]. The medical domain has the most work published on XAI, with a variety of evaluations from statistical analyses to user studies [5] [6]. User studies are essential to evaluating XAI methods, as feedback from the main stakeholders are integral to a well-engineered solution. However, there has only been one such study conducted with lawyers [3]. Considering the economic and societal impacts of legal tasks, XAI methods are still seriously understudied in this domain.

The following work aims to evaluate explanation correctness of machine learning models applied to legal data and, by extension, the reliability of machine learning systems in law.

[*] 14cfl@queensu.ca

Specifically, given a legal task with a curated dataset and a model trained on that data, we seek to evaluate existing post-hoc explanation methods with the help of subject matter experts. All the datasets are tabular — compared to text or image data, which also requires the model to abstract information, tabular annotation data has a higher proportion of facts already deemed important by a lawyer. Removing the layer of complexity presented by natural language understanding or image processing brings this study closer to examining a model's *learned reasoning*. For AI systems that report a high accuracy, it is important to understand if the system's learned space is intuitive — i.e. *plausible* — to human experts. This involves validating whether the explanations are true to the aforementioned learned space, as well as assessing the generated explanations with experienced lawyers.

In summary, the contributions of this paper are as follows:

- The first explanation correctness study on tabular data in high-stakes decision making. Although there is existing work on image and text, tabular data isolates the factors that subject matter experts deemed important. This study brings a more informative evaluation of explanation correctness when both the model and subject matter expert have a definitive shared understanding of relevant facts.
- This evaluation is performed over three legal tasks from different areas of law. Using datasets curated by legal subject matter experts, we compare the faithfulness of two post-hoc, model-agnostic XAI algorithms against a model-specific explanation method.
- A user study was conducted with subject matter experts to evaluate plausibility. They have expertise in the specific tasks of interest, so their feedback is highly valuable to evaluating plausibility. From this study, we propose new methods for quantifying both plausibility and subjectivity.

2. Related Work

2.1. Explainable AI

One of the main goals of XAI is providing *explainability*. In addition to interpreting the results of an AI system, XAI methods need to accurately explain the model's decision making process in a way that humans can understand [7].[1] There are many overarching principles to a high-quality explanation, including correctness, robustness, and simulatability [6]. Within each principle, there are many desirable sub-properties. Existing work [8] defines an explanation's correctness as two components: **faithfulness**, also known as accountability, and **plausibility**, also known as persuasiveness or trustworthiness. Faithfulness is how accurately the explanation reflects the model's internal decision making process, while plausibility is how reasonable the AI's explanation appears to a human. It is possible for an explanation to seem logical to a human while not truly reflecting the model's process, and vice versa, so both must be evaluated concurrently [8]. Although there are many methods of evaluating faithfulness, a popular one is conducting ablation in order of the estimated feature importance, and measure the impact on performance (eg. a decrease in accuracy or change in polarity) [6] [9]. There are many variations — for example, it is also possible to evaluate faithfulness with synthetic data and known feature importance [10].

Faithfulness cannot be measured against human labels by definition [8], but plausibility and other principles of explanation quality are best evaluated through user study. Several interesting user studies have been conducted to evaluate model simulatability using general NLP tasks [11] [12]. In contrast, this work's user study evaluates *explanation correctness*, and focuses on *legal issues*, which requires the help of subject matter experts.

[1] Explanability and Interpretability are interchangably used in some contexts. In this paper, they are different and we follow the definition of explainability [7].

Pursuits into explainability algorithms roughly follow two major categories of approaches: *ante-hoc* and *post-hoc*. Within these approaches, there are two further types of explanations based on their goals: **global explanations** that measure importance over the entire dataset, and **local explanations** that explains specific data samples in terms of their input.

