International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-13

Article

Judicial review and remedies for algorithmic bias in the application of environmental law

Bona Song[1,*]

[1] School of Marxism, Jining Normal University, Ulanqab, Inner Mongolia, 012000, China
* Correspondence author: 302405@jnnu.edu.cn

Abstract: In this paper, the definitions of bias and fairness in deep learning are systematically sorted out, and the specificity of algorithmic bias in environmental law application scenarios is clarified. An algorithmic fairness testing method EIDIG based on gradient search is proposed for uncovering and correcting algorithmic bias.EIDIG adopts the gradient of the model output to replace the gradient of the loss function, which reduces the computational load of the model. Combined with the clustering algorithm to generate diverse individual bias samples as inputs for the next stage, the global search is completed. In the local search, the generated individual bias samples are imported and repeated detection is performed around them. The experimental results show that compared with the comparative fairness testing methods such as ADF and AEQUITAS, the number of bias samples generated by EIDIG, the success rate, and the generation efficiency are all improved to different degrees, and the time taken by EIDIG to generate 1,000 bias samples is only 43.90% and 88.06% of that taken by AEQUITAS and ADF, and EIDIG has achieved a leading position in improving the original model's fairness by achieving leading performance. Finally, this paper proposes judicial remedy strategies such as adjusting the civil law tort liability framework and introducing the public interest litigation system. It provides theoretical support for realizing the rule of law guarantee for environmental disputes in the era of artificial intelligence.

Keywords: Environmental law application; Algorithmic bias; Fairness test; EIDIG; Gradient search; Judicial review and remedy

1. Introduction

As China strengthens its regulation of environmental protection, artificial intelligence algorithms have been widely used in the application of environmental laws, such as remote sensing to identify illegal discharges and air quality testing [1-2]. However, the application of intelligent algorithms also brings about algorithmic bias. Algorithmic bias is different from traditional bias, which has the characteristics of concealment, granularity, structure, polarization and ambiguity, making it difficult to be identified and regulated by people, and increasing the difficulty of regulation [3-4]. The main criteria used in the tradition for bias identification are procedural fairness, data bias, etc., but such criteria can not effectively identify algorithmic bias, in this context, judicial review and relief has become the main direction of development.

Judicial review, as an important mechanism for safeguarding civil rights, needs to clarify the review standard when dealing with algorithmic bias, and it is an important balance between safeguarding civil rights and promoting technological progress [5-6]. Through procedural and substantive review, courts are able to assess the legality, fairness and transparency of algorithms. As for the remedies against algorithmic bias, the current legal remedies in China are mainly categorized into three ways: administrative supervision, judicial litigation and social supervision. In terms of administrative supervision, relevant regulatory agencies should strengthen the management and inspection of algorithm-using units, and the administrative departments can order rectification and punishment based

on the illegal behaviors, which plays a restraining role [7-8]. Judicial proceedings, on the other hand, are an important means to protect the rights and interests of victims. The discriminated person can file a lawsuit involving property rights, fair procedure rights, etc., based on the relevant provisions of the Civil Code, requesting the cessation of discriminatory behavior, elimination of the impact, and compensation for damages [9-10]. And social supervision promotes algorithm transparency through public opinion monitoring and third-party evaluation. The public and civil society organizations can question or suggest algorithm applications, supervise platforms to fulfill their information disclosure obligations, and promote algorithm openness and transparency [11-12]. On this basis, the establishment of industry standards and ethical codes can help regulate algorithm development and use and reduce the risk of discrimination.

The current academic attention to the judicial review of algorithmic bias is relatively small, mainly focusing on the phenomenon of algorithmic bias and its relief. Literature [13] combs through the connotation of algorithmic bias, causes and relief paths, points out that data-driven decision-making may retain or even amplify the discrimination of disadvantaged groups, and emphasizes that the introduction of philosophical analysis is important for the identification and regulation of algorithmic injustice with important moderating significance. Literature [14] analyzes the legal causes of algorithmic bias and remedial dilemma from the framework of constitutional equal protection and civil rights regulations, points out that AI systems are prone to exacerbate structural discrimination and difficult to be held accountable due to training data and feedback loops, and emphasizes that the construction of hierarchical regulation and enforceable fairness standards is a key path to crack algorithmic discrimination. Literature [15] analyzes the historical inequality of AI algorithmic bias stemming from training data and architectural design, points out that facial recognition misclassification and stereotype generation exacerbate racial injustice, and emphasizes that solving algorithmic discrimination is a key prerequisite for ensuring fair application and maintaining procedural justice in the legal field. Literature [16] draws on the psychological traps of human decision-making framework to analyze the pathways of bias generated by data, algorithms, and machine learning in analytics systems, pointing out that blind reliance on system outputs may raise the risk of injustice and lawlessness, and emphasizing that translating remedies to reduce human bias into algorithmic governance tools is a key direction of relief. Literature [17] constructs an analytical framework for the double erosion of algorithmic discrimination through the comparison of multinational cases, points out that opacity makes it difficult to review biased outputs and hinders remedies, and emphasizes that building a comprehensive regulatory system that integrates burden of proof shifting, independent auditing, and participatory governance is a key path to cracking algorithmic injustice. Literature [18] reveals the social harm of algorithmic bias embedded in data and models based on the discussion of COMPAS and other cases, points out that traditional anti-discrimination laws have remedial deficiencies in the face of AI harm, and emphasizes that the construction of a comprehensive framework that integrates transparency, participatory governance, and fair data practices is an urgent path to ensure the goodness of technology. Literature [19] analyzes the legal and technical implications of shifting from correlation to causal inference for circumventing algorithmic bias, and emphasizes that this move can provide protection by seeking a balance between predictive accuracy and compliance remedies. Literature [20] examines the fairness and legitimacy challenges raised by the adoption of automated decision-making tools by government agencies, noting that the intersection of design flaws and implementation policies constitutes a difficulty in remedies, and emphasizing the need to carefully assess the impact of algorithmic power on the allocation of public resources. Literature [21] responded to the criticism of "horse's method" by examining whether algorithmic repair fell into the double dilemma of excessive specialization of methodology and too narrow scope of injury. It was pointed out that although the method was universal, there was a limitation in relief because it only aimed at sufficient and necessary conditions of damage, and it was emphasized that identifying the causal role of algorithmic system in damage was the premise to ensure the effectiveness of repair. Literature [22] systematically reviews the sources and mitigation methods of algorithmic bias by constructing a four-phase framework covering data generation to model deployment, points out the need to integrate the definition of fairness and legal standards from computer science and organizational behavior, and emphasizes that interdisciplinary collaboration is a key path to fill the remedial gaps and achieve legitimate and effective governance. The above research reveals the multidimensional causes and relief dilemmas of algorithmic bias from the multidisciplinary perspectives of technology, law, and ethics, and emphasizes the paths of cross-disciplinary collaboration and layered regulation to achieve effective relief and fair governance, which provides a reference for the field of environmental law application.

