Legal Fairness Analysis via Treatment Effect Estimation

Anonymous ACL submission

Abstract

Legal fairness is one of the most important principles pursued by modern legal systems. Unfortunately, unfairness may be inevitably introduced in real-world cases due to both objective and subjective uncertainty, such as ambiguity in the law or practical bias in judgments. Existing works for fairness analysis mainly rely on labor-intensive element annotation for cases, which suffer from limited generalization ability. To address this issue, we propose to utilize large-scale textual data to perform quantitative legal fairness analysis via our Causal-based Legal Fairness Measuring Framework (CaLF). To verify its effectiveness, we construct a legal-fairness dataset, and experimental results show that CaLF can accurately characterize the unfairness. Further, we adopt CaLF on a large-scale real-world dataset. Based on our settings and within our dataset, we have several interesting experimental observations from the perspective of gender, age, and region.

1 Introduction

Legal fairness is the principle that each individual is supposed to be treated equally before the law without discrimination, and it is regarded as an essential element of advanced law systems (Browne et al., 2001). Nevertheless, it is hard to achieve absolute fairness in the real world, and unfair judgments are sometimes inevitable in reality (Arvey, 1979; Hammond, 1996). Table 1 shows an example of different judgments between two similar cases in the real world. If no reasonable justifications present, such judgment undoubtedly undermines the principle of fairness, whether it results from subjective or objective uncertainty. Therefore, it is crucial to measure the judgment differences caused by specific factors (e.g., region, gender), which may uncover legal unfairness and can help regulate the judicial practice and prevent unfair judgments.

Case A: Alice stole a diamond ring worth 35,000 RMB from her friend. After arrested, Alice returned the stolen goods. Other circumstances of the defendant include confession and obtaining forgiveness.
Prison Term: 1 year 2 months (suspended for 2 years).
Alice: Female, age 20, from Region A.

Case B: Bob secretly stole an car (valued at 35,000 RMB) from his ex-girlfriend and sold it. The stolen property was recovered and returned to the victim. The defendant confessed to the crime.
Prison Term: 4 years 5 months.
Bob: Male, age 39, from Region B.

Table 1: An example of different judgments between two similar theft cases in the real world. Crucial legal elements are denoted with underlines. Case details can be found in the appendix.

Legal fairness analysis has been studied for decades (Douglas, 1949; Sheppard, 1985; Hoff, 1994; Reamer, 2005; Valvoda et al., 2021; Wang et al., 2021). Some quantitative studies attempt to use statistical methods to perform correlation analysis (Grogger and Ridgeway, 2006; Fryer Jr, 2019; Johnson et al., 2019), which cannot capture comprehensive information from complex factors and thus suffer from spurious correlation problem. To tackle this problem, causal inference is introduced to conduct causal effect analysis (Pierson et al., 2020; Gaebler et al., 2020; Knox et al., 2020). However, these works require to represent the cases with a few structured elements. Compared with original legal documents, the elemental representations need time-consuming and labor-intensive annotations. Thus, these methods are extremely restricted with generalization in the large-scale real-world analysis. Therefore, in this work, we aim to utilize large-scale legal documents to measure the unfairness, i.e., the causal effect on the judgment result.

However, the task is non-trivial, and there exist two crucial challenges: (1) Case Representation: Regarding the given textual legal cases, how to effectively generate expressive case representations for downstream analysis is a challenge. (2) Causal Effect Estimation: Legal judgments are usually
influenced by various factors. How to estimate the causal effect between the factors and the prison term is another challenge.

To address these issues, we propose a simple and effective Causal-based Legal Fairness Measuring Framework (CaLF), which uses neural models to extract expressive text representation, and then adopt a re-weighting causal model, inverse propensity weighting (IPW) (Rosenbaum, 1987; Rosenbaum and Rubin, 1983), to estimate the causal effect. Specifically, we first normalize the case distribution across groups by assigning each case a weight calculated by neural models. Then we measure the unfairness as the difference between the weighted average judgment result of different groups. Taking gender as an example, if males often commit more serious crimes than females and thus receive heavier sentences, we adopt re-weighting to balance the proportion of serious cases for the two genders, and the average judgment results can be compared for analysis.

Notably, CaLF can be applied to analyze the outcome of various judicial processes, including arrest, conviction, and sentencing, etc. In this paper, we choose the term of penalty (i.e., the outcome of the sentencing process) as our target for analysis, since it is the main punishment for offenders.

To verify the effectiveness of CaLF, we construct the first legal treatment effect estimation dataset, LegalTrEE. We annotate each case with factual legal elements, and we use a matching algorithm based on elemental trial (Cohen, 1982) to get the counterfactual outcomes. Experimental results on LegalTrEE prove that CaLF can achieve more accurate causal effect estimation than other models.

Furthermore, we adopt CaLF on the large-scale legal dataset from China, CAIL2018 (Xiao et al., 2018), to conduct the real-world legal fairness analysis. The experiment covers the perspective of age, gender, and region, while we also focus on 13 typical charges. From the result, we find some interesting biases in some specific aspects. The young tend to be sentenced to 1 month shorter than others on average, perhaps because of leniency towards students. Criminals in areas with high crime rates tend to be sentenced to 0.8 months longer than ones in areas with low crime rates, reflecting the traditional Chinese concept of “governing the country with severe law during trouble times”.

To summarize, we make several noteworthy contributions in this paper:¹

1. We design a framework, CaLF, which utilizes large-scale legal documents for fairness analysis. Compared with previous works, CaLF has better applicability and performance.
2. We build the first legal-domain treatment effect estimation dataset, LegalTrEE, on which we conduct comprehensive experiments to prove the reliability of CaLF.
3. We perform fairness analysis on large-scale real-world court decision data from the perspective of age, gender, and regional equality.

