An Empirical Study on Group Fairness Metrics of Judicial Data

YANJUN LI1,4, HUAN HUANG2, XINWEI GUO3, AND YUYU YUANT

1Software College, Beijing University of Posts and Telecommunications, Beijing 100876, China
2Shenzhen Research Institute of Nanjing University, Shenzhen 518000, China
3China Justice Big Data Institute Company Ltd., Beijing 100070, China
4China National Accreditation Service for Conformity Assessment, Beijing 100062, China

Corresponding author: Huan Huang (huanghuan@mooctest.com)

This work was supported in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant YFJGJS1.0 and Grant CIGJZD20200617103001003.

ABSTRACT Group fairness means that different groups have an equal probability of being predicted for one aspect. It is a significant fairness definition, which is conducive to maintaining social harmony and stability. Fairness is a vital issue when an artificial intelligence software system is used to make judicial decisions. Either data or algorithm alone may lead to unfair results. Determining the fairness of the dataset is a prerequisite for studying the fairness of algorithms. This paper focuses on the dataset to research group fairness from both micro and macro views. We propose a framework to determine the sensitive attributes of a dataset and metrics to measure the fair degree of sensitive attributes. We conducted experiments and statistical analysis of the judicial data to demonstrate the framework and metric approach better. The framework and metric approach can be applied to datasets of other domains, providing persuasive evidence for the effectiveness and availability of algorithmic fairness research. It opens up a new way for the research of the fairness of the dataset.

INDEX TERMS Data fairness, sensitive attributes, metric, judicial data.

I. INTRODUCTION

Machine learning algorithms are widely used as artificial intelligence (AI) software systems in many fields of our daily life, including the judicial field [1]. Because machine learning algorithms are highly dependent on training data, the quality of the dataset used by algorithms poses threats to AI software systems [2], which may cause them to make unfair or even wrong decisions. Artificial intelligence tends to have discriminatory problems in judicial decisions [3], [4]. Fairness is one of the formal norms that people pay close attention to, particularly the fairness of the judiciary.

There have been many definitions of fairness and fairness metrics in artificial intelligence (AI) [5], [6], such as awareness fairness, unawareness fairness, bayesian fairness, etc. We will introduce them detailly in the fourth section. Although many studies pay attention to the fairness of AI, most studies focus on definitions of fairness and algorithm of training fair models, and very few mention the connection between fairness and data quality.

Experts are inclined to focus on whether the algorithm is fair [7], [8], yet the fairness of data gains less attention in the judicial field. Although algorithms are always considered to be responsible for unfair judgment results [9], the unfair results could occur when using a perfect algorithm and poor-quality data. When an expert trained a model using a dataset and a specific algorithm, how to perform valid experiments to determine if the data is responsible for the unfair circumstances, it always puzzles every researcher.

We propose a new metric to measure the fairness of a sensitive attribute by distinguishing values and a framework to study data fairness, using judicial data as an example. The first step is to figure out the sensitive attributes of the dataset. Calculating the fairness of the attribute is the second step. Then show the impact of the sensitive attribute on the fairness of the experimental data. The second and third steps respectively observe the fairness of sensitive attributes in different ways.

From the perspective of the individual and the group, fairness is classified as individual fairness [10], [11] and group fairness [8], [12]. Individual fairness emphasizes that the output of two similar inputs must be similar. In contrast, group fairness emphasizes that the classification distribution on the
two groups should be as similar as possible. Group fairness is a significant fairness definition, which is widely accepted. It means the different groups have an equal probability of being predicted for one aspect. For example, two different individuals with the same criminal background should have the same predicted probability of recidivism.

Our work can solve the problem through a series of steps on the dataset’s attributes to determine the group fairness of the specific dataset. This type of fairness protects vulnerable groups without discrimination to some degree.

We hope to use our framework and metric mentioned above to study the fairness of data in the judicial field and even extend to applying to other fields, which may improve data quality. We aim to reveal what problems or prejudices exist, provide persuasive evidence for the effectiveness and availability of algorithmic fairness research, and open up a new way to research the fairness of datasets about data. Data scientists can grasp the dataset’s overall situation more clearly after using our framework and metric. Our contributions include:

- A redefinition of sensitive attributes: data-related attributes. The first study is about data fairness with judicial characteristics.
- A Framework for obtaining sensitive attributes: integrated methods. It is used to find what prejudices may exist in the dataset.
- A novel fairness metric: overlapping area (OA). It aims to present the fairness degree of a specific sensitive attribute.