- **Ante-hoc** — Also known as self-explaining methods, ante-hoc implies that there is inherent explainability in the model's architecture, often incorporated by predicting human-understandable concepts as an intermediate step [13]. Models with ante-hoc explainability are considered inherently interpretable and generally desirable [2]. However, they need to be purposefully designed into the model. Methods that initially seemed inherently ante-hoc, such as attention mechanisms, have been found to have issues with faithfulness [9].
- **Post-hoc** — These are methods that explain a model's decisions after training. A typical approach is building a secondary model to simplify or project the learned space of the original, such as deconvolutional networks in computer vision [14].

This paper focuses on **post-hoc, local explanations** — specifically, post-hoc feature attribution methods. Examples include SHapley Additive exPlanation (SHAP) [15], Integrated Gradients [16], and Local Interpretable Model-agnostic Explanations (LIME) [17], among others. Feature attribution methods are desirable because they can produce explanations for specific samples [18]. This allows for more fine-grained understanding compared to global importance or overall trends in the data, which fulfills the expectation of transparency in the decision process for lawyers [3]. Many of these methods are also model-agnostic, i.e. they can be extended to any underlying architecture, so they are more generalizable compared to ante-hoc algorithms. There are drawbacks to post-hoc methods — LIME and SHAP in particular have been shown to be vulnerable to adversarial attacks [19] [2]. However, because of their ease of implementation, post-hoc methods are ideal starting points to quickly adopt explainability into existing technology. Global explanations serve a different purpose in XAI — for example, they are likely preferred for text or image data, where high levels of abstraction might make feature attribution less meaningful. In this study, since the data is tabular, post-hoc methods were chosen for our purposes.

2.2. **Explainability in High-Stakes Industries**

Other high-stakes domains have also stated a need for explainability, with healthcare having the most extensive publications and rigorous evaluations [20] [5]. The financial domain is another important industry for XAI along with medicine and law, with many works noting the pertinence of explainability in their AI applications [4]. To our best knowledge, however, there is no public work on XAI methods for financial data involving subject matter experts — the ones found only perform statistical evaluations [21]. The closest known work to this paper explores explanation correctness with subject matter experts on a medical imaging task [6]. However, their work was on image data instead of tabular. As mentioned in Section 1, images have additional complexity in abstracting facts compared to tabular data, so our study is unique in its findings and challenges.

2.3. **Legal AI**

In our domain of interest, law, the application of machine learning technologies is commonly referred to as *Legal AI*. Though the field is broad, there are two major efforts as determined by dataset curation method: manually annotating datasets, and parsing legal documents automatically with the help of NLP [22]. Likely following the NLP trend, the only known user study published was on a legal text classification task [3]. For text data,

explainability goes beyond feature attribution — lawyers in that study, for example, wanted the model to provide prior case law and citations. This is the major deep learning trend in legal AI, and many works are pursuing ante-hoc explainability designs to that goal [23] [24]. For post-hoc method evaluations, law is similar to the financial domain — most existing publications only include a statistical analysis without user study [25]. Work involving subject matter experts is still sparse, likely because lawyers are more expensive to survey compared to annotators in general machine learning tasks. However, it is still important to evaluate explainability, especially conventional algorithms that are widely adopted in practical settings, because of the potential economic and societal impacts.

3. Evaluating Explanation Correctness in Legal Data

This work's main contribution is a study of multiple XAI methods as applied to legal data. Due to the high stakes nature of the legal industry, professionals require adequate explainability to trust the systems they use. For any machine learning model to be reliable, it is important to validate the learned logic with a lawyer, which consequently brings greater confidence into their implementation and adoption. The definition of explanation correctness used follows previous work [8], which clearly differentiates between *faithfulness* and *plausibility*, both of which are fundamental to this study. This paper's overarching framework also borrows from existing methods [6], with two components: *statistical analysis* for faithfulness and a *user study* for plausibility. For a comprehensive analysis, this work evaluates multiple datasets over several areas of law. Each task has different challenges and requirements from a legal perspective, to provide a more comprehensive evaluation of the XAI methods. First, explanations are generated for all samples in the test set and those are statistically assessed to determine faithfulness. A few samples were organized into a user study, where lawyers validated how closely the explanation aligns with their understanding, and metrics are extracted to create a quantitative measure of plausibility.

