Machine Judges Reduce Sentencing Bias?
A Computational Social Science Evaluation

Mingyang Chen[1*], Zhipeng Wu[2*]
[1]Faculty of Social Science, University of Macau,
Avenida da Universidade, Taipa, 999078, Macau, China
[2]Elite Institute of Engineering, Chongqing University,
No. 9 Liangjiang Avenue, Yubei District, 401120, Chongqing, China
mc55649@um.edu.mo; 202530131014T@stu.cqu.edu.cn

Abstract

Machine learning models have been applied in many criminal justice decisions, and prior research has proved that machine learning models can reduce biases if they are blind. However, prior research focuses on classification tasks in criminal justice. Regression tasks' disparity is much more difficult to be evaluated. Prior research on sentencing bias evaluation only focuses on systematic biases and ignores the individualized biases in cases. In this study, we focus on the sentencing task. We propose a new method to evaluate whether an individual case is biased based on comparing it with all other cases according to Treating Like Cases Alike. We collect all 238,419 theft cases from CJO and extract the legal factors and sentencing results. 159,699 cases are used for building a machine learning model, and we test our model' ability of reducing biases on the rest 78,720 cases. We use XGBoost to train our model. By employing the method, we find if all judges are replaced by machine learning models, the probability of being sentenced an unfair result is 35% lower; if cooperating with judges, 55% biased cases can be sentenced in a more fair way. Machine learning models can reduce individualized biases.

Background

Sentencing prediction models care more on accuracy, which means the smaller differences between the real judge's and machine's decision, the better (Xiao et al. 2018; Hofman et al. 2021). The same applies to other prediction tasks in criminal justice system, and even in other fields. Most research aims for predict same results as real judges, even in other criminal justice sentence field like the bail (Kleinberg et al. 2017). However, judges have biases. If a machine learning (ML) model predicts a sentencing results as a real judges, then there may be some problems: the model may just learn from judges' biases and produce unfair results continuously. Nevertheless, many researchers have found that ML does not have the ability to predict a same results as human judges, and at first they considered that this is because the algorithm is not good enough (Xiao et al. 2018). However, until now, no ML model can predict the same as a human judges. Even a human judges can not predict the same as another. Later, researchers noticed that this gap may be a result of human biases and they try to use other methods to reduce these biases (Zhou et al. 2022). Except for building a better model, researchers also introspect on prior research and discover that ML models, in most cases, only receive legal factors as features for prediction, which have the potential to reduce criminal justice decision biases (Kleinberg et al. 2017; Goel, Rao, and Shroff 2016; Lin et al. 2020), although the data provided contains a series of discrimination or biases, and in order to fit these biased result, ML model may use complex calculus where biases arise Hu et al. (2025). However, other researcher argue that, during the machine learning proceeding, all biases are considered as noise in ML models, and models only try to capture the rule between legal factors and expecting outcomes from biased decisions made by human beings (Kleinberg et al. 2017). For example, Kleinberg et al. (2017) use bail data from New York, and find that ML model produces a better result than real human judges, which means machine judges successfully predict the recidivism. Goel, Rao, and Shroff (2016) also use the data from New York, and find a similar result that by employing ML models, police can stop less, hit more, and reduce more race discrimination. However, Dressel and Fraid (2018) find that ML models perform no differences from lazy people in predicting recidivism. Thus, Lin et al. (2020) replicate the research, and find that if ML models are provided with rich risk factors, or called as legal factors in our context, they can outperform the human beings. Overall, these findings all find that, in certain circumstance, which means rich legal factors are available for ML models and extra-legal factors are excluded, machine judges may perform better than human judges. It seems that this assertion, that machine learning models only use legal factors as prediction features so they can not produce biased results, is self-evident. Nevertheless, ML models employ non-linear models, and create many interactions, or other changes on data in order to fitting the results of hu-

[*] These authors contributed equally.

man judges. For example, in Hu et al. (2025)'s research, they argue that some features may be an indicator to gender information, and through complex ML models, they can be captured. Thus, further evaluation should be employed.

However, These findings have several limitations: first, all research focuses on classification tasks' performance, e.g., recidivism prediction and bail, which means that the ML models, as well as judges only need to decide yes or no, considered as a much more easier decision compared with regression tasks. In other fields, research also focuses on classification tasks, e.g., microlending decision (Cresswell et al. 2024; Hu et al. 2025). However, this prevents generalizing the conclusion that ML models can reduce biases to the regression tasks. A typical field where ML models are used mainly is sentencing, specifically in China (Tahura and Selvadurai 2022).