II. APPROACH

In this section, we define some sensitive attributes that are characteristically related to the specific data that accidentally affect the results. In our ideal expectations, this set of attributes should be irrelevant to the model’s predictions. Besides, we have evaluated the utility of each attribute according to its execution on the judicial decisions. COMPAS is a risk assessment tool designed by Northpointe. It will conduct a risk assessment on the possibility of the defendant committing a crime again based on the information recorded in the defendant’s criminal files and questionnaires. Therefore, the judicial decisions should remain what it is as the sensitive attributes change.

A. MOTIVATION

In most studies, the common sensitive attributes generally are race and nationality, etc. There are discriminatory judgments caused by race and nationality in the judicial field, just like the theft of Brisha Borden in 2014. Borden and her friend Prater grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs. Borden and her friend were arrested and charged with burglary and petty theft for the items. When Borden and Prater were booked into jail, the computer program used for risk assessment gave them a score to predict their possibility of committing a crime in the future. The results showed that Borden—who is black—was rated as high risk by the program, Prater—who is white—was rated as low risk. Two years later, Borden has not been charged with any new crimes, while Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars worth of electronics.

However, the attributes mentioned race and nationality, etc., are more suitable for the social conditions of the multi-ethnic mixed areas. Take China’s judicial data as an example; criminals of other nationalities can almost be neglected based on the sizeable Chinese population. It isn’t very meaningful to regard nationality as a sensitive attribute. We need to find sensitive attributes that are consistent with Chinese national conditions. This paper proposes a framework that can be widely used to obtain the specific sensitive attributes for a specific dataset to develop data fairness research.

In the process of training learners, without sensitive attributes for avoiding discrimination often leads to a worse training effect, so as to lead to unfair results. A group must have some common characteristics, and then AI will use most of these common characteristics as its features. Once an object without the typical features or belongs to a small part, it may be given negative feedback. Researching sensitive attributes is necessary because it is inevitable to use them.

B. ATTRIBUTES

According legal experience and data analysis, select and determine a set of sensitive attributes $A = \{a_i, i \in N^+\}, a_i$ is a suspicious sensitive attribute, such as gender, nationality, region, etc. What we need to find is a sensitive attribute related to data characteristics in this step. For instance, we think the region is one of the sensitive attributes of judicial data. The sentence length for cases that happened in region A and region B should be the same.

Attribute substitution is a crucial step, which would verify the sensitivity of the property. The set $A$ of sensitive attributes determined in the step of initial analysis is based on experience only. Inspired by counterfactual fairness [13], we use the method to perform experiments on the replacement of sensitive attributes. Once we deal with one attribute $a_i$, change its value to an another from the value-set randomly. Here we use the trained model to get a set of indicators about model evaluation to identify sensitive attributes.

After data preprocessing, take the results obtained without changing the features of the dataset as the benchmark experiment. Compared with the results of the benchmark experiment, we can get the gap $\Delta_i$ of both values for accuracy score or RMSE to verify whether the attribute is sensitive:

$$\Delta_i = |\text{Indicator}(BM) - \text{Indicator}(AS)|$$ (1)

$\text{Indicator}(BM)$ is the indicator from the benchmark experiment, and $\text{Indicator}(AS)$ is the indicator from the step of Attribute substitution, $i$ is the sign for identifying indicator.

If the number of $i$ is bigger than one, means there are more than one kind of indicator, such as MSE, RMSE, MAE, R2, etc. We can set thresholds for every $\Delta$. When $\Delta$ is
bigger than the threshold (decided by the researcher), like threshold = 0.1, it means that this attribute may influence the trainer’s result. Then we can say it is one of the sensitive attributes that belong to the dataset. The gap gained in this step is a basis; it cannot identify the sensitive attribute directly. Because of the state, clear criteria do not exist; everyone has a different definition of fairness. At this time, people can set their threshold when they need it.

C. METRICS

Inspired by Gaussian distribution [14], combining with the Law of Large Numbers and Central Limit Theorem, our primary contribution is a metric for measuring the fairness of data-related sensitive attributes. When an attribute is a sensitive attribute, it means that its different values have distinguished effects on the learning process of the trainer, thus affecting the judgment result. So we can say that the attribute is unfair, which needs to be judged the degree of fairness. Suppose it is a binary attribute (A and B); we can explain it by the theory of counter-evidence. If this attribute is fair, regardless of its value A or B, the probability of the result should be the same, namely that \( P(A \cap B) \rightarrow 1 \), A and B are almost coincident. Many datasets typically follow a normal distribution. We used a Gaussian distribution to represent the probability distributions of values about a particular attribute. And then, we compared distributions from multiple values of a particular sensitive attribute.