In this paper, a fairness testing method based on gradient search is proposed, and the process explores unfair regions in the input space through a global search strategy, and then utilizes local search

to refine the detection of the discovered bias sample domains in order to excavate more hidden bias samples. To enhance the detection efficiency of the algorithm, a bootstrapping mechanism based on gradient information is introduced to make the search more directional. A shift from white-box testing to black-box testing is realized, which enhances the operability of the model in real environmental law application scenarios. Comparative tests of the algorithm were carried out on real environment-related datasets, and ADF and AEQUITAS were selected as comparative models, and multidimensional comparisons were made with this paper's method to verify the usability of this paper's method.

2. Algorithmic bias research under deep learning

2.1. Definition of prejudice

Prejudice is usually defined in two main ways, including direct prejudice and indirect prejudice. Direct bias refers to differential treatment, where a person is intentionally treated differently based on his/her membership in a protected class. Indirect bias refers to differential impact, which affects members of a protected class more negatively, even if a policy appears neutral. Algorithms trained for deep learning that do not use sensitive attributes (i.e., attributes that explicitly identify protected and non-protected groups) are unlikely to produce differential treatment, but may induce unintentional bias in the form of differential impact.

2.2. Definition of fairness

The rapid development of artificial intelligence technology exposes the profound contradiction between fairness and ethical risk. For high-risk scenarios such as credit assessment and judicial recidivism prediction, researchers propose a fairness-aware learning framework, which eliminates features related to sensitive attributes through adversarial training or introduces causal inference to decouple biased paths.

Fairness requires neural network models to avoid differential decision-making based on protected attributes, which centers on eliminating the statistical dependence between prediction results and sensitive features. In the field of criminal justice, COMPAS models incorrectly assess risk for black defendants at a rate several times higher than that of the white population, despite no significant difference in actual recidivism rates. Artificial intelligence bias usually forms a feedback loop between the dataset, the model, and the human-computer interaction, in which the bias propagates and is reinforced, which may ultimately lead to the behavior of the AI model completely deviating from its original goal.

In current research, fairness is categorized into group fairness and individual fairness. Group fairness focuses on the statistically significant consistency between different groups. Statistical consistency is used as a measure of fairness, i.e., two groups have the same probability distribution for a given prediction.

Based on the above theory, conditional statistical consistency is further proposed considering that other attributes of the two groups are controlled to remain consistent. Conditional statistical consistency requires that the probability distributions of the prediction results of different groups should be consistent given the feature X and the sensitive attribute A of the groups. The formula is shown below:

$$P\left(Y=1\middle|X,A=a_1\right)=P\left(Y=1\middle|X,A=a_2\right),\ \forall a_1,a_2\in A,X \tag{1}$$

Scholars have proposed the Equal Opportunity Criterion (EO), which takes into account that different groups may have different distributions with respect to the label Y. Where y is the training set exact label, which is performed by comparing the truth rates of different subgroups:

$$\frac{P\left(\hat{y}=1\middle|z=0,y=1\right)}{P\left(\hat{y}=1\middle|z=1,y=1\right)}=1 \tag{2}$$

Research on dataset fairness is often conducted on the basis of group fairness, while fairness errors generated for datasets as well as models are often centered around individual fairness. Individual fairness emphasizes that AI systems should give the same or highly similar decision outcomes for individuals that differ only in sensitive attributes. Based on this idea as well as the robustness of AI and adversarial attacks, the resulting individual bias samples are widely used in AI fairness assessment studies.

2.3. Sample of Individual Bias

Adversarial samples reveal the sensitivity of deep learning models to input perturbations, where even small, imperceptible changes in the input can lead to extreme misclassification results in the model output. This property exposes the vulnerability of neural networks and drives research on model robustness. However, in addition to the safety issue, research on adversarial samples indirectly reveals the potential risk of AI in terms of fairness, especially in individual bias scenarios where the model may make unfair decisions for some individuals due to irrelevant factors.