2 Related Work

2.1 Legal Fairness Analysis

Most of the current works on legal fairness are from a case-by-case or microcosmic perspective (Douglas, 1949; Sheppard, 1985; Tyler, 1988; Browne et al., 2001; Reamer, 2005; Hoff, 1994). Recently, many researchers attempt to analyze legal fairness quantitatively with statistical methods, such as correlation and regression analysis (Grogger and Ridgeway, 2006; Fryer Jr, 2019; Johnson et al., 2019), which cannot capture information from complex factors and suffer from spurious correlation problem. To tackle this issue, some researchers utilize the causal inference theory (Pierson et al., 2020; Gaebler et al., 2020; Knox et al., 2020). However, these works simplify the cases’ facts to a few structured elements, which need high-cost annotation. Besides, Wang et al. (2021) attempt to analyze legal fairness from large-scale textual data, but the method is limited by the unsatisfactory performance of sentencing prediction models (Zhong et al., 2020b). These existing methods are restricted with generalization in practice.

2.2 Treatment Effect Estimation

Treatment effect estimation aims to evaluate the causal effect of a given treatment on the outcome (Yao et al., 2020). Previous works mainly use elementary vectors as covariates, so they cannot be applied to our textual study (Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Rosenbaum and Rubin, 1985; Nie and Wager, 2017). In recent years, many researchers start to employ neural networks for text-oriented treatment effect estimation (Keith et al., 2020; Pham and Shen, 2017; Veitch et al., 2019). However, these works rely on the counterfactual

¹We will release our source code and the dataset once this paper is accepted.
outcome prediction, which is greatly challenging, especially in the legal domain. Due to the unsatisfactory performance of existing prison term prediction models, introducing outcome prediction in our task will bring bias to the results.

2.3 Legal AI

Legal AI focuses on applying artificial intelligence technology to help legal tasks (Zhong et al., 2020b). In recent years, with the development of deep learning, many researchers introduce natural language processing (NLP) technology to Legal AI and achieve remarkable progress on many tasks, such as legal judgment prediction (Chen et al., 2019; Zhong et al., 2020a; He et al., 2019), similar case matching (Tran et al., 2019; Xiao et al., 2019), legal information extraction (Chen et al., 2020; Shen et al., 2020), and jurisprudential perspectives verification (Valvoda et al., 2021). However, few works attempt to employ advanced NLP technologies to analyze legal fairness.

3 Methodology

In this section, we first describe notations and the problem formulation of legal fairness, and then introduce the proposed Causal-based Legal Fairness Measuring Framework (CaLF).

3.1 Notations

We formalize the problem as a treatment effect estimation task. We use the triplet \((X, Y, T)\) to represent a case:

**Covariate (background)** \(X\). In causal inference theory, covariate \(X\) is the background information of each sample. In our problem, the covariate \(X = (w_1, w_2, \ldots, w_l) \in \mathbb{R}^l\) represents the case’s factual information in plain text, where \(l\) denotes the text length and \(w_i\) denotes the \(i\)-th token.

**Outcome** \(Y\). We let the outcome \(Y \in \mathbb{R}\) to denote the judgment result. To better quantitatively measure the unfairness, we take the prison term (unit: month) as the judgment result in this paper, so we have \(Y \geq 0\). In practice, the outcome can also indicate other judgment results, such as fine, charged rate, etc.

**Treatment** \(T\). The treatment \(T \in \{0, 1\}\) is the potential unfair factor we study. In this paper, we take age \((\leq 28 \text{ or } > 28)\), gender (male or female), and region (south or north, GDP high or low, etc.) as \(T\) to detect the unfairness. In this way, samples

![Figure 1: A schematic diagram of CaLF. We employ IPW to estimate ATE and use a neural model to estimate the propensity score.](image)

The critical challenge in treatment effect estimation is that data distribution differs from groups, so we cannot simply compare the two groups’ mean values. Inverse propensity weighting (IPW) (Rosenbaum, 1987; Rosenbaum and Rubin, 1983) use re-weighting to balance the data distribution between groups and thus to get accurate measured value.
As the gender example in the introduction, fairness does not mean having identical average sentences for men and women. If a case is likely to be in the male group but actually in the female group, IPW will adjust its weighting upwards to balance the bulk of similar male cases. More generally, the more abnormal the factual treatment is, the more the case weights. Specifically, the re-weight for each sample \((x, y, t)\) is the inverse of the conditional probability \(\Pr(T = t | X = x)\). Finally, the ATE is estimated as the weighted mean prison term.

Formally, we are to estimate the ATE given an estimating dataset \(C^n = \{(x^{(i)}, y^{(i)}, t^{(i)}\})_{i=1}^{CN}\). Following previous works, treatment effect estimation relies on the two assumptions. One is unconfoundedness, which means legal documents contain sufficient information:

\[
T \perp Y (T = 0), Y (T = 1) | X.
\]

The other is overlap, which means no case definitely belongs to a specific group:

\[
0 < \Pr(T = 1 | X) < 1.
\]

In practice, both two assumptions can be satisfied for fairness analysis, when we employ the factual description as the covariate and the prison term as the outcome. Based on the two assumptions, we can employ inverse propensity weighting to estimate the ATE as:

\[
\text{ATE}_\text{IPW} = \frac{1}{CN} \sum_{i=1}^{CN} y^{(i)} \left( \frac{t^{(i)}}{e(x^{(i)})} - \frac{1 - t^{(i)}}{1 - e(x^{(i)})} \right).
\]

Here \(e(x)\) represents the propensity score (Rosenbaum and Rubin, 1983), defined as the conditional probability of treatment given covariates:

\[
e(x) = \Pr(T = 1 | X = x).
\]

### Estimating Propensity Score

Following Equation 4, we can estimate the ATE with propensity score. In this paper, since we are to encode plain-text legal documents, we employ neural models to estimate propensity scores. For previous works in the field of causal inference, topic models and word counts are widely adopted to deal with texts. However, these methods will lose much of the complex semantic information in legal documents and thus are not suitable for our work.