We propose a new metric to measure the degree of a particular attribute by calculating the area of the overlap between two normal distributions. The theoretical maximum value of OA is 1, which means that the model has good fairness. Here are two Gaussian distributions \( x \sim N(\mu_1, \sigma_1^2) \) and \( x \sim N(\mu_2, \sigma_2^2) \). The intersecting part of the value-conditional densities is the probability that both occur in the same condition.

The area formula about the probability density curve for the group is:

\[
F_i(x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dX
\]

If two distributions intersect (when two groups are completely fair, they should be almost coincident) and \( \sigma_1 < \sigma_2 \), they would have two points of intersection, \( m_1 \) and \( m_2 \). We suppose \( m_1 > m_2 \). OA for two values of a specific attribute is:

\[
OA_j = 1 - F_1(m_1) + F_1(m_2) + F_2(m_1) - F_2(m_2)
\]

For the scenario where most attributes have numbers of values, we improve the metric for two values. OA for all values of a specific attribute is:

\[
OA\text{attribute} = mean(\sum OA_j)
\]

After identifying sensitive attributes, it is clear to calculate the fairness degree of a particular property. Take the predicted sentence as an example, the abscissa is the sentence, and the two cases of sensitive attributes (Han and Chinese minorities) are two groups. The fairness degree can be calculated by the formula. When \( OA \rightarrow 1 \), two normal distributions almost coincide, there is no unfairness. When \( OA \rightarrow 0 \), two normal distributions are almost disjoint. The range of OA is from 0 to 1. What is mentioned above are two extreme situations.

III. EXPERIMENT EVALUATION

For our analysis and experiment, we performed a case study on the law case dataset CAIL2018 [15], including 2.68 million short criminal judgment documents, involving 202 crimes and 183 laws, with sentences ranging from 0-25 years to life and death penalty.

We first measure the dataset’s fairness to provide better service for the fairness of algorithms. Then, we decide on an
algorithm at first and use the original dataset to train a model. Next, we can choose and get indicators about model evaluation, such as accuracy for a classification model, RMSE (Root Mean Square Error) for the regression model.

The dataset’s fairness test is ultimately applied to an algorithm for training, and the sensitive attributes are relevant to the user model. Thus we need to specify the model before the experiment. In our experiment, we choose text CNN as our model, defining it as a regression problem. Therefore we used the common indicators as our loss functions, Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-Squared (R2). MSE can evaluate the degree of change of the data. The smaller the value of MSE, the better the accuracy of the prediction model describes the experimental data. RMSE is used to measure the deviation between the observed value and the actual value. It is susceptible to excessively large or excessively small errors in a set of measurements. MAE can better reflect the actual situation of the predicted value error. The range of R2 is between 0 and 1, the closer to 1, the better the regression fitting effect.

We employ these indicators to describe the trained model. Nevertheless, the indeed indicators we used to define sensitive attributes are the gaps between every indicator with itself in two experiments.

**A. EXPERIMENT DESIGN**

Here are the detailed steps of our experiment. First, we choose an initial set of sensitive attributes A, consisting of name, gender, region, color, and year related to dataset characters. According to legal experience, a case that occurs, whatever in which year, should not affect its sentence. However, the year may be used as a characteristic basis to predict the sentence in the process of model training. When this happens, we call “year” a sensitive attribute.

Second, after data preprocessing, do the benchmark experiment without changing the features of the dataset. Then the evaluation indicators gained, the related indicators of benchmark experiment can be seen in Table 1, and a trained model M (BM) gained as well. In the step of attribute substitution, we experiment respectively for the sensitive attribute from A. For instance, we assume that the attribute “year” has eighteen different values (from 2000 to 2017). If a case occurred in 2017, we randomly change the attribute “year” into another value from its value-set, like 2003. Use the former model M (BM) to predict the length of the sentence about the changed case. Then we can get another set of evaluation indicators about the initial sensitive attribute set, MSE, RMSE, MAE, R2, as shown in Table 1.

Table 2 shows the impact of attributes from the initial sensitive attribute set to the model’s predictions. If the value of any indicator is bigger, the attribute is more likely to be sensitive. In Table 2, the experiment about gender does not influence the predictions from the model, and the gaps of all indicators about company attribute approaches to 0. Both gender and company are not sensitive attributes. The region, color, and year are sensitive attributes. Another interesting phenomenon can be seen that if sensitive attribute A’s MSE is bigger than attribute B’s, the other indicators of A will be bigger than B’s corresponding indicators.