Second, there is no suitable tools to identify sentencing tasks' performance. If we want to study on continuous outcomes, especially sentencing, we can not use the method Kleinberg et al. (2017) use as we mentioned in the first point because we do not know the post-hoc results of sentencing. The only method is to use models to estimate whether disparity exists. However, biases occur in various forms, while previous research only focuses on race or other demographic features of offenders. It leads to a question that in all social science research, is it possible for researchers to control all the variables? If not, then there exists an endogeneity problem. This is also why sentencing disparity or criminal justice decision disparoity research is always inconclusive. There are so many factors that may affect sentencing but usually ignored by previous research: Kim, Spohn, and Hedberg (2015) and Wu (2021) use multi-level regression models and found judges themselves produce biases; Enough and Mussweiler (2001) also proves that the prosecutors' sentencing recommendation will make judges produce different results even when they are dealing same cases; Danziger et al. (2011), Heyes and Saberian (2019) and (Ludwig and Mullainathan 2024) also find many factors, including food supply, environment, and even the outlook of offenders other than demographic features will affect final decisions, producing biased sentencing results. What is more, (Ludwig and Mullainathan 2024) find that the outlook of offenders, in their context, the well-groomed and heavy-faced, has a higher impact than other demographic features, and they are highly related to gender issues. Moral institutions also have an impact on criminal justice decisions according to Silver, Ulmer, and Silver (2023). All these findings then make us aware of a problem that researchers can not control many variables. This causes serious endogeneity problem. For example, a social science researcher wants to investigate the sentencing disparity between male and female criminals, but there are too many variables that a researcher can not control all of them, e.g., the food supply, the moral institutions, the environment and even the face of offenders (Enough and Mussweiler 2001; Danziger et al. 2011; Heyes and Saberian 2019; Silver, Ulmer, and Silver 2023; Ludwig and Mullainathan 2024). As a result, the findings that gender contributes to a sentencing disparity may be wrong because in their model they do not control all rival factors, leading to a biased causal effect identification. For instance, researcher may find judges discriminate the male, but in fact, it is because in the sample, most male offenders would have a heavy-faced feature.

Third, these research focuses more on systematic biases based on equality before law philosophy (Beccaria and Voltaire 1876), instead of individualized biases based on Treating Like Cases Alike (Hart 1996). Systematic biases here are defined as the disparity that happens in the whole sentencing system, e.g., most judges sentence black people a longer imprisonment (Baumer 2013), or even the ideology of the whole justice system (Kennedy 1976, 1997). Individualized biases are defined as biases in an individual case compared with other cases. Research on ML model's biases also pay more attention to systematic biases, e.g., Goel, Rao, and Shroff (2016) focus more on the race disparity from a systematic view by comparing black people's stop percentage with white people's, and Larson et al. (2016) also evaluate the race problem using similar methods. Their research is usually conducted as following steps: first, they find a model and produce results; next, they devide cases into two groups and evaluate whether these cases show a different prediction results. Theoretically, sentencing biases can be divided into two types: one is from systematic discrimination (Seniuk 2016), and one is from individual cases, which is usually ignored by social science academia because they are hard to generate a theory. Individualized biases can generate to a systematic one, while a systematic one can contribute to an individual case's sentencing disparity. A typical social science research starts with dividing cases into groups by some extra-legal factors and then observe the difference between groups either using regression models or matching comparisons. But, individualized biases require observing cases one by one and compare them from a micro view. However, it is hard for scholars to answer which case is biased and which one is not, which is a straight answer to individual biases, because it is impossible for scholars to read such many cases and even remember all details. In fact, most social science tools are designed for identifying systematic issues, because that is what they mainly concern. To summary, concentrating on the systematic biases can generate social science theory, but the endogenous issue prevent us to produce reliable evaluation on sentencing. Social science tends to view the disparity from a macro view, but as for defendants, they only care about themselves. Thus, this research only focus on individualized biases. The theory frame in this part is showed in Figure 1. Then, the main problem here is, how can we develop a method to estimate individualized biases, and the method should be a relatively explanable one?

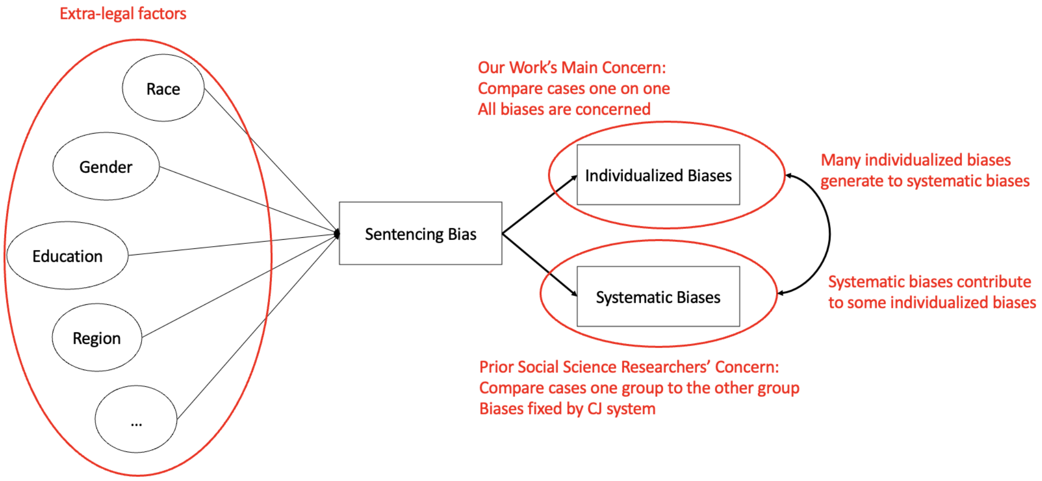

Extra-legal factors
Race
Gender
Education
Region
...
Sentencing Bias
Individualized Biases
Systematic Biases
Our Work's Main Concern:
Compare cases one on one
All biases are concerned
Many individualized biases
generate to systematic biases
Systematic biases contribute
to some individualized biases
Prior Social Science Researchers' Concern:
Compare cases one group to the other group
Biases fixed by CJ system