Specifically, we formalize the task as a binary classification problem. We train the model predicting treatment \(T\) with the covariate \(X\) as input, and the propensity score \(e(X)\) represents the output probability of \(T = 1\).

In practice, we can employ BERT (Devlin et al., 2019) or any other NLP models to get text encoding, and then we use a linear layer and a softmax layer to predict the propensity score \(e(x) = \hat{t}\). Besides, for training, we employ the cross-entropy loss function to optimize the model.

Moreover, there is another challenge for the neural network. As the sample numbers of \(T = 0\) and \(T = 1\) are usually unbalanced, neural models will overfit to the label with more samples, and the estimation of propensity score and ATE will be seriously affected. To resolve this problem, we use dropout and early stop to mitigate overfitting to prevent models from being overconfident. Details and analysis can be found in the appendix.

### 4 Dataset: LegalTrEE

To evaluate the effectiveness of CaLF and baselines, we construct the first legal causal dataset, Legal Treatment Effect Estimation Dataset (LegalTrEE). Based on the elemental trial theory, we manually annotate the legal elements of each case and match cases with similar elements in two groups. Then judgment differences between matched cases can be regarded as ground-truth unfairness, and thus the accuracy of model estimation can be evaluated.

China is the country we focus on in this paper. The Supreme People’s Court of China has published a large-scale legal document dataset, CAIL2018 (Xiao et al., 2018), which is currently one of the largest legal datasets and consists of millions of cases. It provides great data support for our work. Moreover, China has a large population and a vast territory, so there exist many complex factors (e.g., race, region) that may cause unfairness. Therefore, exploring judicial unfairness in the Chinese legal system is an interesting topic. Therefore, we construct LegalTrEE based on CAIL2018 to verify the effectiveness of CaLF. Notably, the CAIL2018 dataset is also used for our large-scale analysis.

Theft is the charge we focus on in LegalTrEE because it has the most cases in the CAIL2018 dataset. We select only one charge because involving multiple charges in the dataset require too many legal elements to be considered and annotated. Gender
is the treatment we focus on in LegalTrEE because it is one of the most-talked-about forms of discrimination. We define \( T = 1 \) to represent the defender in the case is male, and \( T = 0 \) for female.

To build a treatment effect estimation dataset, how to get the counterfactual outcome (i.e., \( Y(T = 1 - t) \)) is a challenge because it is often non-observable. Existing works mainly use domain-specific knowledge to build semi-synthetic datasets (Yao et al., 2020). In this paper, the counterfactual outcome denotes what the judgment will be if the treatment is reversed and the covariate remains. According to elemental trial theory in the legal domain (Tadros and Tierney, 2004; Cohen, 1982; Quintard-Morénas, 2010), judgments must be solely based on crucial legal elements extracted from the case fact. Therefore, the case in which the treatment is reversed and the elements are close enough can represent the counterfactual outcome. On this basis, we use matching to find such cases to build the complete LegalTrEE dataset.

Referring to the relevant articles and with the help of legal professionals, we enumerate 14 essential legal elements for theft cases’ sentencing. Then we pick thousands of theft cases from the CAIL2018 dataset and annotate them with these elements. We use a matching algorithm based on elemental trial to obtain these cases’ counterfactual outcomes, i.e., \( Y^{c}\text{-factual} = Y(T = 1 - t) \). Specifically, we match cases where the amount of stolen money (or value of the stolen property) is close and other elements are identical. Please refer to the appendix for more details of the legal elements and the matching algorithm.

Removing atypical cases that cannot be matched, we finally obtain our LegalTrEE dataset with 3,266 cases, of which 1,674 are female cases (\( T = 0 \)), and 1,592 are male cases (\( T = 1 \)). The statistics are shown in Table 2. From the table, we find males’ average prison term (6.207 months) is longer than females’ (4.033 months). However, there is an ATE of 0.985 months, which means that males expect to be sentenced to 0.985 months longer than females in the same criminal background.

<table>
<thead>
<tr>
<th>Group</th>
<th># Cases</th>
<th># Chars</th>
<th># Words</th>
<th>( Y^{\text{factual}} )</th>
<th>( Y^{c}\text{-factual} )</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3,266</td>
<td>520.51</td>
<td>292.84</td>
<td>6.207</td>
<td>5.119</td>
<td>1.088</td>
</tr>
<tr>
<td>Female (( T = 0 ))</td>
<td>1,674</td>
<td>489.44</td>
<td>276.77</td>
<td>4.033</td>
<td>4.921</td>
<td>0.887</td>
</tr>
<tr>
<td>Male (( T = 1 ))</td>
<td>1,592</td>
<td>504.58</td>
<td>284.60</td>
<td>5.093</td>
<td>5.017</td>
<td>0.985</td>
</tr>
</tbody>
</table>

Table 2: Statistics of LegalTrEE. Here \( Y^{\text{factual}} = Y(T = t) \) represents the factual outcome, that is, the factual judgment of the case in the real world. \( Y^{c}\text{-factual} = Y(T = 1 - t) \) represents the counterfactual outcome that we matched following elemental trial-based matching algorithm. The unit for \( Y^{\text{factual}} \), \( Y^{c}\text{-factual} \), and ATE is month.

We plot the distribution of \( Y(T = 0) \), \( Y(T = 1) \), and individual treatment effect (ITE) in Figure 2. Here ITE = \( Y(T = 1) - Y(T = 0) \) for each case, denoting the unfairness for each individual defendant. From the figure, we can find that the distribution of prison terms for males and females is similar, with females being slightly lower. Meanwhile, from the distribution of ITE, we can find that judgments are generally fair, although females are more often to receive a shorter sentence.