The fairness of neural networks can be regarded as a conceptual extension of robustness in the sociological dimension. In AI system design, generalized robustness refers to the ability of a model to maintain functional stability under internal parameter perturbations or external environment variations, while adversarial robustness specifically refers to the ability of a model to defend against adversarial sample attacks under adversarial disturbances. In contrast, the fairness problem focuses on the systematic impact of non-task-relevant attributes on decision boundaries: adversarial samples can lead to model prediction bias through small perturbations in the input space, while fairness deficits manifest themselves as systematic decision biases triggered by protected attributes, both of which reveal the vulnerability of the model in the presence of non-ideal data distributions.

In tasks such as image classification, credit scoring, and face recognition, it has been found that deep learning models may make unfair predictions for specific groups or individuals due to biases in training data or feature sensitivity. For example, a face image may have small variations in skin color or gender features causing the model to produce significantly different classification results across individuals. Similar to adversarial samples, individual bias samples may influence classification decisions by adjusting for specific attributes (e.g., skin color, gender characteristics), while the model itself may still make systematically unfair decisions due to data bias without explicitly learning these attributes. These issues have motivated researchers to explore fairness enhancement techniques to mitigate the phenomenon of individual bias in decision making.

Based on the individual fairness its properties derived from adversarial samples, from the study of adversarial samples, we can not only optimize the robustness of the model, but also further enhance the fairness of the AI system to reduce the unfair decision-making due to unconscious bias. Therefore, the individual bias sample, which is widely used in deep learning fairness assessment and enhancement, has much in common with the confrontation sample definition, and the individual bias sample definition is shown below:

Let the sample instance $x = \{x_1, x_2, \cdots, x_n\}$, where x_i denotes the values of the attribute A_i. If there exists another instance $x' \in \mathrm{I}$ is satisfied:

Related attribute discrepancy: $\exists p \in \mathrm{P}$ such that $x_p \neq x'_p$;

Unrelated attribute agreement: $\forall q \in \mathbb{N}\,\mathrm{P}$ such that $x_q = x'_q$;

Decision discrepancy: $D(x) \neq D(x')$.

Then x is said to be the individual bias sample of model D and x and x' are said to be individual bias sample pairs.

Take the original data of the German Credit dataset as an example. In practical applications, often preprocessing before constructing individual bias samples, this use of raw data, the expression is more intuitive. Therefore, for this data, x and x' make a pair of individual bias sample pairs:

Example x: $[1,6,4,12,5,5,3,4,1,67,3,2,1,2,1,0,1,0,0,1,0,1,0,1]$

Example x': $[1,6,4,12,5,5,3,4,1,35,3,2,1,2,1,0,1,0,0,1,0,1,0,1]$

where x and x' are individual bias sample pairs for the sensitive attribute S (age). They differ only in age, and if a neural network model N gives different predictions for x and x', then x and x' are said to be a pair of individually biased sample pairs for the sensitive attribute S.

3. Gradient search based algorithmic fairness testing approach

Algorithmic bias detection in the application of environmental laws requires a test method that can both efficiently generate individual bias samples and migrate from a white box to a black box, so as to provide technical support to the judiciary in identifying algorithmic bias and reviewing the fairness of decision-making.

3.1. Methodological framework for fairness testing

The fairness test method framework is shown in Fig. 1, and the EIDIG fairness test framework

proposed in this paper mainly consists of two parts: generating individual bias samples and eliminating bias. The focus of this paper is on the generation of individual bias samples, which is divided into two sequential phases, global generation phase and local generation phase, to systematically search for individual bias samples to identify the existence of algorithmic bias for the judiciary.