3.1. Our Approach

This study primarily evaluates post-hoc feature attribution methods. As previously mentioned, these algorithms seek to explain the output in terms of the significance of each input feature, which allows for feedback on individual samples through a user study. These two methods also have well-formulated open source packages, which make them more likely to be implemented in the industry. Each method was fit on the train set of the respective dataset and evaluated on the test set.

- **LIME** — An early model-agnostic method that randomly samples perturbations around the input to create a linear approximation of the local neighbourhood [17]. The learned weights of each feature in the linear approximation is the assigned importance.
- **SHAP** — An algorithm using Shapley values from cooperative game theory. Viewing high and low probability as two competing entities, the method adds and removes input features to measure the impact of each component on the model outcome [15]. This work specifically uses TreeExplainer [26].

Faithfulness will be evaluated with the **diffAUC** metric used in previous work [6]. DiffAUC is computed as follows: for the list of feature importances produced by an XAI method, features are iteratively removed from the input from most to least important. The model is retrained and the F1 score is obtained after each removal. The f1 scores are plotted over the features against a random ablation baseline. Then, the difference in AUC between the methodical and random ablation is scaled by the number of features to produce a score

in the range [-1, 1]. The more informative a feature importance ranking is, the faster the F1 score is expected to decrease, so a lower score is better.

This metric is chosen as it uses global feature importance, which accumulates explanations over all available data points. For SHAP and LIME, the feature importance of one feature, $feat_i$ is defined in Equation 3.1.

$$feat_i = \frac{\sum_{n=0}^{N} |w_i n|}{I} \tag{3.1}$$

N is the number of samples in the test set, $w_i n$ is the importance of $feat_i$ in the nth sample as calculated by the XAI method, and I is the total number of features.

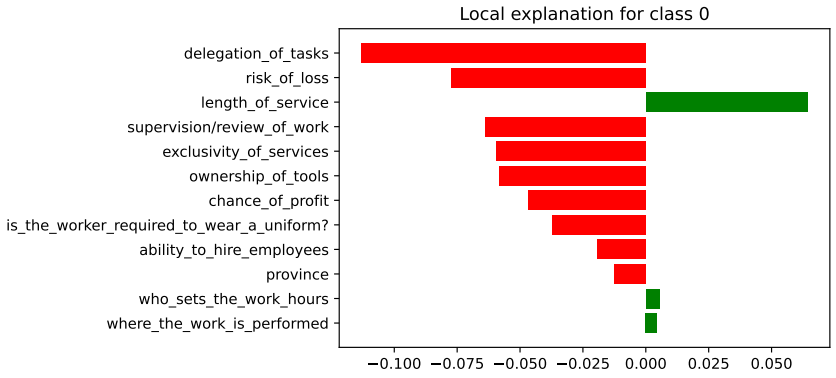
Plausibility was evaluated with lawyers via an online survey. The study proceeded with 11 lawyers over the three tasks, with each lawyer receiving 5 questions per XAI method, for 10 total. Some results from Section 4.2 indicate the length might have been too short, and this should be revisited in future work. The samples were randomly chosen from the test set for variety, although we generally kept a balance between positive and negative predictions. Each lawyer only evaluated tasks where they had expertise.