5 Experiment

In this section, we evaluate the performance of CalF on LegalTrEE, and then we apply CalF to analyze legal fairness on the CAIL2018 dataset.

5.1 Experimental Settings

Neural models. We employ CNN (Kim, 2014) and BERT (Devlin et al., 2019) as the encoder for CalF’s propensity score estimation. In Section 5.2, we test these models and find CNN outperforms other models, so we employ CNN as the encoder for the large-scale analysis in Section 5.3.

Baselines. We compare our proposed method with several representative baselines. We test traditional element-oriented methods as baselines, where linear regression is used for the prison term estimation, and logistic regression is used for the propensity score estimation (Yao et al., 2020). Further, we also introduce two neural causal methods as baselines: (1) Regression only (Regr.) (Keith et al., 2020). This method only uses regression to predict factual and counterfactual prison terms and simply subtracts them to obtain ATE. (2) Targeted maximum...
likelihood estimator (TMLE) (Van Der Laan and Rubin, 2006). TMLE is a doubly robust method that models both propensity score and outcome prediction to get better and more robust estimation performance. More specifically, TMLE subtracts the estimated prison term to get ATE like regression only but further use propensity score and well-designed methods to adjust the regression-predicted prison term.

**Dataset.** LegalTrEE is used for the effectiveness evaluation of CaLF and baseline models. CAIL2018 is used for the large-scale fairness analysis. We totally introduce more than 1.7 million cases for the large-scale analysis. Since the treatments that we concern involves age, gender, and region, we match all data with cases on China Judgment Online\(^2\) to obtain their source region and the defendants’ gender and age. In Section 5.2, we employ 3-fold cross-validation and randomly divide the train and test set by 2 : 1. In Section 5.3, for each experiment, we randomly select 50% of data for training and leave the remainder for analysis.

**Multiple experiments.** We repeat each experiment 50 times and guarantee the models to be well trained. We ensure the results pass the normality test (Shapiro-Wilk test). Suppose \((\mu, \sigma^2)\) are the mean and variance of the results, we report the 95% confidence interval of the results as \(\mu \pm 1.96 \frac{\sigma}{\sqrt{50}}\). Please refer to the appendix for more detailed settings and dataset statistics.

### 5.2 Experiments on LegalTrEE

<table>
<thead>
<tr>
<th>Method</th>
<th>avg ATE</th>
<th>avg(δ)</th>
<th>std(δ)</th>
<th>(δ = ATE - GT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth (GT)</td>
<td>0.985</td>
<td>0</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Element -Oriented</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regr.</td>
<td>0.095</td>
<td>0.080</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>TMLE</td>
<td>0.904</td>
<td>0.081</td>
<td>0.217</td>
<td></td>
</tr>
<tr>
<td>IPW</td>
<td>2.075</td>
<td>1.090</td>
<td>0.431</td>
<td></td>
</tr>
<tr>
<td>Neural Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regr. + CNN</td>
<td>0.546±0.070</td>
<td>0.439</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>Regr. + BERT</td>
<td>3.562±0.363</td>
<td>2.577</td>
<td>0.313</td>
<td></td>
</tr>
<tr>
<td>TMLE + CNN</td>
<td>1.626±0.036</td>
<td>0.641</td>
<td>0.315</td>
<td></td>
</tr>
<tr>
<td>TMLE + BERT</td>
<td>2.784±0.218</td>
<td>1.799</td>
<td>0.347</td>
<td></td>
</tr>
<tr>
<td>CaLF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPW + CNN</td>
<td>1.339±0.052</td>
<td>0.354</td>
<td>0.318</td>
<td></td>
</tr>
<tr>
<td>IPW + BERT</td>
<td>1.848±0.252</td>
<td>0.863</td>
<td>0.509</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Experimental results for CaLF and baseline methods on LegalTrEE (unit: month). We employ 3-fold cross-validation and report the average result, average error, and errors’ standard deviation.

We test the performance of CaLF and the baseline methods on LegalTrEE using 3-fold cross-validation, as shown in Table 3. From the results, we can observe that CaLF with CNN can outperform other text-oriented methods, and the average analysis error is less than 11 days.

Notably, both two neural baselines have worse performance than CaLF. This problem is likely to be caused by the unsatisfied performance of the prison term prediction model (Zhong et al., 2020b; Chen et al., 2019), which brings bias to the results. Therefore, these baselines that need prison term prediction is not suitable for our work.

We notice that the performance of BERT-based methods is worse than CNN-based methods. From the observation, we find that BERT suffers from the overfitting problem and usually captures subtle features that are irrelevant to the judgment. Thus, BERT can predict treatment labels accurately but fails to accurately estimate the propensity scores.

Besides, we also find that element-oriented Regr. and TMLE achieve the best performance among all methods, even better than CaLF with CNN. Since element-oriented approaches introduce legal knowledge to the problem and simplify cases to a few elements, they can perform well on the regression task of prison term prediction. However, these methods are not comparable to text-oriented methods. Text-oriented methods can be easily applied to the analysis of large-scale textual legal documents. In contrast, if we want to use element-oriented methods for such analysis, the high-cost manual annotation must be introduced, which makes the task highly unacceptable.

### 5.3 Large-Scale Analysis on Real-World Data

From the results mentioned above, we can find that CaLF with CNN can accurately estimate the legal fairness. Thus, in this section, we perform the large-scale analysis on the CAIL2018 dataset using CaLF with CNN. Specifically, we measure both overall unfairness and unfairness for the 13 typical charges with sufficient data in the CAIL2018 dataset. Please refer to the appendix for the detailed descriptions of the charges.