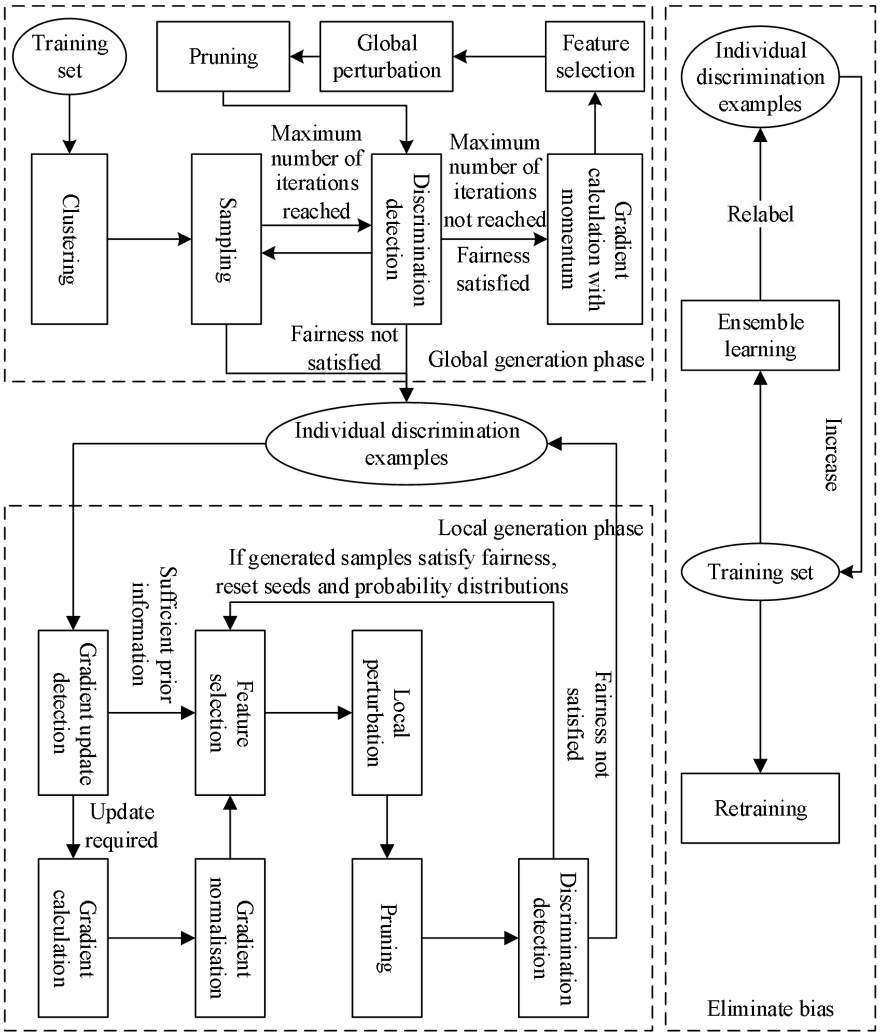

Figure 1. The EIDIG neural network equity test framework

3.2. Gradient Search Improvement

In this section, the paper will demonstrate how a more direct and precise mapping relationship between input perturbations and output changes can be constructed to serve as an effective guideline for generating individual bias samples via perturbations.

Suppose a neural network model N classifies an input sample x into kind p. The previous state-of-the-art ADF algorithm uses the gradient of the loss function over the input features $\nabla_x L(F(x), y)$ to provide guiding information for the generation of individual bias samples. The direction to which the gradient points is the direction in which the function value grows fastest. Thus, if any attribute is perturbed in the direction of $\nabla_x L(F(x), y)$, the prediction error $L(F(x), y)$ will go up, and the model's confidence in the current prediction $F_p(x)$ will fall, and vice versa. The magnitude of the effect on the model's prediction when each attribute of the sample is perturbed is proportional to the absolute value of the corresponding gradient component. Accordingly, any seed input can be subjected to a series of perturbation operations to generate new samples while acquiring new target properties or keeping the original properties unchanged.

Unlike the ADF algorithm, EIDIG directly derives the neural network model and utilizes the gradient $F(x)$ of the model output over the input attributes to establish a direct and accurate mapping

between the input perturbations and the output changes. From a computational point of view, compared to the gradient of the loss function, the gradient of the prediction result eliminates the need to do backward propagation of the loss function. (According to the chain rule, EIDIG does not have to compute the derivative $\frac{\partial \mathrm{L}(\mathrm{F}(x), y)}{\partial \mathrm{F}(x)}$ of the loss function with respect to the model output, nor does it have to compute the gradient $\frac{\partial \mathrm{F}(x)}{\partial \mathrm{H}(x)}(\frac{\partial \mathrm{F}_p(x)}{\partial \mathrm{H}(x)}$ of the model output with respect to the output of the penultimate layer, with the exception of the gradient of the model output with respect to the output of the penultimate layer.)

3.3. Global search phase

In the global search phase, the EIDIG algorithm aims to find a small set of diverse individual bias samples as fast as possible. Therefore, each search iteration perturbs the potential individual bias samples as much as possible in the direction of the decision boundary of the neural network model.

In the EIDIG algorithm, the momentum term takes into account the gradient information of the previous iteration steps to help stabilize the direction of the perturbation, escape from the local optimum point, and find individual bias samples faster. Therefore, the introduction of the momentum term accelerates the iterative search process and improves the success rate of generating individual bias samples in the global search phase.

To generate diverse individual bias samples, EIDIG first performs cluster partitioning on the training set X using the K-Means clustering algorithm, and then samples from the partitioned dataset in a polled manner.

The individual bias samples (g_id) generated in the global search phase will be used as seed inputs for the local search phase.

3.4. Local search phase

In the local search phase, the EIDIG algorithm attempts to quickly generate as many individual bias samples as possible around the small set of individual bias samples generated in the global search phase. The motivation for doing so in this paper stems from considerations of neural network robustness; a well-trained neural network model should be robust, i.e., similar inputs should be mapped to similar or identical outputs.

In the local search phase, EIDIG minimally alters the model's prediction confidence for potential individual bias sample pairs, thus attempting to maintain the model's original predictions. In this way, the perturbed sample pairs still yielded different model predictions while maintaining that they differed only in some of the sensitive attributes, i.e., new individual bias samples were found around the original individual bias samples. Therefore, non-sensitive attributes that have little effect on the model prediction results should be preferentially selected as perturbation objects.

In the local search phase, since each iteration step makes the smallest possible perturbation, the guideline attribute selection and the perturbed gradient information between neighboring iterations are highly correlated or even overlapped with each other with high probability.

3.5. Increased equity

After generating a sufficient amount of individual bias samples, the next step is to enhance the fairness of the model under test, which is the point of generating individual bias samples.

Adversarial training refers to mixing the original training set with the generated adversarial samples and retraining the neural network model with the augmented dataset. According to experiments in the fields of adversarial attacks and testing, adversarial training has been found to be effective in improving the robustness of neural network models. Similarly, in this paper, the individual bias samples generated by EIDIG will be utilized to mitigate the degree of bias of the original model.

3.6. Conversion of white-box testing to black-box testing

Compared with white-box testing, black-box testing methods do not require information about the structure of the neural network model, hyperparameter settings, training methods, etc., but only need to be able to ask the model's prediction results for a certain input. In some cases where the internal information of the neural network is not available, such as when commercial secrets are involved, the white-box fairness test method EIDIG proposed in this paper is not able to compute the gradient

through backward propagation, and therefore cannot be applied. Therefore, the basic idea of converting EIDIG into a black-box testing algorithm, i.e., zero-order optimization, will be given here.