For each sample, the lawyer was first shown the input and asked to *make their own prediction* on the sample's outcome. This allows them to read through the facts and build their own reasoning, as well as give a measure of subjectivity. They were not told the sample's true outcome. Then, they were shown the model's prediction and accompanying explanation, and asked to rate how closely these aligned with their understanding of the case. Every participant received verbal instructions on how to interpret the explainability graphs outputted by each method before beginning the survey. A rating of 1 indicated they disagreed with the model's prediction and explanation, while a rating of 5 meant they agreed with both completely. After rating, the participants were asked to give written feedback on the explanation — for example, to list the feature(s) they disagreed with and why. Requesting comments gave insight to the participants' thought processes, as well as how their reasoning aligns with one another. The questions alternated between the LIME and SHAP outputs to limit bias towards one or the other. From the survey, there are three main pieces of information: the expert's **predicted outcome**, **rating of the explanation** the model produced, and their **written feedback**. The inter-rater agreement for all of these pieces was evaluated with Fleiss' kappa and Krippendorff's alpha. For the written feedback, the features criticized are manually extracted and treated as options in a multi-selection question, and only Krippendorff's alpha is calculated with MASI distance, since Fleiss' kappa cannot be used with missing values. These agreements are another measure of subjectivity - if the agreement is low, the task is more likely to be subjective to lawyers.

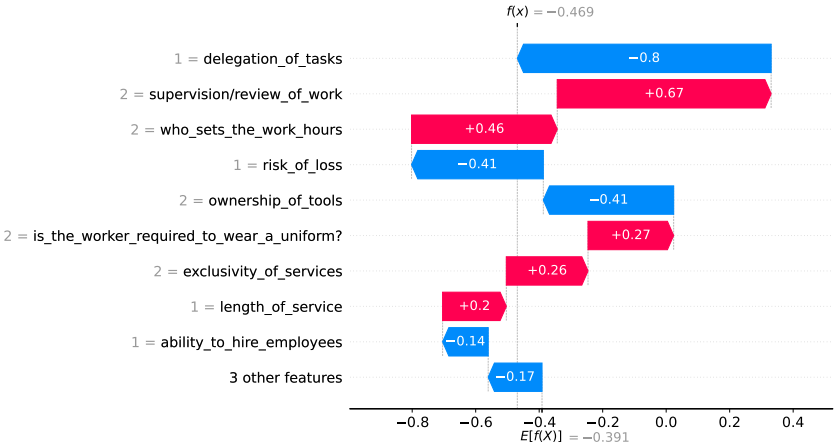
Finally, the plausibility is quantified as the statistical correlation between absolute error (distance of model prediction from the true label) and the average explanation rating to see if the quality of explanation relates to the model's certainty. Intuitively, an explanation should be **less plausible** as the model is more incorrect. For a correlation in the range [-1, 1], the perfect plausibility score is therefore -1. This metric p_{rate} is defined in Equation 3.2.

$$p_{rate} = corr(\sum_{i=1}^{N} |y_i - f(x_i)|, \mu_{si}) \tag{3.2}$$

where N is the number of samples (in this survey, $N = 5$ for each XAI method), $f(x_i)$ is the model's prediction for $x_i \in [x_1, x_2, ..., x_N]$, μ_{si} is the mean rating, and y_i is the true label.



(a) WC - 1



(b) WC - 2

Figure 1. Example explanation graphs shown to the annotators, corresponding to comments in Table 6

4. Experiments

4.1. Experiment Setup

Datasets. This paper evaluates legal tasks from three different areas of law: personal injury, intellectual property, and employment. There are three corresponding sets of historical legal decisions, either court proceedings or internal cases, and all relevant information pertaining to decision was annotated by subject matter experts into a tabular format. In this study, the lawyers who annotated this data also volunteered for the survey, so they are verified to have expertise in these specific tasks. The tasks are as follows:

- **Personal Injury Negotiation (PIN)** — For a personal injury dispute, predict whether the total negotiated settlement amount will be above or below the median average. Although the target is a continuous monetary value, the task was initially defined as classification at the clients' request.[2]
- **Trademark Confusion (TC)** — For a trademarks dispute case, predict whether there is confusion or no confusion between two pieces of intellectual property (IP) [27]. Lawyers chose five measures of similarity as the most important features [23].

[2]This dataset is not available for distribution due to its sensitive nature.