Then, setting thresholds for every indicator (or setting a threshold for one indicator only for the regulation we said before), like \( \text{threshold}(\text{MSE}) = 0.1 \). The final sensitive attribute set for the criminal judgment document dataset includes name, color, region, and year.

After identifying the set of sensitive attributes, we can find its distribution in Figure 4. The number of groups is equal to the number of distinct values for the attribute “year”. The graph shows the relationship between different years of imprisonment. As an almost universal rule, the same crime in different years should be sentenced to the same length of years, while the fact is unsatisfactory in terms of fairness.

According to the distribution of each group, we use Equations 2 and 3 to calculate OA for attribute “year”, which is
0.7962 in Figure 2. It shows the OAs of sensitive attributes from the judicial dataset. Attribute “name” is the most unfair attribute for its lowest OA. Figure 3 presents partial data of different years for comparison, which is OA about two groups. What is interesting in Figure 3 is that as the year increases, the OA of “2001” and “year” have steadily decreased. The OA between 2001 with 2017 is only 0.5137. It means that the trend about the sentence of cases changed a lot, and the fairness of attribute “year” for 2017 is half of 2005’s. If the year is closer to 2001, the value of its OA is closer to 1. If two years are similar, their OAs with 2001 are similar. It means the two groups are almost fair.

**B. EXPERIMENT RESULTS**

This step wants to help people to know more detail about the dataset. After calculating the fairness degree about a particular attribute by OA, we can observe the influence of the attribute on data from the overall view. Visualization is an easy way for scholars to understand the label’s trend for a specific property, then find the regulations.

To present the relationship between sensory attributes and result (the sentence is the result in our experiment), we construct a box plot (Figure 5). Box plot can display discrete distribution of data effects. It captures the median, 25% of the distributions, 75% of the distributions, as well as the minimum and maximum values. What can be seen in this chart is the phenomenal decline of the sentence in general as the year increases from the mean line, with some fluctuations for some years.

**IV. RELATED WORK**

There are various definitions of fairness. In previous studies [5], [6], we can find an overview of fairness research and their comparisons. Dwork et al. [10] proposes fairness through awareness based on the idea that individuals should be treated equally. As a comparison, Kusner et al. [16] and Grgic-Hlaca et al. [17] focus on unawareness fairness, defining fairness as the absence of the protected attribute in the model features. Besides, existing studies also provide many other perspectives to investigate fairness. For example, Williamson and Menon [18] generalizes some existing proposals and proposes a new definition of fairness, enforcing that the expected losses (or risks) across each subgroup induced by the sensitive feature are commensurate. Dimitrakakis et al. [19] argues that contemporary notions of fairness in machine learning need to incorporate parameter uncertainty explicitly and introduced bayesian fairness as a suitable candidate for fair decision rules.

In addition to the generalized definition of fairness, there are algorithm-specific definitions and domain-specific definitions. About the former, Jabbari et al. [20] defines fairness in reinforcement learning as an algorithm never prefers one action over another if the long-term (discounted) reward of choosing the latter action is higher. Chierichetti et al. [21] defines fairness in clustering as each protected class must have approximately equal representation in every cluster and formulate the fair clustering problem under both the k-center and the k-median objectives. For the latter, Yao and Huang [22] focuses on fairness in collaborative filtering which means collaborative filtering methods should not make unfair predictions for users from minority groups.
Y. Li et al.: Empirical Study on Group Fairness Metrics of Judicial Data

Fazelpour and Lipton [23] connects the recent literature on fair machine learning and the ideal approach in political philosophy and works this analysis through different formulations of fairness.

Around the topic of fairness in AI, most of them focus on model and algorithm fairness, whether in deep learning or traditional machine learning.

Similarly, most current fairness metrics focus on model and algorithm fairness. For example, Bose and Hamilton [24] presents a compositional approach to ensuring fairness for graph embedding when building blocks for machine learning models that operate on graph data. Liu et al. [25] shows that unconstrained learning on its own implies group calibration and thinks of group calibration as a byproduct of unconstrained machine learning. There are still a few studies that notice the influence of data quality on fairness. Chouldechova and Roth [26] proposes that the available training data contaminated by bias lead to the fairness problem. Berk [3] presents how to use unfair datasets to train for fairer results in a case study. Ma et al. [27] flags thousands of discriminatory inputs that can cause fairness violations but only treats them as test cases to enhance their model. However, no one proposes a method or takes operation on a specific dataset to judge its fairness. Besides, Kallus and Zhou [28] notices residual unfairness on datasets and proposes metrics to evaluate it. Still, these study only focus on datasets censored by humans for fairness adjustment but remains unfair in practice. It is easy to say that discrimination exists on datasets involving race attributes, but finding out the prejudice on other area datasets is for further study.