Figure 1: Two Perspectives on Sentencing Biases

Methodology

China Judgement Online Data

Our China Judgement Online (CJO) data ranges from 2018 to 2021, which is before the Amendment XI to Criminal Law of PRC took effect, and after the Amendment X to Criminal Law of PRC took effect. This makes sure all sentencing decisions are under same legal regulations. Next, we extract all legal factors from the reasoning part. We imitate the method Xiao et al. (2021)'s LawFormer used to extract legal factors. We first build regular expressions and ask legal students to read the judicial documents and ask them whether there is a need to change the regular expressions. Sometimes, judges produce incorrect reasoning or do not provide any reasoning. Thus, there will be some wrong cases, e.g., some cases with no legal factors or some cases with contradict factors. Under the rule built by our legal research assistants, we delete these cases. We thus get a total of 238,419 cases. Then, we go to the facts description parts and use the Task Augmentation method proposed by Li, Qin, and Yu (2022) to extract the theft amount data. We also use a generative LLM model Qwen 3 to help us extract some obviously wrong data, e.g., cases with extremely small amount or the amount is inconsistent with its sentencing length.

Machine Learning Model

159,699 cases are used for building a model, and the rest of cases are used as validation dataset. Among 159,699 cases, 80% are used for training and 20% are used for testing. Our research employs the XGBoost model proposed by Chen and Guestrin (2016):

$$f(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

, where $f(x)$ denotes the XGBoost function, M represents the number of decision trees included in the model, $h_m(x)$ is the m-th tree, and γ_m is the corresponding weight. The loss function is:

$$\ell(y, f(x)) = \frac{1}{2}(y - f(x))^2$$

, where $\ell(y, f(x)$ is the loss function; y denotes the judges' decisions; $f(x)$ denotes the prediction results of XGBoost model. To prevent overfitting, an L2 regularization term is incorporated:

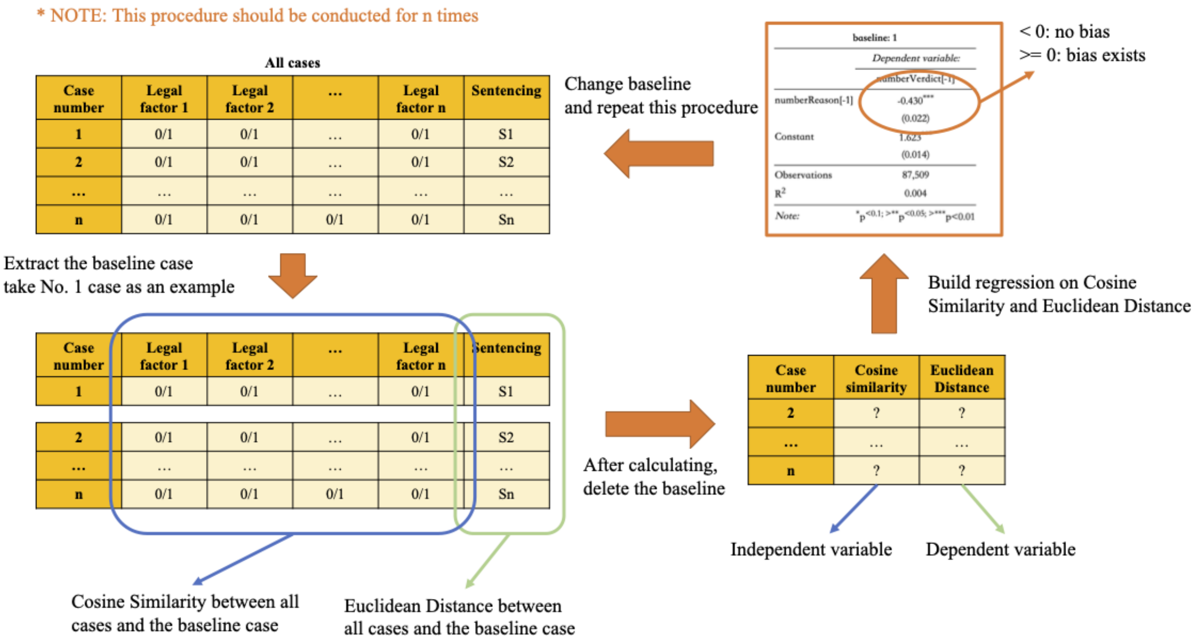
$$\Omega(f) = \lambda \sum_{m=1}^{M} \|h_m\|^2$$

To evaluate our model's performance, we use the CAIL2018 score proposed by Xiao et al. (2018) and MAE score. The CAIL2018 score ranges from 0 to 100. We also employ the LightGBM model as a robustness check and we find that they have similar performance. The MAE of XGBoost model is 3.406 months, and the CAIL2018 score is 80.8. The MAE of LightGBM is 3.347 months, and the CAIL2018 score is 81. Compared with similar models, our performance is good enough to be used for testing the ability of reducing sentencing disparity (Li et al. 2020). Besides, the XGBoost and LightGBM model's MAE is only about 0.37 months.