Table 1. The dataset composition, including the number of features available for the task, total number of cases, and distribution between positive and negative samples

Dataset	# Features	Positive/Negative Label	Positive Samples	Negative Samples	Total
PIN	37	Above Average/Below Average	325	311	636
TC	5	Confusion/No Confusion	331	126	456
WC	12	Indep. Contractor/Employee	171	240	411

- **Worker Classification (WC)** — For an employee contracts dispute, predict whether a worker hired by a company is an independent contractor or an employee.[3]

The distribution of different classes can be found in Table 1. All tasks are binary classification, and the label of interest is referred to as the positive class. PIN is the most balanced by design of the task, while the majority class of TC and WC comprise 78% and 58% of their datasets respectively. Personal Injury Negotiation (PIN) is the only dataset made of internal dispute claims, and are dependent on the firm's internal decision making processes alongside legislature or case law. Consequently, the data is not as consistent as those from court trials, which only use prior case law. The PIN dataset is being re-annotated at the time of writing. Also, a feature in WC ("Length of Service") was included in the model and survey but later noted to be unimportant to the decision by the annotators, so this dataset's scores might be biased due to this spurious feature rather than poor plausibility. The results are reported as is, and we discuss how bias and subjectivity can be identified from the study in Section 4.2.

Model Settings. The model used for the survey is an XGBoost gradient-boosted decision tree trained with binary cross-entropy loss [28]. For tabular data, XGBoost models achieve comparable or improved performance compared to deep learning for lower computational cost [29]. The XGBoost library includes a gain-based feature importance measure, but it is a global explanation and does not provide feature importance for individual samples, so there is still merit to adopting LIME or SHAP.

For each dataset, the data is divided into a 80:20 train:test split with stratified sampling to preserve data balance. The tree is trained for 10 boosting rounds with a maximum depth of 2, except for TC which used a maximum depth of 3 and 20 boosting rounds. These settings were empirically shown to give the best performance, but the difference is most likely due to TC having significantly fewer features.

The model performance on each dataset is summarized in Table 2. The F1 score has a high variation between datasets, which is to be expected considering the differences in data balance and consistency. Figure 2 shows the histograms of prediction probabilities for the various test sets. While this is not a perfect indication of sample difficulty, predicted probability is shown to demonstrate model confidence, which in turn might affect the survey and plausibility scores. In TC and WC, for example, most predictions are between [0,0.2] and [0.8,1.0], which shows that the model was more confident in its decisions compared to PIN. However, the graph outputs on TC is highly skewed to the positive class — likely because of the unbalanced classes preserved by stratified sampling.

4.2. **Results and Discussion**

The average diffAUC scores were taken over 25 trials to account for variations in the random ablation baseline, and summarized in Table 3. Overall, each explainability method improves over the random baseline, performing better than what was observed with image

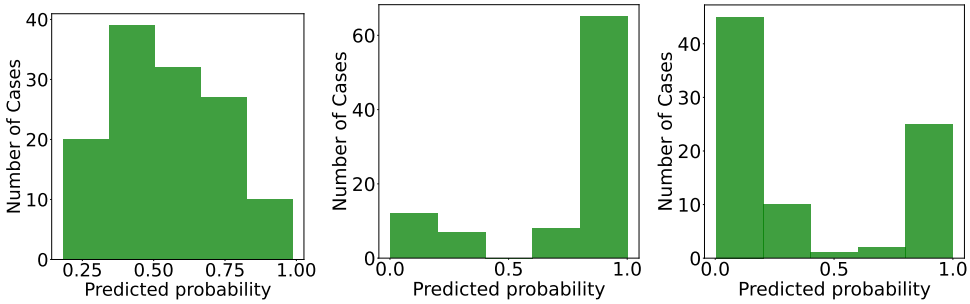
[3]This dataset will be released separately in the near future.

Table 2. The model's performance results over all 3 datasets, reporting precision, recall, accuracy, and F1 score.