It is worth emphasizing that the EIDIG algorithm requires knowledge of the structure and weights of the neural network model only when computing the gradient of that item. Thus, the key to converting EIDIG to a black-box testing algorithm is to approximate the gradient by relying only on the predictions of the neural network model. The most basic approximation is to estimate the gradient $\nabla_x \mathrm{F}_p(x)$ using a symmetric difference quotient:

$$\nabla_x \mathrm{F}_p(x) = \left[\nabla_{x_1} \mathrm{F}_p(x) \quad \nabla_{x_2} \mathrm{F}_p(x) \quad \cdots \quad \nabla_{x_n} \mathrm{F}_p(x) \right] \tag{3}$$

where n is the dimension of the input sample, which for structured datasets is the number of attributes, and the i th term is:

$$\nabla_{x_i} \mathrm{F}_p(x) = \frac{\partial \mathrm{F}_p(x)}{\partial x_i} \approx \frac{\mathrm{F}_p(x + \delta e_i) - \mathrm{F}_p(x - \delta e_i)}{2\delta} \tag{4}$$

From the above equation, given an input sample, for each attribute of that input, a separate estimation needs to be done, and each estimation requires the neural network model to predict each of the two inputs in the neighborhood of the original input. For an n -dimensional input, the neural network model is then required to predict $2n$ times. Obviously, such a transformation method is required to consume close to $2n$ times the computational resources for the purpose of black-box testing.

In order to improve the computational efficiency, many black-box adversarial attack methods take this as the core of their research, and propose some approximations that significantly reduce the number of prediction times, and they are all able to extend to the EIDIG algorithm proposed in this paper.

4. Experiments and analysis of results

4.1. Experimental environment and experimental data set

The hardware implementation of the algorithms in this study is mainly based on GPUs and the software implementation is mainly the TensorFlow deep learning framework. The following comparison models were used for comparison tests.

AEQUITAS: The AEQUITAS algorithm improves the THEMIS algorithm by employing a two-stage generative framework.AEQUITAS is a systematic generation algorithm.In the first stage, AEQUITAS generates a set of random biased samples in the input space as seeds. In the second stage, AEQUITAS devises three different strategies - random, semi-directed and fully directed - to update the probabilities used to guide the selection of perturbation attributes by randomly adding perturbations to the unprotected attributes of the samples, thus searching for more biased samples around the seed inputs found in the first stage.

ADF: A scalable gradient-based algorithm (ADF) that efficiently generates biased samples for deep learning models through adversarial sampling. The algorithm uses a combination of a global generation phase and a local generation phase to systematically search the input space for biased samples guided by the gradient.

In consideration of the fact that the biasiness of tabular data is more common in real-life applications and that there are relatively few text-based public datasets, three public fairness test datasets, Census Income, German Credit, and Bank Marketing, are selected as the experimental datasets in this study in order to facilitate the evaluation of the experimental results.

4.2. Comparative Experimental Results

The performance of the EIDIG model is first compared with the performance of models such as AEQUITAS and ADF for tabular datasets. The specifics of sample generation for each model are shown in Table 1.

EIDIG and ADF are significantly better at generating samples than AEQUITAS, while the EIDIG model is slightly better at generating samples than ADF.In terms of the number of samples, the number of unique samples generated by EIDIG and ADF is several times the number of unique samples generated by AEQUITAS, and the corresponding number of biased samples are several times the number of biased samples generated by AEQUITAS, respectively. From the perspective of the success rate of generating biased samples as measured by the ratio of the number of biased samples to the

number of unique samples generated, the success rate of EIDIG on the Age protected attribute of the Census dataset is 52.41%, which is slightly higher than that of ADF's 52.13%, and is much higher than AEQUITAS's success rate of 10.73%.

Table 1. The specific conditions of the samples generated by each model

Data set	Protected property	AEQUITAS		ADF		EIDIG	
		The only sample number	Discriminatory sample number	The only sample number	Discriminatory sample number	The only sample number	Discriminatory sample number
Census	Age	57226	6141	491364	256140	592133	310311
Census	Race	53657	4736	344256	140232	402099	156212
Census	Gender	33422	2863	260142	47496	332566	55341
Bank	Age	31740	9039	698164	363949	753141	394162
Credit	Age	99411	39039	398048	233245	450173	265361
Credit	Gender	33577	4963	312054	72386	383145	81149

On this basis, we also analyze the performance effect of each model in the global generation stage and local generation stage separately. Set each model to generate 1,000 samples (set to 600 samples on the CREDIT dataset) in the global generation stage respectively, and calculate the proportion of biased samples among the samples generated by each model in this stage.

The comparison results are shown in Table 2, where it can be observed that EIDIG generates the highest number of biased samples in the global generation stage, slightly higher than ADF by 5.29% on the Age protected attribute on the Census dataset, and on average by a factor of 5.3 compared to AEQUITAS. We believe this demonstrates the effectiveness of the gradient-based guided search approach during global generation.

Table 2. Discriminatory samples generated at the global generation stage

Data set	Protected property	AEQUITAS	ADF	EIDIG
Census	Age	107	643	677
Census	Race	96	478	484
Census	Gender	43	340	368
Bank	Age	39	807	793
Credit	Age	164	614	632
Credit	Gender	52	474	483

In order to be able to accurately compare the performance of each model in the local generation phase, this experiment uses the same set of randomly selected biased samples as input for each model in the local generation phase and generates 1,000 samples respectively, so as to be able to accurately assess the effect of the local generation phase. The comparison results are shown in Table 3, from which it can be seen that EIDIG still has a significant performance advantage over AEQUITAS in the local generation stage, generating 570 more biased samples than AEQUITAS on the Age protected attribute of the Census dataset; compared to ADF, it is an improvement of about 5.29%, but in the experiments of the partially protected attribute the performance on the effect is close to that of ADF.