As Table 4 shows, we select gender, age, and region as factors (treatments) for experiments. Gender is defined biologically as male or female. Age is divided as \(\leq 28\) or \(> 28\) because the age of 28 is considered as the standard of whether a citizen is mature enough to take the responsibilities (Zhou, 2018). Region is used to test if the human geographical environment, regional economic status, cultural development, and crime rate will affect the

\(^2\)http://wenshu.court.gov.cn/
As mentioned above, the age of ATE measured by CaLF in the Chinese legal system, so we use it as our reference threshold. For example, in drug specific charges, young are often sentenced shorter. Moreover, we also find some interesting phenomena with distinct regularity:

### Table 4: The descriptions of the factors.

<table>
<thead>
<tr>
<th>Charge</th>
<th>Average Treatment Effect (ATE)</th>
<th>Gender</th>
<th>S or N</th>
<th>Region Split by EDU</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-1.0±0.2</td>
<td>-0.1±0.2</td>
<td>-0.3±0.3</td>
<td>-1.0±0.3</td>
<td>0.2±0.3</td>
</tr>
<tr>
<td>Drug Trafficking</td>
<td>-3.7±0.3</td>
<td>0.3±0.7</td>
<td>1.1±0.6</td>
<td>-4.0±0.7</td>
<td>-4.1±0.6</td>
</tr>
<tr>
<td>Theft</td>
<td>-1.3±0.1</td>
<td>2.2±0.2</td>
<td>0.5±0.2</td>
<td>0.3±0.2</td>
<td>0.0±0.2</td>
</tr>
<tr>
<td>Disrupting Public Service</td>
<td>-0.8±0.2</td>
<td>1.3±0.1</td>
<td>-0.5±0.1</td>
<td>-0.1±0.1</td>
<td>-0.1±0.1</td>
</tr>
<tr>
<td>Possession of Illegal Drugs</td>
<td>-9.2±1.0</td>
<td>-2.0±0.7</td>
<td>0.0±0.4</td>
<td>3.8±0.3</td>
<td>3.5±0.6</td>
</tr>
<tr>
<td>Unlawful Detention</td>
<td>-2.1±0.2</td>
<td>-4.0±0.1</td>
<td>0.4±0.1</td>
<td>0.4±0.1</td>
<td>0.6±0.1</td>
</tr>
<tr>
<td>Intentional Injury</td>
<td>-0.2±0.2</td>
<td>0.5±0.4</td>
<td>0.0±0.3</td>
<td>1.3±0.5</td>
<td>0.1±0.3</td>
</tr>
<tr>
<td>Traffic Offence</td>
<td>-2.2±0.5</td>
<td>1.2±0.4</td>
<td>-1.2±0.5</td>
<td>1.4±0.4</td>
<td>0.2±0.3</td>
</tr>
<tr>
<td>Robbery</td>
<td>-17.6±1.2</td>
<td>-0.7±0.7</td>
<td>-1.2±0.9</td>
<td>1.2±0.8</td>
<td>9.7±1.3</td>
</tr>
<tr>
<td>Racketeering</td>
<td>-11.9±0.4</td>
<td>-5.5±0.4</td>
<td>-3.1±0.2</td>
<td>0.2±0.2</td>
<td>0.9±0.2</td>
</tr>
<tr>
<td>Providing Venues for Drug Users</td>
<td>-7.2±0.1</td>
<td>0.5±0.1</td>
<td>-0.1±0.1</td>
<td>1.0±0.1</td>
<td>0.3±0.2</td>
</tr>
<tr>
<td>Picking Quarrels and Provoking Trouble</td>
<td>-1.8±0.2</td>
<td>-3.7±0.3</td>
<td>-0.3±0.1</td>
<td>0.8±0.3</td>
<td>0.7±0.3</td>
</tr>
<tr>
<td>Fraud</td>
<td>-6.8±1.0</td>
<td>-3.6±0.8</td>
<td>-1.6±0.6</td>
<td>3.0±0.9</td>
<td>0.5±0.8</td>
</tr>
</tbody>
</table>

Table 5: The experimental results of the average treatment effect (ATE) (unit: month). For example, in drug trafficking cases, the youth (age ≤ 28) expect to be sentenced to 3.7 ± 0.3 months shorter than others (age > 28) in the same criminal background. The results show that there is little judicial unfairness in China generally. The results which are considered unfair (with absolute values over 3 months) are denoted with underlines.

judgment results. In this paper, regions are divided by the provincial administrative units of China.

Table 5 shows the experimental results of the ATE measured by CaLF in the Chinese legal system. If there are no lost confounders in legal documents and CNN can accurately give propensity scores, the results represent how much more the treatment group will be sentenced than the control group on average. For example, gender only causes 0.1 months’ bias (favoring men) for all cases, and for possession of illegal drugs cases, the number increases to 2. From the results, we can find that:

From the results, we can observe that the overall sentencing biases for focused treatments are all shorter than about 1 month. Moreover, for specific charges, the sentencing bias does not exceed 3 months in more than 80% of results. And there are still some significant differences that can be observed. For example, age plays a significant role in the robbery. Notably, here 3 months is generally the minimum unit of sentencing in the Chinese legal system, so we use it as our reference threshold.

Moreover, we also find some interesting phenomena with distinct regularity:

### The youth are favored. For either overall or specific charges, young are often sentenced shorter. As mentioned above, the age of 28 is thought as the standard of whether a citizen is mature (Zhou, 2018). Further, those not older than 28 include a large group of students. Therefore, the observation that youth are often favored can be explained that judges tend to give more forgiveness and leniency to immature young people and students.

“Governing the country with severe law during trouble times”. Overall, criminals in areas with high crime rates tend to be sentenced 0.8 months longer than ones in areas with low crime rates, as is the situation for most charges. This traditional Chinese concept is recorded in the Rites of Zhou. In the modern Chinese legal system, it is also well documented. The thought of retribution sets the upper limit of a crime, while the aim of prevention might reduce the sentence (Zhang, 2011). In other words, it is necessary for judges to have discretion power for the purpose of prevention. In western criminal policy theory, the deterrence theory is a similar concept (Paternoster, 2010). The core idea of deterrence is that offenders may weigh the costs and benefits of crime, so when people feel that security is deteriorating, it is easy to think that “crime can be reduced by increasing penalties.”