The attribute is a crucial link between data and algorithm, the sensitive attribute particularly. In many studies, sensitive attributes serve for algorithm fairness. Menon and Williamson [29] related the tradeoff between accuracy and fairness to the alignment between the target and sensitive features’ class probabilities. We also study data fairness based on sensitive attributes, and the final goal of data fairness is to serve algorithm fairness. Noted that sensitive attributes are defined later in the paper, we focus on data characters instead of personal attributes. There has been much recent work on fair algorithms. Thus we use algorithm-related research and methods to promote the research of data fairness conversely.

V. CONCLUSION

Our framework is an innovative method for studying dataset fairness. It helps researchers to address the problem of bias from AI software systems. We first propose the metrics which can be applied to multi-value attributes. With experiments across the dataset in the judicial field, we have proved unfairness exists in the data. It means researchers need to address the problem of unfairness about the dataset before studying algorithms’ fairness. Our empirical results and value-related figure show how OA can measure group fairness with a specific multi-value attribute and measure fairness about a sensitive attribute in a dataset.

In the experiment, we calculated many times, removed the extreme value to get the average value, and observed whether there is an offset value to avoid errors caused by misoperation to eliminate the threat of internal effectiveness. To eliminate external threats of validity, the data set we selected is from the actual criminal legal documents published by the China Judgments Online to ensure the reliability and credibility of the data source. In addition, many of us participate in constructing the code to avoid the threat of construction effectiveness. These works ensure the reliability of our experimental results.

In the future, we plan to improve the dataset’s fairness using our criterion as a regularizer to improve the machine learning algorithms.
REFERENCES

[1] Q. Gu, W. Cai, S. Yu, and Z. Chen, “An exploratory study on judicial image quality assessment based on deep learning,” in Proc. IEEE 19th Int. Conf. Softw. Qual., Rel. Secur. (QRS), Jul. 2019, pp. 300–305.

[2] Y. Zhao, Y. Feng, Y. Wang, R. Hao, C. Fang, and Z. Chen, “Quality assessment of crowdsourced test cases,” Sci. China Inf. Sci., vol. 63, no. 9, pp. 1–16, Sep. 2020.

[3] R. Berk, “Accuracy and fairness for juvenile justice risk assessments,” J. Empirical Legal Stud., vol. 16, no. 1, pp. 175–194, Mar. 2019.

[4] M. A. Beghouda, “Software engineering teamwork data understanding using an embedded feature selection,” Int. J. Performability Eng., vol. 17, no. 5, pp. 464–472, May 2021.

[5] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, “Fairness in criminal justice risk assessments: The state of the art,” Socio. Methods, vol. 50, no. 1, pp. 3–44, Feb. 2022.

[6] S. Verma and J. Rubin, “Fairness definitions explained,” in Proc. Int. Workshop Softw. Fairness, Y. Brun, B. Johnson, and A. Meliou, Eds., Gothenburg, Sweden, May 2018, pp. 1–7.

[7] P. Gajane and M. Pechenizkiy, “On formalizing fairness in prediction with machine learning,” 2017, arXiv:1710.03184.

[8] M. P. Kim, A. Ghorbani, and J. Zou, “Multiaccuracy: Black-box post-processing for fairness in classification,” in Proc. AAAI/ACM Conf. AI, Ethics, V. Conitzer, G. K. Hadfield, and S. Vallor, Eds., Honolulu, HI, USA, 2019, pp. 247–254.

[9] S. A. Seshia, A. Desai, T. Dreossi, D. J. Fremont, and X. Yue, “Formal specification for deep neural networks,” in Proc. 16th Int. Symp. Automated Technol. Verification Anal., USA, Springer, 2018, pp. 20–34.

[10] C. Xiao, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in Proc. 3rd Innov. Theor. Comput. Sci. Conf., 2012, pp. 214–226.

[11] M. Joseph, M. J. Kearns, J. Morgenstern, S. Neel, and A. Roth, “Rawlsian fairness for machine learning,” CoRR, vol. abs/1610.09559, pp. 1–27, Feb. 2016.