Dataset	Precision	Recall	Accuracy	F1 Score
PIN	0.703	0.800	0.726	0.748
TC	0.876	0.955	0.870	0.914
WC	0.882	0.851	0.867	0.859

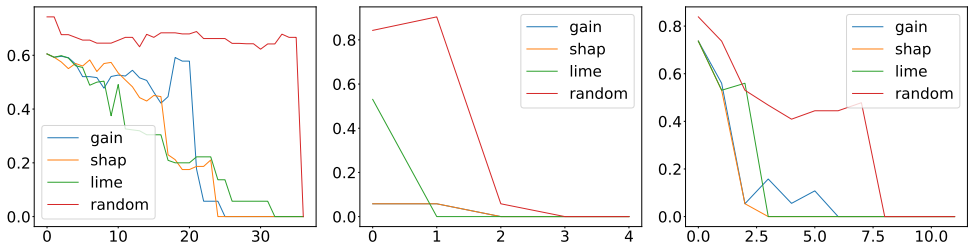
Table 3. DiffAUC metric summarized from the 3 datasets and the 3 explanation methods.

Dataset	# Features	Gain	LIME	SHAP
PIN	37	-0.205 ± 0.08	-0.247 ± 0.09	-0.236 ± 0.08
TC	5	-0.189 ± 0.13	-0.145 ± 0.13	-0.183 ± 0.13
WC	12	-0.297 ± 0.13	-0.284 ± 0.13	-0.326 ± 0.13



(a) Personal Injury Negotiation *(b)* Trademark Confusion *(c)* Worker Classification

Figure 2. Prediction probabilities of the model over the test set.



(a) PIN *(b)* TC *(c)* WC

Figure 3. Sample plots of F1 score over 1 trial, comparing the different feature importances to the random baseline. For TC, SHAP and GAIN are overlapped.

data [6]. However, their performance varies on different datasets. SHAP performs best on WC, while LIME scores highest for PIN, and Gain on TC. It seems like Gain's feature importance is most accurate on TC, but inconsistencies start to show with higher dimensionality data — the F1 score plot of Gain fluctuates significantly for PIN, returning to the original F1 score with half of the features removed. There are also minor fluctuations in SHAP and LIME, although they are visually smoother. Similar to findings from existing work [10], the faithfulness of these XAI methods are inconsistent on decision tree models.

The inter-rater agreements for the survey questions are reported in Table 4. The PIN lawyers had the highest agreement for explanation rating, but lowest when predicting the

Table 4. Inter-rater agreements over all multiple choice questions in the survey, measured with Krippendorff's alpha and Fleiss' kappa. For explanation ratings, the mean rating is also reported with the agreements.

Outcome Prediction			
Dataset	**Krippendorf's Alpha**	**Fleiss' Kappa**	**Mean Rating**
PIN	0.18	0.16	-
TC	0.35	0.33	-
WC	1	1	-
Explanation Rating			
Dataset	**Krippendorf's Alpha**	**Fleiss' Kappa**	**Mean Rating**
PIN	0.24	0.22	4.2
TC	0	-0.04	3.37
WC	0.02	0	4.28

Table 5. Inter-rater agreements for features mentioned in the comments. The mean # of features indicates the number of features criticised per comment, per rater.

Dataset	**Krippendorf's Alpha**	**Mean # Features**
PIN	0.03	0.5
TC	0.28	0.8
WC	0.26	1.03

outcome. This might be due to the subjectivity of the task — while there is a higher tolerance for what is considered a plausible explanation, it is difficult to predict the outcome. With TC, there was marked disagreement in explanation rating, but relatively strong agreement when predicting the outcome. Finally, WC achieved perfect agreement on predicting the outcome, but explanation rating had similar agreement values to TC.

This behaviour can be explained with the written feedback. Two comments from each survey are provided in Table 6, with specific features being criticised highlighted in the text. Upon closer inspection, disagreement in explanation rating does not necessarily reflect subjectivity. For example, in WC-1, all the lawyers mention the "Length of service" feature, but this sample had disagreement in the explanation rating. This was likely caused by differences in understanding the rating system rather than true disagreements. As another example, WC-2 in Table 6 is interesting, because the model made a mistake on this sample. However, the experts rate the agreement strongly, and one person noted that the explanation was "pretty accurate." Either the annotators themselves were overly confident in the model, or the sample was an outlier in the dataset.