### The judicial local protectionism is not significant. The judicial local protectionism is a controversial issue in China, referring to the local authori-
ties’ intervention in the judgment for the protection of various local interests, which affects the fairness and justice of the legal system (Liu, 2003, 2016). However, according to our experimental results, this phenomenon is not significant as described by the jurisprudence from the perspective of sentencing. Specifically, it is not systematically found in general, regardless of region’s location, GDP, educational status, or crime rate.

Moreover, to detect the fine-grained judicial local protectionism, we pick 8 representative provincial-level administrative regions, which are different in terms of geographic location, GDP, and educational status, for more detailed experiments. Please refer to the appendix for the descriptions for them. We use the one vs. all to perform a fine-grained experiment to measure the unfairness of these regions. The results are shown in Table 6. This result further demonstrates that judicial local protectionism in China is not a systematic problem, for the overall sentencing ranges of each crime tested are all shorter than the 3 months standard.

To sum up, in this section we mainly employ CaLF to conduct real-world legal fairness analysis on the CAIL2018 dataset from the perspective of gender, age, and region. Based on our settings and dataset, we find some interesting phenomena. Besides, there are also some significant variances in some specific aspects, which cannot be explained right now. We will leave these as our future work and to the legal community.

### 6 Discussion

Since legal fairness is a principled, serious, and sensitive topic, it is necessary to further discuss the potential limitations and risks of our approach. Here we list several important issues which may lead to biased conclusions.

1. **Data collection.** We collect our dataset from the cases published by the Chinese government. Due to confidentiality, there are still some non-public cases, which means that there may be a difference between the distribution of the collected data and that of the real world. If such differences exist, the results will be unreal. 2. **The subjectivity in legal documents.** Legal practitioners strive to follow the guidelines of objectivity and comprehensiveness in the process of writing legal documents. However, there are no golden rules for writing legal documents and it is difficult to achieve absolutely objective. The inevitable subjectivity in the legal documents may introduce bias to the conclusion.

3. **The two assumptions.** As mentioned in Section 3, IPW is based on the unconfoundedness and overlap assumption. If relevant criminal information is missing (unconfoundedness violated) or the case distribution of the two groups does not overlap (overlap violated), then reweighting will not work and the measured ATE will be biased. 4. **Limitations of models.** Regardless of the model employed, there are inevitably prediction errors, leading to biased propensity scores and thus affect measured results. Furthermore, pre-trained word embedding or language models may suffer from ethical issues, and we have more discussions about this in the Ethical Considerations.

### 7 Conclusion and Future Work

In this paper, we formalize legal fairness analysis as the treatment effect estimation task and propose CaLF, a Causal-based Legal Fairness Measuring Framework. We construct the first legal treatment effect estimation dataset to verify the effectiveness of CaLF. Then we conduct analysis on the large-scale legal dataset, CAIL2018, from the perspective of gender, age, and region. From the experiment, we find some interesting phenomena, such as youth tend to be mitigated, and troubled regions tend to have harsher penalties.

We will explore the following directions in the future: 1. We will combine legal knowledge to carry out more in-depth analysis and give comprehensive explanations of the experimental observation. 2. Since the legal systems of different countries are very different, we will attempt to conduct legal fairness analysis for other countries. Given sufficient open data, such analysis and comparison will be interesting, as well as of great importance.

We hope with the development of legal fairness analysis, legal judgments around the world can become more transparent and fair, and equality before the law can be truly achieved.

---

Table 6: Experimental results on 8 specific regions (unit: month).

<table>
<thead>
<tr>
<th>Region</th>
<th>Shanghai</th>
<th>Fujian</th>
<th>Shandong</th>
<th>Liaoning</th>
<th>Jilin</th>
<th>Henan</th>
<th>Sichuan</th>
<th>Guangxi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity</td>
<td>-2.4±0.1</td>
<td>-1.9±0.1</td>
<td>0.8±0.2</td>
<td>0.6±0.3</td>
<td>1.0±0.1</td>
<td>-1.2±0.1</td>
<td>2.3±0.2</td>
<td>1.4±0.2</td>
</tr>
</tbody>
</table>
**Ethical Considerations**

Our approach CaLF mainly focuses on utilizing large-scale legal documents to analyze legal fairness. To verify the effectiveness of our model, we construct a dataset LegalTrEE via manual annotation. During the annotation stage, we first annotate some cases on our own to approximate the workload, and then we determine annotators' wages based on local standards. All the legal documents we used in our work are published by the Supreme People’s Court of China, and the participant names are anonymized.

The goal of this work is to leverage AI technology for legal fairness analysis and promote equality and non-discrimination around the world, instead of praising or criticizing any country’s legal system or for any political purpose. The misuse of our method for political or private interest will have a negative impact. We encourage the regulators to make efforts to mitigate this risk.

We aim to help regulate the judicial practice and prevent unfair judgments via fairness analysis. Unfortunately, the failure tolerance is a problem that even the best techniques may fail. For instance, the pre-trained word embedding or pre-trained language models may mislead the results. But our method is agnostic to the specific choice of neural encoders, and the ethically-compliant encoders can be easily adopted in our method. We also encourage the community to devote effort to mitigate the bias for legal fairness analysis.

**References**


Congqing He, Li Peng, Yuquan Le, Jiawei He, and Xiangyu Zhu. 2019. *Secaps: A sequence enhanced capsule model for charge prediction.* In *Proceedings of ICANN,* pages 227–239. Springer.


Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. *Analogical reasoning on Chinese morphological and semantic relations.* In *Proceedings of ACL,* pages 138–143.