[12] J. R. Foulds and S. Pan, “An intersectional definition of fairness,” CoRR, vol. abs/1807.08362, pp. 1–14, Jun. 2020.

[13] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4066–4076.

[14] F. L. Y. Qu and X. Chen, “LAL: Meta-active learning-based software defect prediction,” Int. J. Performability Eng., vol. 16, no. 2, pp. 203–213, Feb. 2020.

[15] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, and H. Wang, “Cail2018: A large-scale legal dataset for judgment prediction,” CoRR, vol. abs/1807.02478, pp. 1–4, Jul. 2018.

[16] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” CoRR, vol. abs/1703.06856, pp. 1–18, Mar. 2017.

[17] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, “The case for process fairness in learning: Feature selection for fair decision making,” in Proc. NIPS Symp. Mach. Learn. Law, vol. 1, 2016, p. 2.

[18] R. Williamson and A. Menon, “Fairness risk measures,” in Proc. 36th Int. Conf. Mach. Learn., vol. 97, Jun. 2019, pp. 6786–6797.

[19] C. Dimitrakakis, Y. Liu, D. C. Parkes, and G. Radanovic, “Bayesian fairness,” in Proc. 34th Int. Conf. Mach. Learn., Aug. 2017, pp. 1617–1626.

[20] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, “Fair clustering through fairlets,” CoRR, vol. abs/1802.05733, pp. 1–13, Feb. 2018.

[21] S. Yao and B. Huang. “Beyond parity: Fairness objectives for collaborative filtering,” CoRR, vol. abs/1705.08804, pp. 1–10, May 2017.

[22] S. Fazelpour and Z. C. Lipton, “Algorithmic fairness from a non-ideal perspective,” in Proc. AAAI/ACM Conf. AI, Ethics, SoC., Feb. 2020, pp. 57–63.

[23] A. Bose and W. Hamilton, “Compositional fairness constraints for graph embeddings,” in Proc. 36th Int. Conf. Mach. Learn., Jun. 2019, pp. 715–724.

[24] L. T. Liu, M. Simchowitz, and M. Hardt, “The implicit fairness criterion of unconstrained learning,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4051–4060.

[25] A. Chouldechova and A. Roth, “The frontiers of fairness in machine learning,” CoRR, vol. abs/1810.08810, pp. 1–13, Oct. 2018.

[26] P. Ma, S. Wang, and J. Liu, “Metamorphic testing and certified mitigation of fairness violations in NLP models,” in Proc. 25th Int. Joint Conf. Artif. Intell., 2020, pp. 458–465.

[27] N. Kallus and A. Zhou, “Residual unfairness in fair machine learning from prejudiced data,” in Proc. 35th Int. Conf. Mach. Learn., Jul. 2018, pp. 2439–2448.

[28] A. K. Menon and R. P. Williamson, “The cost of fairness in binary classification,” in Proc. Conf. Fairness, Accountability Transparency, 2018, pp. 107–118.

YANJUN LI was born in Hebei, China, in September 1981. Since 2012, he has been engaged in the research of software testing laboratory accreditation management and software testing standardization in CNAS, mainly studying the system testing and evaluation in the field of judicial big data. Since 2014, he has been engaged in software testing and evaluation research at the School of Posts and Telecommunications. His research interests include software testing and evaluation, software testing laboratory accreditation management, and intelligent justice.

HUAN HUANG was born in Guizhou, China, in June 1998. She received the bachelor’s degree in engineering from Central South University, in 2020. She is currently a Research Assistant with the Shenzhen Research Institute of Nanjing University. Her research interests include software engineering, data analysis, and artificial intelligence.

XINWEI GUO received the master’s degree in software engineering from the School of Software Engineering, Beijing University of Technology, in 2016. He is currently working as the Director with the Information Evaluation Center, China Justice Big Data Institute, responsible for the quality standard formulation and system evaluation of the court information systems. He has presided over or participated in compiling more than ten national standards and court industry standards and specifications. His research interests include artificial intelligence testing, big data testing, other testing technologies, and the research of quality standards of court information technology. He is a Technical Committee Member of the Software and System Engineering Sub-Technical Committee, National Information Technology Standardization Technical Committee.

YUYU YUAN received the M.S. degree from the University of Electronic Science and Technology of China and the Ph.D. degree from the Research Institute for Fiscal Science, Ministry of Finance, China. She is currently a Professor with the Beijing University of Posts and Telecommunications, Beijing, China. She is also an Expert at the International Standards Workgroup. Her research interests include software quality, trustworthy service, and software testing.