Bias Identification Method

We then try to find a way, hoping that by using the legal factors only, we can find whether there is a sentencing disparity in each individualized case. We then get an idea from Zhou et al. (2022)'s paper as well as Seniuk (2016)'s idea. The foundation of our method is Treating Like Cases Alike theory in legal academia (Hart 1996, 1958). We then test the ability of reducing biases of machine judges on 78,720 validation cases. Our method is illustrated by Figure 2.

By employing the method illustrated, we can identify which case is biased, compared with all other cases


* NOTE: This procedure should be conducted for n times
Change baseline
and repeat this procedure
< 0: no bias
>= 0: bias exists
Extract the baseline case
take No. 1 case as an example
Build regression on Cosine
Similarity and Euclidean Distance
After calculating,
delete the baseline
Independent variable
Dependent variable
Cosine Similarity between all
cases and the baseline case
Euclidean Distance between
all cases and the baseline case

All cases

Case number	Legal factor 1	Legal factor 2	…	Legal factor n	Sentencing
1	0/1	0/1	…	0/1	S1
2	0/1	0/1	…	0/1	S2
…	…	…	…	…	…
n	0/1	0/1	0/1	0/1	Sn

Case number	Legal factor 1	Legal factor 2	…	Legal factor n	Sentencing
1	0/1	0/1	…	0/1	S1

2	0/1	0/1	…	0/1	S2
…	…	…	…	…	…
n	0/1	0/1	0/1	0/1	Sn

Case number	Cosine similarity	Euclidean Distance
2	?	?
…	…	…
n	?	?

	baseline: 1
	Dependent variable:
	numberVerdict[-1]
numberReason[-1]	-0.430***
	(0.022)
Constant	1.623
	(0.014)
Observations	87,509
R^2	0.004
Note:	*p<0.1; >**p<0.05; >***p<0.01

Figure 2: Method to Identify Sentencing Biases

by running this procedure for every case. We then get 78,720 coefficients for every case. We conduct this procedure for the machine predictions as well. We assume that the treatment group is the machine predictions, while the control group is the judge predictions. We employee two comparison strategies. More exactly, we first compare the human judges' decisions with other human judges' decisions as the data of control group, and also compare the ML predictions with other human decisions as the data of treatment group. This strategy assumes that if all judges employed the machine's outcomes, what the counterfactual will be compared with if they do not use machine learning models. We also propose a framework for human-machine cooperation. We consider the other strategy, which is an imitation to the appeal in court system. Usually, if a defendant thinks the sentencing is biased, he will appeal. Or if the upper level court finds a case is biased, it can start a retrial. We then assume that, for all biased cases, the defendants appeal, or the court decides to retrial under the instruction of similarity results. And in the second trail, judges use the machine to assist their decisions. If the human decision is unbiased, then there is no need to use machine judges. We call this group as a machine-modify group, and this group is our new treatment group. In this case, we compare the machine's predictions with all human judges' decisions, so in this experiment, all cases are compared with the human judges' decisions. We then can calculate how many biased cases are there in machine predictions and judge predictions groups. We use the Risk Ratio (RR) to calculate the difference between machine judges and human judges. which is often used in medical tests. The equation of RR is:

$$RR = \frac{\frac{b_t}{b_t+u_t}}{\frac{b_c}{b_c+u_c}}$$

, where b_t means the number of biased cases in the machine predictions, considered as a treatment group, and u_t is the number of unbiased cases in the machine predictions; b_c is the number of biased cases in real human judge predictions, considered as a control group, and u_c is the number of unbiased cases in judge predictions. The baseline of RR is 1, meaning that machine predictions are tantamount to human predictions, and if the RR is lower than 1, it means machine judges perform better than human judges. We also report the Odds Ratio (OR), which is often used together with RR. The baseline of OR is also 1. and if the OR is lower than 1, it also means machine judges perform better than human judges. The equation of OR is showed as below, and the b_t and u_c have same meanings as they are in RR formula:

$$OR = \frac{\frac{b_t}{u_t}}{\frac{b_c}{u_c}}$$

Robustness Check: Gini Method

Anderson, Kling, and Stith (1999) proposed a method using Gini coefficient to evaluate inequality in sentencing. They consider sentencing decisions as enumeration data, and use a negative binominal model to estimate judges' effect. By imitating their method, we also employ a negative binominal model. We treat this method as a robustness check. A problem in social science model is, the relationship between legal factors and sentencing may not be linear, which means the negative binominal model may not capture the real residuals. Our explanation is: according to Supreme People's Court of PRC (2014)'s rule, For cases with multiple sentencing factors, the sentence is generally adjusted by adding those factors in the same direction and subtracting those in the opposite direction, according to the respective adjustment ratios of each sentencing factor., which indicates that judges' decisions are without any complex interactions. Besides, the adjustment ratios in Supreme People's Court of PRC (2014)'s announcement is presented in the form of percentage, so negative binominal model is much more suitable because in social science research, many researchers interpret the logged data into a percentage form. Based on above discussion, the log-link function equation of our model is set as:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

, where μ_i denotes the sentencing length of case i; X_{ik} denotes the legal factor k of case i; β_0 represents the intercept of our model; β_k denotes the coefficients estimated for legal factor k.

We use the excepted sentencing length to represent the sentencing of each case if they have same legal factors. We assume all cases share the same legal factors (all factors are zero), and thus the residuals represent that in each case, what length of sentencing are decided by extra-legal factors. The intercept is the expected sentencing lengths when all cases with no any sentencing legal factors. Thus, our residuals should plus the intercept to get the expected sentencing (Anderson, Kling, and Stith 1999). Then we use Gini to evaluate the inequality as a robustness check for new method. If all cases' residuals are roughly the same, then we can say there is no individual cases' biases. We use Lorenz curve to estimate the Gini:

$$G = 1 - 2\int_0^1 L(x)\,dx$$

, where G denotes the Gini coefficient; $\int_0^1 L(x)\,dx$ is the area under the Lorenz curve, and we use the integral to calculate it; $L(x)$ is the function of Lorenz curve. Lorenz curve is produced by cumulative proportion of sentencing decisions. To compare different decision makers' biases, we use machine judges' decisions to calculate the residuals again and calculate the Gini. Our method is: first, we use the human judges' decisions to build our negative binominal model, and get residuals of each case. Next, we use the machine predictions to minus model's excepted sentencing length, and get residuals of machine judges. Last, we use the two set of data to analysis the Gini coefficient.

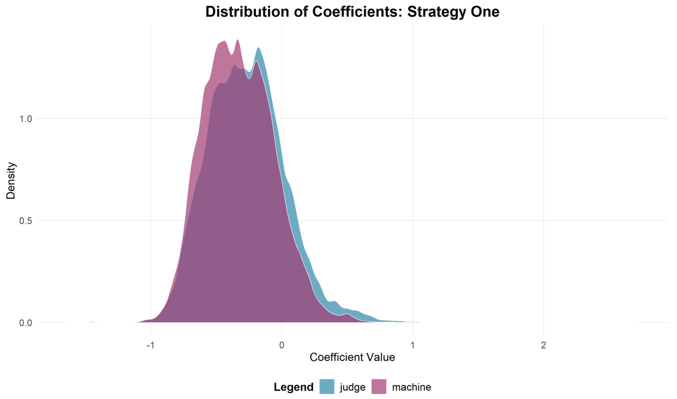


Figure 3: Distribution of Coefficients: Strategy One

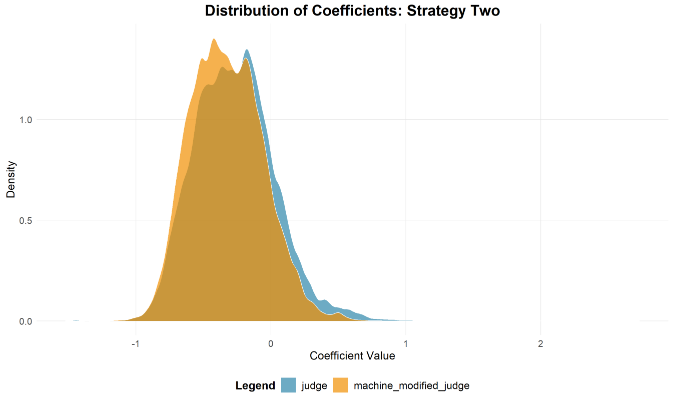


Figure 4: Distribution of Coefficients: Strategy Two

Findings

Main Findings

The first RR and OR result, showed in Table 1, uses the first strategy we proposed. The RR is 0.65, and statistically significant, which means if all defendants are sentenced by machine judges instead of human judges, the probability of being sentenced an unfair result is 35% lower. The OR also shows a similar result. The results of strategy two are showed in Table ??. The RR is 0.45, and statistically significant, which means if judges use a ML model to assist their work when the cases are appealed by defendants due to that defendants feel they are treated unequally, 55% biased cases can be sentenced in a more fair way. Our results also show that, by cooperating, the probability of producing biased cases drop by 57.14% compared with only using machines.