A LegalTrEE: Legal Treatment Effect Estimation Dataset

In this part, we describe the details of our new dataset, the Legal Treatment Effect Estimation Dataset, LegalTrEE.
A.1 Elements Description

For each case, we represent it as an element vector, $X \in \mathbb{R}^{14}$. In other words, there are 14 elements related to the theft cases’ sentencing, according to the Chinese Criminal Law, and “The interpretation of several issues on the application of the law in handling criminal cases of theft” published by the Chinese Supreme People’s Court and Chinese Supreme People’s Procuratorate. Here we describe these elements in turn.

Amount of theft: $x_1 \in \mathbb{R}$. The amount of stolen money or properties. The core element for theft cases’ sentencing. For the stolen objects, the value is based on the valuation in the legal instrument.

Ratio of refund: $x_2 \in [0, 1]$. If a person commits theft, he/she may be mitigated if he/she returns the stolen goods or makes restitution. We define $x_2$ to represent the restitution as a percentage of the amount of theft.

Level of theft amount: $x_3 \in \{0, 1, 2, 3\}$. The level of theft amount is divided into relatively large, huge, and especially huge. The standards of these levels vary slightly from region to region. For this dataset, we select cases from several specific regions, so that the standards for relatively large, huge, and especially huge are 2,000, 60,000, and 400,000 yuan, respectively. Here we use $x_3 = 1$ to represent the amount is relatively large, use $x_3 = 2$ to represent the amount is huge, use $x_3 = 3$ to represent the amount is especially huge, and use $x_3 = 0$ to represent the amount does not reach a relatively large amount.

Burglary: $x_4 \in \{0, 1\}$. The element is used to represent whether the criminal intrudes into another person’s residence to steal.

Multiple thefts: $x_5 \in \{0, 1\}$. The element is used to represent committed thefts more than three times within two years.

With a murder weapon: $x_6 \in \{0, 1\}$. The element is used to represent whether theft with firearms, explosives, control knives, other instruments prohibited by the government, or other instruments sufficient to endanger others’ safety.

Pickpocketing: $x_7 \in \{0, 1\}$. The element is used to represent theft of property carried by others in public places or public transport.

Minors: $x_8 \in \{0, 1\}$. Minors under the age of 18 are persons of limited criminal responsibility and shall be punished less severely.

75 years old or older: $x_9 \in \{0, 1\}$. A person over the age of 75 is of limited criminal responsibility and shall be punished less severely.

Psychosis: $x_{10} \in \{0, 1\}$. Psychosis who have not yet completely lost the ability to recognize or control can be punished less severely.

Voluntary surrenders: $x_{11} \in \{0, 1\}$. (Article 67 of the Chinese Criminal Law) Voluntary surrenders refers to the act of voluntarily delivering oneself up to justice and truthfully confessing one’s crime after one has committed the crime. Any criminal who voluntarily surrenders may be given a mitigated punishment. The ones whose crimes are relatively minor may be exempted from punishment.

Recidivism: $x_{12} \in \{0, 1\}$. (Article 65 of the Chinese Criminal Law) If a criminal commits another crime punishable by fixed-term imprisonment or heavier penalty within five years after serving his sentence of not less than fixed-term imprisonment or receiving a pardon, he is a recidivist and shall be given a heavier punishment.

Other aggravating circumstances: $x_{13} \in \{0, 1\}$. The element is used to represent whether there are other aggravating circumstances.

Other mitigating circumstances: $x_{14} \in \{0, 1\}$. The element is used to represent whether there are other mitigating circumstances.

For those binary elements that take a value in $\{0, 1\}$, we have 1 means yes, and 0 means no. For $x_{13}$ and $x_{14}$, the description in the judgment document prevails.

A.2 Inter-annotator Agreement

The Krippendorff’s alpha of the annotation is 0.94.

A.3 Matching Algorithm

We first define the matching score of cases and then find matched cases between the treated and control group according to the matching score.

Definition of Matching Score

For case $c_A = (x^A, y^A, t^A)$, and case $c_B = (x^B, y^B, t^B = 1 - t^A)$, we specify that they are matchable if and only if $x^A = x^B, \forall i = 4, 5, \ldots, 14$. Then, for the two matchable cases, we define the matching score of $c_B$ to $c_A$ as:

$$\text{match}_{c_A}(c_B) = \theta - \alpha \frac{|x^A_i - x^B_i|}{\max \{x^A_i, x^B_i\}}$$

$$- \beta |x^A_i - x^B_i|$$

$$- \gamma |x^A_i - x^B_i|.$$

(6)

Here $\alpha, \beta, \gamma, \delta, \theta$ are parameters.

The main idea of the matching score is based on the elemental trial, i.e., judgment should be
only correlated with the legal element. Therefore, for those $x_i (i = 4, 5, \ldots, 14)$ do not match, we consider that they are unmatchable; otherwise, we hope the other three elements, the amount of theft ($x_1$), the ratio of refund ($x_2$), and the level of theft amount ($x_3$) to be as close as possible.

**Matching for Counterfactual Outcomes**

For each case $c$ in the treated group, we will try to find a matched case $c_M$ in the control group. Vice versa, for each case $c$ in the control group, we will try to find $c_M$ in the treated group. If we find such a well-matched case, we set the counterfactual outcome of case $c$ as the factual outcome of case $c_M$, i.e., we set $Y^{c\text{-factual}} (c) = Y (c_M)$.