Table 3. Discriminatory samples generated at the local generation stage

Data set	Protected property	AEQUITAS	ADF	EIDIG
Census	Age	107	643	677
Census	Race	96	478	484
Census	Gender	43	340	368
Bank	Age	39	792	793
Credit	Age	164	614	632
Credit	Gender	52	474	483

In order to better analyze the performance gap between EIDIG and ADF, we conducted a comparative experiment on the time required for these three methods to generate the same number of biased samples as well, and the comparative results of the experiments are shown in Table 4.EIDIG has a significant advantage in generating biased samples, and in the case of generating the same 1,000 biased samples, the average time spent by EIDIG is only 43.90% and 88.06% of the time required by AEQUITAS and ADF, which indicates that EIDIG's selection of perturbation objects can obtain more

prejudicial samples in one iteration, and the optimization processing of the algorithm can also effectively reduce the time complexity of the algorithm.

Table 4. Compare the time of producing 1000 discriminatory samples

Data set	Protected property	AEQUITAS	ADF	EIDIG
Census	Age	175.34	67.16	57.26
Census	Race	131.72	68.20	60.42
Census	Gender	163.37	79.38	67.68
Bank	Age	195.75	110.59	98.03
Credit	Age	181.44	70.23	61.21
Credit	Gender	164.41	108.93	99.66
	Mean	168.67	84.08	74.04

4.3. Dynamic fairness test results

Accuracy (ACC for short): $ACC = (TP + TN) / (TP + TN + FP + FN)$, where TP denotes the number of samples in the positive sample that were actually predicted to be classified as positive; TN denotes the number of samples in the negative sample that were actually predicted to be classified as negative; FP denotes the number of samples in the negative sample that were actually incorrectly predicted to be positive; FN denotes the number of samples in the positive sample that were actually incorrectly predicted to be negative. the number of samples that were actually incorrectly predicted as positive samples; and FN the number of samples that were actually incorrectly predicted as negative samples in positive samples. The matrix consisting of these four metrics is called the confusion matrix and reflects the rate at which the classifier accurately identifies true positives and false negatives. This metric may seem to accurately reflect the performance of a classifier, but it is important to note that the value of ACC is a judgment made only for the current set of input data. This can easily be skewed by the data, and therefore it is inaccurate to observe accuracy alone when actually judging how well a model is predicting.

TPR, also known as recall, is one of the performance metrics of a binary classification model. It indicates the proportion of true positive examples that the model is able to identify. Specifically, TPR is equal to the number of true positive cases (TP) divided by the total number of all true positive (TP) and false negative (FN) cases:

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

where TP denotes the number of true positive examples that are correctly predicted as positive examples and FN denotes the number of true positive examples that are incorrectly predicted as negative examples. In binary classification problems, the predictions are usually categorized into positive and negative examples. When the model predicts a true positive example as a positive example, it is called a single correct prediction (TP). When the model predicts a true positive example as a negative example, it is called one false prediction (FN).TPR measures the ability of the model to identify all true positive examples, i.e., for true positive examples, how many positive examples the model is able to predict as positive examples.The higher the TPR, the more true positive examples the model is able to identify, i.e., the higher the recall. The recall of the model is very important in some applications, such as medical diagnosis, security monitoring, etc., because the cost of missed diagnosis or under-reporting in these applications is often very high. It is important to note that recall and accuracy (are two different metrics. Accuracy is the proportion of correct classifications by the model, while recall is for real positive examples. In some cases, the model may only be able to identify a small number of true positive examples, but the accuracy of the predicted results is very high, in which case the recall will be low, but the accuracy will be high.

The Census Income dataset is first divided into Cluster A and Cluster B based on gender, with Cluster A representing males and Cluster B representing females. Then the two groups of data are divided into training set and test set, in this paper, the commonly used 80%-20% division principle is adopted, that is, 80% of the data is used for model training, 20% of the data is used for model testing.

To address unfair or biased behaviors exhibited between different groups, equal opportunity fairness metrics are commonly used. To determine the dynamic fairness of the model in this paper, the definition of S-TPR is used.TPR is the percentage of actual positives that are correctly predicted as positive. In this paper, a positive prediction is assumed to result in a benefit.TPR can be interpreted as the percentage of people who legitimately benefit from the model. In the case of equal opportunity, the

model is considered fair if the TPR is equal for both groups. In this paper, the variable error is defined as the difference between the TPR of the two cohorts, and the closer the error is to 0, the fairer the model is considered to be. the error trend graph is shown in Fig. 2, and it can be seen that at the beginning of the test, the value of error fluctuates considerably. As the model continues to be trained, the error value gradually converges to 0. This indicates that the model has achieved good results in improving the fairness of equal opportunity.