Another interesting finding is, in these two strategies, we all find that machine judges still have biases. In the first strategy, machine judges still produce 9689 biased cases, and back to the data itself, we find that some unbiased cases produced by judges, if re-decided by machine judges, will be biased. This means ML models also have biases, and they produce some new biases,

Table 1: Machine v.s. Judges: Strategy One

Strategy	RR (95% CI)	OR	Decision	Biased	Unbiased	Total
Strategy One	0.65 (0.64 to 0.67)	0.61 (0.59 to 0.63)	Machine	8644	70076	78720
			Human	13217	65503	78720
Strategy Two	0.45 (0.44 to 0.47)	0.41 (0.40 to 0.42)	Machine	5992	72728	78720
			Human	13217	65503	78720

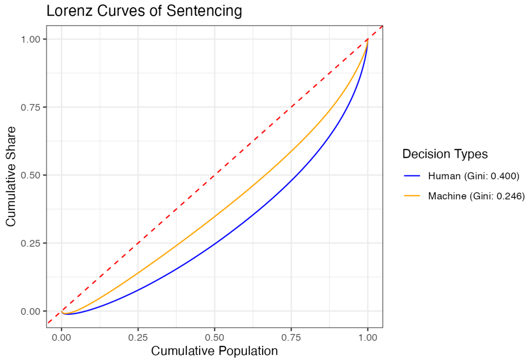


Figure 5: Lorenz Curves for Two Decisions

even though they can to some extend modify human's wrong decisions. They may have learned some wrong rules from biased cases.

Robustness Check

The Gini coefficient results are showed in Figure 5 in the form of Lorenz Curves. The red line denotes an ideal situation that all cases are treated equally. The blue curve and the orange curve is the Lorenz Curve for human and machine's predictions. Overall, by using machine judges, the inequality index for sentencing can be reduced by 38.5%.

Discussion and Conclusion

Our study results are consistent with Kleinberg et al. (2017), Lin et al. (2020), Goel, Rao, and Shroff (2016) and Hu et al. (2025) and strongly support Kleinberg et al. (2018)'s idea on that machine learning should be used instead of regulated to promote equality. ML model can reduce biases simply by building a model that excludes extra-legal factors. What is more, our research first compare ML model and human judges' decisions in the regression task, to be more exact, the sentencing task in criminal justice, advancing this area of research further. Machines can learn from biased human decisions and though treating biases as noise, they can reduce sentencing disparity to some extend. What is more, if they can cooperate with humans, the performance can be better.

Our contributions include: first, we proposed a new method, allows researchers to evaluate an individual case's bias based on comparison, making evaluating regression results possible. And we provide a method to evaluate sentencing disparity for researchers from other countries where extra-legal factors are difficult to get. This method overcomes a series of problems that prior social science research can not solve, especially the problem that it is impossible to control all extra-legal factors. Second, prior research considers that by excluding extra-legal factors, the ML model can produce a more fair result than human beings (Kleinberg et al. 2017, 2018; Hu et al. 2025), so as human beings (Chohlas-Wood et al. 2021). However, to delete extra-legal factors in sentencing is impossible because judges always need to interact with defendants or their lawyers. Thus, we focus on the model. Our research proves that it is also true in the regression task. We generate this theory to the regressor ML model, instead of the classifier ML model. We successfully proved that, ML model can reduce biases, because it does not take extra-legal factors into consideration. Besides, we also find that by cooperating with humans, the machine can perform even better. Third, our new method can be another evaluation for sentencing prediction models. As we have mentioned, most sentencing prediction models focus on accuracy (Xiao et al. 2018). However, the social impact of models should also be considered. ML studies have enough samples to conduct this method in a reliable way. We recommend that all sentencing prediction studies should employ the comparison-based method to evaluate the social impact of their models. Last, for legal researchers, the new disparity is worth considered that whether a CJ system should employ machine learning models and experts need to provide a cost-effectiveness analysis on using machines. Simply regulating machine judges is not a long way to go. We proved that machine perform better. A more proper way is to consider whether it is worth to introduce machine models to reduce sentencing biases in sacrifice of some new biased results and how to compensate the biased results made by machines while they are employed.

References

Anderson, J.; Kling, J.; and Stith, K. 1999. Measuring Interjudge Sentencing Disparity: Before and after the Federal Sentencing Guidelines. The Journal of Law and Economics, 42(S1): 271–308.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L.; and ProPublica. 2016. Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.

Anwar, S.; Bayer, P.; and Hjalmarsson, R. 2016. The Impact of Jury Race in Criminal Trials. The Quarterly Journal of Economics, 127(2): 1017–1055.

Barry, B. 2020. How Judges Judge: Empirical Insights into Judicial Decision-Making (1st ed.). London: Informa Law from Routledge.

Baumer, E. P. 2013. Reassessing and Redirecting Research on Race and Sentencing. Justice Quarterly, 30(2): 231–261.

Beccaria, C.; and Voltaire. 1876. An Essay on Crimes and Punishments With a Commentary by M. de Voltaire. Albany: W.C. Little & Co.

Blume, J.; Eisenberg, T.; and Wells, M. T. 2004. Explaining Death Row's Population and Racial Composition. Journal of Empirical Legal Studies, 1(1): 165–207.