Specifically, for case $c$, we first get the maximum matching score that other cases can achieve with it:

$$m_c = \max_{c_m \in \text{matchable}} Y^{c\text{-match}} (c_m).$$

As a particular case, for those where no matchable cases can be found, and for those where $m_c < 0$, we simply remove them from the dataset. Then, we construct the candidate set containing cases with a similar matching score:

$$C_{\text{candidate}} = \{c_m | Y^{c\text{-match}} (c_m) + \epsilon \geq m_c\}.$$ (8)

Here $\epsilon$ is another parameter for the algorithm. Finally, we randomly select a case in the candidate set with equal probability as the matching case for $c$. Since the cases in the candidate set are all similar enough to the cases to be matched, the random selection simulates the randomness of judges.

**Parameter Values**

For the matching algorithm, we have 6 parameters, $\alpha, \beta, \gamma, \delta, \theta,$ and $\epsilon$. The value of these parameters when we build LegalTREE is shown in Table 7.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\delta$</th>
<th>$\theta$</th>
<th>$\epsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>2.5</td>
<td>0.5</td>
<td>100</td>
<td>1.0</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 7: Parameter values of matching in practice.

**C Baselines and Models**

In this section, we introduce the two neural models used in our experiments.

**CNN (Kim, 2014)**: This work proposes Convolutional Neural Networks with multiple filter widths specifically for text classification.

**BERT (Devlin et al., 2019)**: BERT is the model formed by multiple bidirectional Transformer layers. The parameters of BERT has been fully pre-trained on large-scale text corpora. In this paper, we employ the BERT pre-trained on Chinese corpora for experiments.

**D Experimental Settings**

Here we mainly introduce the training settings such as the optimizer, neural models’ hyper-parameters, and how we guarantee that the models are well trained. Other experimental settings can be found in the main text.

**D.1 Basic Training Settings**

We use Adam (Kingma and Ba, 2015) to train CNN and use BertAdam (Devlin et al., 2019) to train BERT. We employ character-level embedding to train BERT and use external Chinese word vectors (Li et al., 2018) to train CNN.

We train models with NVIDIA GTX 2080 Ti.

**D.2 Hyper-Parameters of Neural models**

The hyper-parameters of neural models are shown in Table 11.

**D.3 How We Ensure the Models Are Well Trained**

In the experiment, we guarantee that our model is neither over-fitted nor under-fitted.

To prevent models from over-fitting, we employ the dropout layer and the early stop mechanism. The dropout ratio is shown in the model hyper-parameter part. The early stop mechanism force the training to stop when the loss does not decrease on the valid set. In these ways, the model does not output a too extreme distribution of propensity scores and thus produces more stable ATE results.

To prevent models from under-fitting, when taking the results of 50 replications of each experiment, we only select the models that are well-performed on the treatment assignment’s classification task. Specifically, we ignore those results

---

3 https://github.com/Embedding/Chinese-Word-Vectors
where the macro-f1 is below 95% of the optimal value on the classification task.

## E Plot of Propensity Scores

Figure 3 shows propensity scores for the two groups when using CNN. We can find that with the help of early stop and dropout, the CNN does not fall into overconfident prediction, so the reliability of the IPW scheme is guaranteed.

## F Charge Description

In this section, we describe the criminal charges mentioned in this paper. All the descriptions refer to Chinese Criminal Law.

**Drug Trafficking**: The act of knowingly smuggling, trafficking, transporting, or manufacturing drugs.

**Theft**: The act that, for the purpose of illegal possession, steals a relatively large amount of public or private property or commits theft repeatedly.

**Disrupting Public Service**: The act that, by means of violence or threat, obstructs a functionary of a State organ from carrying out his functions according to law.

**Possession of Illegal Drugs**: The act that ille-
gally possesses drugs (e.g., opium, heroin, methyl aniline) knowingly of relatively large quantities.

**Unlawful Detention**: The act that unlawfully detains another person or unlawfully deprives the personal freedom of another person by any other means.

**Intentional Injury**: The act that intentionally inflicts injury upon another person.

**Traffic Offence**: The act violates regulations governing traffic and transportation and thereby causes a serious accident, resulting in serious injuries or deaths or heavy losses of public or private property.

**Robbery**: The act that, for the purpose of illegal possession, robs public or private property by violence, coercion, or other methods.

**Racketeering**: The act that, for the purpose of illegal possession, extorts public or private money or property by blackmail.

**Providing Venues for Drug Users**: The act that provides shelter for another person to ingest or inject narcotic drugs.

**Picking Quarrels and Provoking Trouble**: Committing any of the following acts of creating disturbances, thus disrupting public order: (1) beating another person at will and to a flagrant extent; (2) chasing, intercepting, or hurling insults to another person to a flagrant extent; (3) forcibly taking or demanding, willfully damaging, destroying or occupying public or private money or property to a serious extent; or (4) creating disturbances in a public place, thus causing serious disorder in such place.

**Fraud**: The act that, for the purpose of illegal possession, swindles public or private money or property, and if the amount is relatively large.

### G Details of Example Case in Introduction

Here we show the detailed description of the two real-world example cases in the introduction. Due to anonymity, some details (such as names, locations, dates, etc.) have been manually omitted.

**Case A.** In a karaoke room in Region A, when the defendant Alice was helping the victim Carol find a lost ring, she took it away in her purse when Carol was not paying attention. A few months later, the defendant Alice was found by Carol wearing her lost ring while she was playing with her cell phone in Carol’s store. The stolen ring was tested by a mineral testing center and found to be a diamond ring. The price certification center determined that the stolen diamond ring’s retail market price was RMB 35,000 on the day of the crime. After the crime, the relatives of the defendant Alice compensated Carol on her behalf, and Carol expressed her understanding to the defendant Alice. The defendant Alice confessed to the crime in a good manner.

**Case B.** In Region B, defendant Bob stole his ex-girlfriend Daisy’s car parked on the north side of the neighborhood gate and sold the vehicle for RMB 20,000, squandering the proceeds. The stolen vehicle was appraised to be worth RMB 35,000. The stolen vehicle was extracted and returned to the owner after the crime. The defendant Bob confessed to the crime in a good manner.