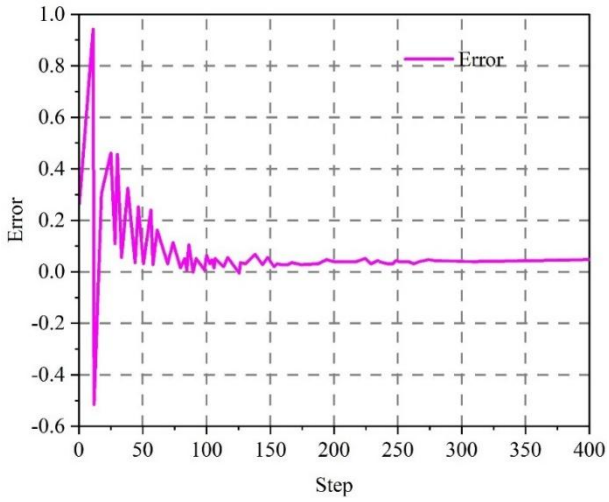

Figure 2. Error chart

In this paper, the Census Income dataset is categorized into two groups by gender, i.e., Cluster A represents males and Cluster B represents females. Each time the environment is adjusted, the metrics related to fairness also change dynamically, such as the accuracy (acc) value. Figure 3 shows the change in acc value over time before the training of the intelligentsia, and this dynamic monitoring method can help to understand how the fairness metrics change in a dynamic environment.

It can be noticed that there are many overlapping regions in the curves for the male and female clusters, which are consistent with the definition of s-Accuracy in Equal Opportunity, indicating that the model under test is fair under this definition. However, as a whole, the data distribution changes as time advances, and it can be found that the male cohort continues to have an overall higher average accuracy than the female cohort.

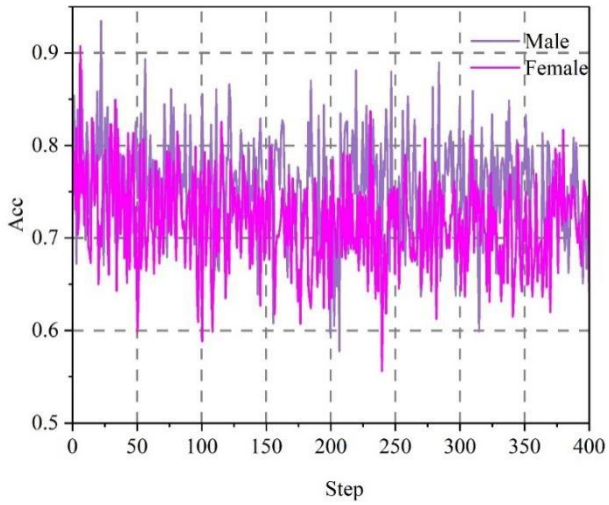

Figure 3. The first two group accuracy changes are at any time

The plot of the accuracy of the two clusters over time after the training of the intelligences is shown in Fig. 4, and compared with Fig. 3, there is a significant increase in the overlapping regions compared with the pre-training period. This means that the model has a significant improvement for the equal opportunity fairness s-Accuracy metric.

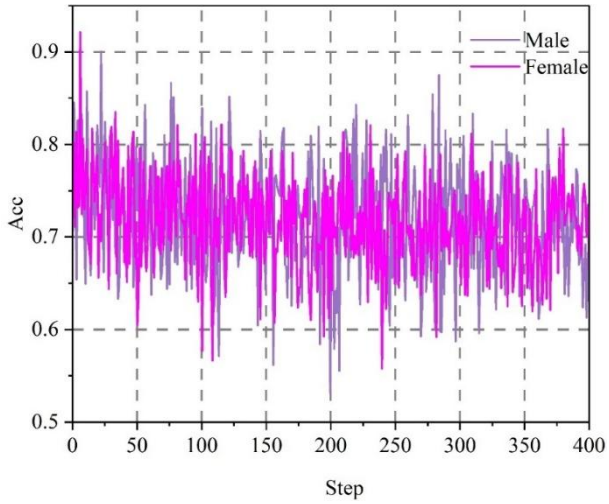

Figure 4. The two group accuracy changes at any time after training

5. Paths to judicial redress for algorithmic bias

The ex ante regulation of algorithmic bias focuses on risk prevention, but because of the hidden and extensive nature of algorithmic bias, it may still cause infringement on users, therefore, it is necessary to introduce the ex post relief system of algorithmic bias, so that when the individual rights and interests of the individual are damaged, they can get effective judicial relief.

Algorithmic bias embodied in the subject plurality, algorithmic technology opaque, algorithmic autonomous computing and other characteristics, to the traditional civil law tort liability theory to bring challenges, there is a need to adjust and broaden the basis of the traditional civil law tort liability framework, to clarify the main body of the algorithmic bias in the assumption of responsibility, so that it is better applied to the algorithmic bias infringement cases. In view of the natural information superiority status of algorithm designers and users, the principle of presumption of fault can be applied. When an incident of algorithmic bias occurs, the platform operator is required to prove that it does not have a causal relationship with the damage caused by the algorithmic bias; otherwise, it is presumed that the behavior of the said subject has a causal relationship with the result of the damage caused by the algorithmic bias.

The behavior of algorithmic bias infringing on consumer rights and interests is insidious and difficult for ordinary consumers to detect. At the same time, because algorithmic bias is characterized by universality, the infringement of individual consumer rights and interests is often small, so most consumers will choose to give up prosecution. However, from the perspective of the overall interests of society, if the algorithmic bias behavior is allowed to occur, it will affect the market economic order on the one hand, and lead to a huge loss of social interests on the other. Therefore, when individuals are negligent in seeking remedies, it is necessary to explore the path of after-the-fact remedies from the framework of civil public interest litigation. The introduction of algorithmic bias public interest litigation system can effectively regulate algorithmic bias behavior from the outside. From the scope of filing, initiation procedures, litigation process, from litigation professionals to algorithmic technology professionals to link, from the legal professional level to build algorithmic bias relief public interest litigation procedures, so that public interest litigation has become a powerful guarantee of algorithmic bias relief, to better ensure the legitimate rights and interests of the victims.