Boyd, C. L.; Epstein, L.; and Martin, A. D. 2010. Untangling the Causal Effects of Sex on Judging. American Journal of Political Science, 54(2): 389–411.

Capurso, T. J. 1998. How Judges Judge: Theories on Judicial Decision Making. University of Baltimore Law Forum, 29(1): 5–15.

Chen, H.; Chen, Y.; and Yang, Q. 2025. Women in the Courtroom: Technology and Justice. The Review of Economic Studies, rdaf066.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Chohlas-Wood, A.; Nudell, J.; Yao, K.; Lin, Z. J.; Nyarko, J.; and Goel, S. 2021. Blind Justice: Algorithmically Masking Race in Charging Decisions. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, 35–45. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Costa, D.; and Zhao, Z. 2025. EQUALITY BEFORE THE LAW IN U.S. CIVIL WAR COURTS-MARTIAL. NBER Working Paper, (34184).

Cox, J. M.; Goldstein, N. E. S.; Dolores, J.; Zelechoski, A. D.; and Messenheimer, S. 2010. The Impact of Juveniles' Ages and Levels of Psychosocial Maturity on Judges' Opinions About Adjudicative Competence. Law and Human Behavior, 1–9.

Cresswell, J. C.; Sui, Y.; Kumar, B.; and Vouitsis, N. 2024. Conformal Prediction Sets Improve Human Decision Making. arXiv:2401.13744.

Danziger, S.; Levav, J.; and Avnaim-Pesso, L. 2011. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences of the United States of America, 108(17): 6889–6892.

Danziger, S.; Levav, J.; Avnaim-Pesso, L.; and Kahneman, D. 2011. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences of the United States of America, 108(17): 6889–6892.

Dressel, J.; and Fraid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1): eaao5580.

Engel, C.; Linhardt, L.; and Schubert, M. 2025. Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. Artificial Intelligence and Law, 33(2): 383–404.

Enough, B.; and Mussweiler, T. 2001. Sentencing Under Uncertainty: Anchoring Effects in the Courtroom. Journal of Applied Social Psychology, 31(7): 1535–1551.

Fan, Y. 2017. What should I do if I encounter a case that I have never dealt with before? | iCourt.

Fishman, G.; Rattner, A.; and Turjeman, H. 2006. Sentencing Outcomes in a Multinational Society: When Judges, Defendants and Victims Can Be either Arabs or Jews. European Journal of Criminology, 3(1): 69–84.

Franklin, T. W.; Dittmann, L.; and Henry, T. K. S. 2017. Extralegal Disparity in the Application of Intermediate Sanctions: An Analysis of U.S. District Courts. Crime & Delinquency, 63(7): 839–874.

Goel, S.; Rao, J. M.; and Shroff, R. 2016. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. The Annals of Applied Statistics, 10(1): 365–394.

Harris, A. P.; and Sen, M. 2019. Bias and Judging. Annual Review of Political Science, 22: 241–259.

Hart, H. L. A. 1958. Positivism and the Separation of Law and Morals. Harvard Law Review, 71(4): 593–629.

Hart, H. L. A. 1996. The Concept of Law. Oxford: Oxford University Press.

Haynes, S. H.; Ruback, B.; and Cusick, G. R. 2010. Courtroom Workgroups And Sentencing: The Effects Of Similarity, Proximity, And Stability. Crime and Delinquency, 56(1): 126–161.

Hester, R.; and Sevigny, E. L. 2016. Court Communities In Local Context: A Multilevel Analysis Of Felony Sentencing In South Carolina. Journal of Crime and Justice, 39(1): 55–74.

Heyes, A.; and Saberian, S. 2019. Temperature and Decisions: Evidence from 207,000 Court Cases. American Economic Journal: Applied Economics, 11(2): 238–65.

Hofman, J. M.; Watts, D. J.; Athey, S.; Garip, F.; Griffiths, T. L.; Kleinberg, J.; Margetts, H.; Mullainathan, S.; Salganik, M. J.; Vazire, S.; Vespignani, A.; and Yarkoni, T. 2021. Integrating explanation and prediction in computational social science. Nature, 595(7866): 181–188.

Hu, X.; Huang, Y.; Li, B.; and Lu, T. 2025. Human–Algorithmic Bias: Source, Evolution, and Impact. Management Science, Articles In Advance: 1–20.

Kennedy, D. 1976. Form and Substance in Private Law Adjudication. Harvard Law Review, 89(8): 1685–1778.

Kennedy, D. 1997. A Critique of Adjudication [Fin de Sicle]. Massachusetts: Harvard University Press.

Kim, B.; Spohn, C.; and Hedberg, E. C. 2015. Federal Sentencing as a Complex Collaborative Process: Judges, Prosecutors, Judge–Prosecutor Dyads, and Disparity in Sentencing. Criminology, 53(4): 597–623.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 133(1): 237–293.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2018. Discrimination in the Age of Algorithms. Journal of Legal Analysis, 10: 113–174.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How We Analyzed the COMPAS Recidivism Algorithm.