To distinguish between the two possibilities, the inter-rater agreement on the features criticized in their written feedback is considered as another measure of subjectivity, and reported in Table 5. With this new metric, WC-2 scores low since all lawyers commented on different features. PIN, which was mentioned to be more subjective, now scores the lowest, while WC and TC score similarly. This measure is more intuitive to defining subjectivity when the annotators criticize an explanation to the same degree but for different reasons. However, this metric discards more general comments, like those for TC-2 in Table 6.

The plausibility metric calculations are included in Table 7. Over all the legal tasks, **SHAP has better plausibility** than LIME. PIN has the most pronounced difference between SHAP and LIME, although the reason is not clear, and TC had the least. TC also has the lowest correlation overall, while WC has the highest. This is likely due to cognitive bias, considering their agreements in predicting the outcome, or the higher probability confidence in the model predictions. Aside from one false negative, the absolute error in the other samples were low, which suggests there was not enough variance for a proper

Table 6. Examples of survey responses across the various tasks. The rating is the average rating across the annotators for that sample's explanation quality.

Example #	Rating	Comments
PIN - 1	3.25	Surprised **age** wasn't a factor Would have assumed that **longer-lasting injuries and psychological diagnoses** would contribute more to the prediction probability than shown in the model. Thought **occupational status** would increase the value.
PIN - 2	4.25	does **income** lower value? Matches my judgement of the case! **Long standing** was surprising as to how much it affected the prediction
TC - 1	3.0	**Conceptual similarity** carried more weight than expected. Surprising that it is confusion given the lack of similarity. Also second graph is counterintuitive. Lack of **visual and conceptual similarity** are the same but point in different directions? The **conceptual similarity** looks really high although it's the same value as the visual and phonetic similarities (2).
TC - 2	4.0	**Visual similarity** seems higher than reflected in the first chart Low similarities = unlikely that there is confusion This one could also go either way.
WC - 1	4.0	**Length of service** was not a big determining factor. Employee could work for a short period of time. I don't think **length of service** should be contributing strongly either way This might be a reoccurring theme, but I am still not aware that **length of service** is that determinative of a factor (I could be wrong though). However, everything else is pretty spot on.
WC - 2	4.75	I agree that **"hirer setting work hours but not when work is done"** is more in the red direction than sample 1 **"hirer only for setting work hours"**. But I would imagine that hirer setting work hours would be in the blue more and if worker only it would be more in the red? The only factors that don't contribute to the person being considered an employee are the fact that they have **no supervision**, and **don't have to wear a uniform**. The rest should all contribute to employee This one is pretty accurate. Although, the only thing I wanted to note is that **"who sets the work hours"**, might be a bit too biased towards contractors

Table 7. The calculated p_{rate} values. Correlations are reported over all 10 questions, then for the 5 individual samples of each explanation method.

Dataset	$p_{rate,lime}$	$p_{rate,shap}$	$p_{rate,overall}$
PIN	0.38	**-0.82**	0.07
TC	-0.65	**-0.89**	-0.72
WC	0.47	**0.029**	0.40

correlation. This survey would benefit from a second run with more samples and a simpler format, eg. only rating explanations, in order to get a more stable measure of plausibility.

5. Conclusion

Explainable AI is an important consideration when implementing AI systems in legal tasks. The ability to explain a decision leads to higher trust in the AI system, which in turn brings about further adoption of the technology overall. This work evaluated explanation correctness of XAI methods on three legal tasks from various areas of law. We evaluated

two post-hoc explanation methods, LIME and SHAP, with a user study involving subject matter experts and quantitative analysis. Similar to existing work, it was concluded that explanation faithfulness can vary depending on the data, but all are significantly more informative than a random baseline. Different areas of law can vary in subjectivity, so this work also proposed quantitative measures of subjectivity and bias. From the user study, we obtained measurements of plausibility with a newly proposed metric, and concluded that SHAP is consistently more plausible than LIME regardless of subjectivity. However, due to the inconsistent faithfulness results, it is recommended to exercise caution with these methods. We hope that this work brings attention to the importance of evaluating explainable AI for legal data, and will lead to further exploration in the future.