6. Conclusion

Algorithmic bias is not a by-product of technological neutrality, but a continuation and reinforcement of the imbalance of interests in traditional environmental governance laws in the digital form. In this paper, we address the above challenges and propose EIDIG, a sample algorithmic bias testing framework for fairness violations applicable to environmental law scenarios.

On three publicly available fairness test datasets, Census Income, German Credit, and Bank Marketing, EIDIG is compared with the more advanced fairness test methods, ADF and AEQUITAS. The experimental results show that the success rate of EIDIG is 52.41% on the Age protected attribute of the Census dataset, which is higher than that of the ADF algorithm and the AEQUITAS algorithm, and this paper's algorithm exhibits a superior success rate on different datasets and protected attributes.

The overlapping areas of cluster A and cluster B accuracy over time in the Census Income dataset after algorithm training are significantly more than before training. It shows that the model has a

positive contribution to equal opportunity fairness, and can realize the fairness adjustment of algorithmic bias samples in environmental legal scenarios to maintain the fairness of the tested model.

Algorithmic bias is not an airy grave, it is an appendage of the rapid development of information technology in the era of big data. Algorithmic bias infringes on the legitimate rights and interests of the general public, and we should cautiously look at the negative impacts it brings, and actively promote the updating of judicial review standards, the enhancement of algorithmic transparency, and the establishment of technical relief mechanisms to protect the legitimate rights and interests of citizens.

About the Author

Bona Song was Born in Ulanqab, Inner Mongolia, China in 1985.She received her Doctor of Laws (LL.D.) degree from China University of Political Science and Law. She is currently working at Jining Normal University. Her main research interests including environmental and resources protection law, constitutional law, jurisprudence and legal practice.

References

1. Mandu, L. A., & Racoveanu, A. C. (2025). Artificial Intelligence and environmental law enforcement: can technology improve compliance with climate law?. Law Rev., 15, 267.
2. Chukaieva, A., & Matulienė, S. (2023). Possibilities of applying artificial intelligence in the work of law enforcement agencies. Scientific Journal of the National Academy of Internal Affairs, 3(28), 28-37.
3. van der Zee, E. (2025). Will new technologies improve environmental decision-making? Strategies to mitigate biases in EU legislation. Law, Innovation and Technology, 1-21.
4. Trehan, A. (2025). Bias in green AI addressing disparities in data and algorithms. In Advancing Social Equity Through Accessible Green Innovation (pp. 63-76). IGI Global Scientific Publishing.
5. Waldron, J. (2022). The core of the case against judicial review. DIREITO GV L. Rev., 18, 1.
6. Lustig, D., & Weiler, J. H. (2018). Judicial review in the contemporary world—Retrospective and prospective. International Journal of Constitutional Law, 16(2), 315-372.
7. Kazim, T., & Tomlinson, J. (2023). Automation bias and the principles of judicial review. Judicial Review, 28(1), 9-16.
8. ERKAN, F. (2026). Reconfiguring Judicial Legitimacy: Artificial Intelligence, Cognitive Bias, and The Rise of Algorithmic Authority. IJSS, 10(42), 159-220.
9. Kharitonova, Y. S., Savina, V. S., & Pagnini, F. (2021). Artificial intelligence's algorithmic bias: ethical and legal issues. Perm U. Herald Jurid. Sci., 53, 488.
10. Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409.
11. Shneiderman, B. (2016). The dangers of faulty, biased, or malicious algorithms requires independent oversight. Proceedings of the National Academy of Sciences, 113(48), 13538-13540.
12. Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., ... & Kuflik, T. (2022). Mitigating bias in algorithmic systems—A fish-eye view. ACM Computing Surveys, 55(5), 1-37.
13. Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. Philosophy Compass, 16(8), e12760.
14. Cromvelle, D. S. (2025). Algorithmic Discrimination and Equal Protection Law: Legal Remedies For Ai Bias In Automated Decision Making. International Journal of Law, Policy and Scientific Research, 1(1), 1-12.
15. Rankin, S. M. G. (2024). MITIGATING ALGORITHMIC BIAS. Scitech Lawyer, 20(4), 26-32.
16. Edwards, J. S., & Rodriguez, E. (2019). Remedies against bias in analytics systems. Journal of Business Analytics, 2(1), 74-87.
17. Johari, B. (2026). Dual erosion of equality and remedy in algorithmic decision systems. Discover Artificial Intelligence.

18. Kumar, A. B., & Sanjaya, K. (2026). Addressing Algorithmic Bias: Legal Challenges and Solutions. In Leveraging AI for Inclusive and Equitable Development (pp. 253-276). IGI Global Scientific Publishing.
19. Xiang, A. (2020). Reconciling legal and technical approaches to algorithmic bias. Tenn. L. Rev., 88, 649.
20. Sun, M., & Gerchick, M. (2019). The scales of (algorithmic) justice: Tradeoffs and remedies. AI Matters, 5(2), 30-40.
21. Doyle, C., Alvarez-Garcia, M., Tracey, P., Grill, G., Whitney, C., & Chambers, L. M. (2024). Reparations of the horse? Algorithmic reparation and overspecialized remedies. Big Data & Society, 11(3), 20539517241270670.
22. Hickman, L., Huynh, C., Gass, J., Booth, B., Kuruzovich, J., & Tay, L. (2024). Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. Journal of Applied Psychology.