Legal Application Research Center. 2021. Criminal Procedure Document Templates of the Supreme People's Court: Production Standards and Legal Basis. Beijing: China Legal System Press.

Li, F.; Qin, Y.; and Yu, S. 2022. Theft or Felony? Task Augmentation for Criminal Amount Calculation in Judgment Documents. In 2022 International Conference on Asian Language Processing (IALP), 14–19. Shenzhen, Guangdong, China.

Li, S.; Zhang, H.; Ye, L.; Su, S.; Guo, X.; Yu, H.; and Fang, B. 2020. Prison Term Prediction on Criminal Case Description with Deep Learning. Computers, Materials & Continua, 62(3): 1217–1231.

Lin, J.; Xia, Y.; and Cai, T. 2024. Tip of The Iceberg? An Evaluation of the Non-uploaded Criminal Sentencing Documents in China. Asian Journal of Criminology, 19(3): 373–395.

Lin, Z. J.; Jung, J.; Goel, S.; and Skeem, J. 2020. The Limits of Human Predictions of Recidivism. Science Advances, 6(7): eaaz0652.

Liu, J. Z.; and Li, X. 2019. Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges. Journal of Empirical Legal Studies, 16(3): 630–670.

Ludwig, J.; and Mullainathan, S. 2024. Machine Learning as a Tool for Hypothesis Generation. The Quarterly Journal of Economics, 139(2): 751–827.

Mitchell, O. 2005. A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies. Journal of Quantitative Criminology, 21(4): 439–466.

Seniuk, G. T. 2016. SYSTEMIC INCOHERENCE IN CRIMINAL JUSTICE: FAILING TO TREAT LIKE CASES ALIKE. Canadian Bar Review, 93(3): 747–792.

Silver, E.; Ulmer, J. T.; and Silver, J. R. 2023. Do moral intuitions influence judges' sentencing decisions? A multilevel study of criminal court sentencing in Pennsylvania. Social science research, 115: 102927.

Supreme People's Court of PRC. 2014. Supreme People's Court's Sentencing Guidelines on Common Crimes.

Supreme People's Court of PRC. 2021. Supreme People's Court's Implementation Measures for the Unified Application of Law.

Tahura, U. S.; and Selvadurai, N. 2022. The use of AI in judicial decision-making: the example of China. International Journal of Law, Ethics, and Technology, 2022(3): 1–20.

Ulmer, J. T. 2012. Recent Developments and New Directions in Sentencing Research. Justice Quarterly, 29(1): 1–40.

Ulmer, J. T. 2019. Criminal Courts As Inhabited Institutions: Making Sense Of Difference And Similarity In Sentencing. Crime and Justice, 48: 483–522.

Ulmer, J. T.; and Bradley, M. S. 2006. VARIATION IN TRIAL PENALTIES AMONG SERIOUS VIOLENT OFFENSES. Criminology, 44(3): 631–670.

Ulmer, J. T.; Light, M. T.; and Kramer, J. H. 2011. Racial disparity in the wake of the Booker/Fanfan decision. Criminology & Public Policy, 10(4): 1077–1118.

Voeten, E. 2020. Gender and judging: evidence from the European Court of human rights. Journal of European Public Policy, 28(9): 1453–1473.

Weber, M. 1978. Economy and Society. Berkeley: University of California Press. Translated by Ephraim Fischoff, et al.

Wei, S.; and Xiong, M. 2020. Judges' Gender and Sentencing in China: An Empirical Inquiry. Feminist Criminology, 15(2): 217–250.

Wu, Y. 2020. Is a Plea Really a Bargain? An Empirical Study of Six Cities in China. Asian Journal of Criminology, 15(3): 237–258.

Wu, Y. 2021. The Boundaries of Sentencing Discretion: Collective Experience, Individual Decision-Making, and Bias Identification. China Journal of Law (Chinese Journal), 43(6): 109–129.

Xiao, C.; Hu, X.; Liu, Z.; Tu, C.; and Sun, M. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. AI Open, 2: 79–84.

Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; and Xu, J. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. arXiv:1807.02478.

Xin, Y.; and Cai, T. 2020. Paying Money for Freedom: Effects of Monetary Compensation on Sentencing for Criminal Traffic Offenses in China. Journal of Quantitative Criminology, 36(1): 1–28.

Xiong, M.; Xia, Y.; and Yu, X. 2025. Sentencing Equilibrium in Rape Cases: A Legal and Political Explanation of Jurisdictional Uniformity in China. Humanities and Social Sciences Communications, 12(1): 59.

Zhou, S.; Liu, Y.; Wu, Y.; Kuang, K.; Zheng, C.; and Wu, F. 2022. Similar Case Based Prison Term Prediction. In Fang, L.; Povey, D.; Zhai, G.; Mei, T.; and Wang, R., eds., Artificial Intelligence, 284–297. Cham, Kanton Zug, Switzerland: Springer Nature Switzerland.