Acknowledgements

This research is supported by NSERC Discovery Grants and DAS supplement, as well as the MITACS Accelerate program. The datasets and user study participants were provided by the Queen's University Conflict Analytics Lab. We thank the three anonymous reviewers for their careful feedback on this work.

References

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and L. Kagal. "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning". In: *CoRR* abs/1806.00069 (2018). arXiv: 1806.00069. URL: http://arxiv.org/abs/1806.00069.

[2] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[3] Ł. Górski and S. Ramakrishna. "Explainable artificial intelligence, lawyer's perspective". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 60–68.

[4] A. Adadi and M. Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[5] E. Tjoa and C. Guan. "A survey on explainable artificial intelligence (xai): Toward medical xai". In: *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[6] W. Jin, X. Li, and G. Hamarneh. "Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?" In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. 2022. URL: https://www2.cs.sfu.ca/~hamarneh/ecopy/aaai2022.pdf.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[8] A. Jacovi and Y. Goldberg. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386.

[9] S. Serrano and N. A. Smith. "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. URL: https://aclanthology.org/P19-1282.

[10] A. Yasodhara, A. Asgarian, D. Huang, and P. Sobhani. "On the Trustworthiness of Tree Ensemble Explainability Methods". In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2021, pp. 293–308.

[11] S. Arora, D. Pruthi, N. Sadeh, W. W. Cohen, Z. C. Lipton, and G. Neubig. "Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. 2022. arXiv: 2112.09669 [cs.LG].

[12] P. Hase and M. Bansal. "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" In: *arXiv preprint arXiv:2005.01831* (2020).

[13] D. Rajagopal, V. Balachandran, E. H. Hovy, and Y. Tsvetkov. "SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 836–850. DOI: 10.18653/v1/2021.emnlp-main.64. URL: https://aclanthology.org/2021.emnlp-main.64.

[14] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[15] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[16] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.

[18] D. Janzing, L. Minorics, and P. Blöbaum. "Feature relevance quantification in explainable AI: A causal problem". In: *International Conference on artificial intelligence and statistics*. PMLR. 2020, pp. 2907–2916.

[19] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 180–186.

[20] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang. "The practical implementation of artificial intelligence technologies in medicine". In: *Nature medicine* 25.1 (2019), pp. 30–36.

[21] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock. "Explainable AI in fintech risk management". In: *Frontiers in Artificial Intelligence* 3 (2020), p. 26.

[22] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence". In: *CoRR* abs/2004.12158 (2020). arXiv: 2004.12158. URL: https://arxiv.org/abs/2004.12158.

[23] R. Bhambhoria, H. Liu, S. Dahan, and X. Zhu. "Interpretable Low-Resource Legal Decision Making". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. 2022. arXiv: 2201.01164 [cs.LG].

[24] S. Paul, P. Goyal, and S. Ghosh. "LeSICiN: A Heterogeneous Graph-based Approach for Automatic Legal Statute Identification from Indian Legal Documents". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*. 2022.

[25] R. Bhambhoria, S. Dahan, and X. Zhu. "Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction". In: *The 34th Canadian Conference on Artificial Intelligence*. 2021.

[26] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. "Explainable AI for trees: From local explanations to global understanding". In: *arXiv preprint arXiv:1905.04610* (2019).

[27] S. Dahan. *Analytics and EU Courts: The Case of Trademark Disputes*. 2021. DOI: 10.2139/ssrn.3786069. URL: https://europepmc.org/article/PPR/PPR385653.

[28] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[29] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. "Neural additive models: Interpretable machine learning with neural nets". In: *Advances in Neural Information Processing Systems* 34 (2021).