# Automated Repression: Ethnic Discrimination in AI-assisted Criminal Sentencing

Eddie Yang*

November 2024

**Abstract**

This paper presents evidence of systematic ethnic bias in Artificial Intelligence (AI) used to assist judges in criminal sentencing in the People's Republic of China. Using an audit experiment on a commercial criminal sentencing software, I show that defendants with ethnic minority status on average receive upward of 6.2 percent longer predicted sentences than ethnic majority Han defendants who are described to have committed the exact same crimes. Ethnic cues such as names and ethnic identities each contribute to this discriminating effect. These findings hold across multiple ethnic groups and crime types. In addition to the main results, I show that the use of AI may introduce new forms of bias likely not seen in human data. Regardless of the intentionality of the ethnic bias in criminal sentencing AI, it points to one troubling venue where authoritarian politics manifest in new, more insidious forms through AI.

*Department of Political Science, Purdue University. Email: eddieyang@purdue.edu

# 1. Introduction

From algorithms making bail decisions in New Jersey to facial recognition systems used to verify voter identity in India, artificial intelligence (AI) has become an increasingly important decision-maker in political institutions across the world. Arguably, this is even more so in authoritarian regimes in which AI is often used for repressive purposes, such as protest suppression (Beraja et al., 2021) and surveillance (Xu, 2021). Yet despite alarms against the use of AI in authoritarian regimes (Diamond, 2019), we lack an understanding of AI as a decision-maker in authoritarian contexts. What kind of biases do AI systems in authoritarian regimes exhibit? How are such biases mediated by existing political institutions? What are the implications of (bias in) AI for authoritarian control?

This paper studies these questions by exploiting a rare opportunity in direct access to AI developed for sentencing in China's criminal courts. While the judicial system has historically been a part of the autocrats' repressive toolkit (Shen-Bayh, 2022), the use of AI to automate criminal sentencing decisions has been unprecedented. Drawing on existing literature showing ethnic discrimination in sentencing decisions by criminal court judges in China (Hou and Truex, 2022), I study whether the use of AI in criminal sentencing replicates ethnic bias of human judges. Additionally, I test if the use of AI introduces new patterns of discrimination not seen in human data.

To test ethnic discrimination in criminal sentencing AI, I perform an algorithmic audit experiment on a Chinese commercial criminal sentencing software. Specifically, I compare AI's sentence length predictions for defendants of different ethnic backgrounds, while holding all other relevant variables of the criminal cases fixed. Analysis of 35,000 experimental criminal cases and their associated sentencing predictions reveals systematic ethnic bias - selected ethnic minority groups receive predicted sentences that are 1.4% to 6.2% longer than Han ethnic majority defendants who are described to have committed the exact same crimes. The discriminating effect holds across multiple crime types. Ethnic cues such as

defendants' names and stated ethnic identities each contribute to this effect. The findings share similarities with ethnic discrimination found in judges' sentencing decisions in the Chinese criminal courts (Hou and Truex, 2022) but also reveal discrimination patterns that are unique to AI. Among other unique patterns, discrimination seems to concentrate on ethnic groups that have historically been subject to repression in China.

To my knowledge, this paper presents the first quantitative evidence of systematic ethnic bias in AI used by an authoritarian regime. The paper contributes to several strands of research. It joins a long line of research on judicial discrimination against racial/ethnic minorities in both democratic and authoritarian regimes (Alesina and La Ferrara, 2014; Grossman et al., 2016; Cohen and Yang, 2019; Choi et al., 2022) but is distinct from the existing literature by its focus on AI decision-makers instead of human judges. Beyond the judicial system, it adds to the literature documenting forms of ethnic discrimination in China, such as in the labor market (Hou et al., 2020) and in settlement (McNamee and Zhang, 2019). Lastly, the paper relates to a broader literature on the bias and fairness of AI systems[1] and adds further evidence to an emerging theme on how AI may perpetuate, exacerbate, or attenuate existing bias in political institutions (Angwin et al., 2016; Kleinberg et al., 2018; Ben-Michael et al., 2021; Yang and Roberts, 2021).

## 2. Background

**AI in the Chinese courts.** The Supreme People's Court of China has been a vocal advocate for the use of AI in the judicial system, stating the goal of using AI is to "digitize and modernize sentencing procedures and make citizens feel that every court case is handled with fairness and justice.[2]" As a result, prosecutors and judges in local courts have been using AI to assist with various workloads. For example, prosecutors in 2016 started using an AI tool, known as System 206, to evaluate the strength of evidence, conditions for an

[1] See e.g., Mehrabi et al. (2021) for a review.

[2] http://v5.pkulaw.cn/fulltext_form.aspx?Db=chl&Gid=293616

arrest, and how dangerous a suspect is to the public.[3] Local courts have also experimented with AI software to assist with various parts of the court system. For example, different AI systems have been used to verify defendant identity, transcribe court hearings and rulings, and recommend sentencing decisions. Although there is no official statistics, at least dozens of local courts have announced the use AI predictions in the criminal sentencing procedure. In one court, AI has been used for criminal sentencing for 204 criminal cases in 2020 and 98.91% of the sentencing predictions were adopted by the court.[4]

**Ethnic minorities in China.** China is a multi-ethnic state, with the Han - the largest ethnic group - and 55 ethnic minority groups recognized by the state. While ethnic minorities are granted policy privileges and some regional autonomy, ethnic tensions have been high and minorities can face discrimination in the labor market and court. For example, the Bai and the Yi in Yunnan province have been found to receive longer sentences for drug-related cases than Han defendants who committed similar crimes (Hou and Truex, 2022).

# 3. Data and Research Design

To test ethnic bias in criminal sentencing AI, I focus on a commercial criminal sentencing software and compare its sentencing and charge predictions for defendants of different ethnic backgrounds in China. To do so, I leverage a rare opportunity in which predictions from a commercial criminal sentencing software were publicly accessible during a short time window. The company that develops the software in question holds partnerships with the Supreme People's Court as well as a top Chinese university. While it is difficult to verify the extent to which it is used in courts, given the level of expertise involved and the fact that various criminal sentencing AI uses training data collected from the same source, it is reasonable to

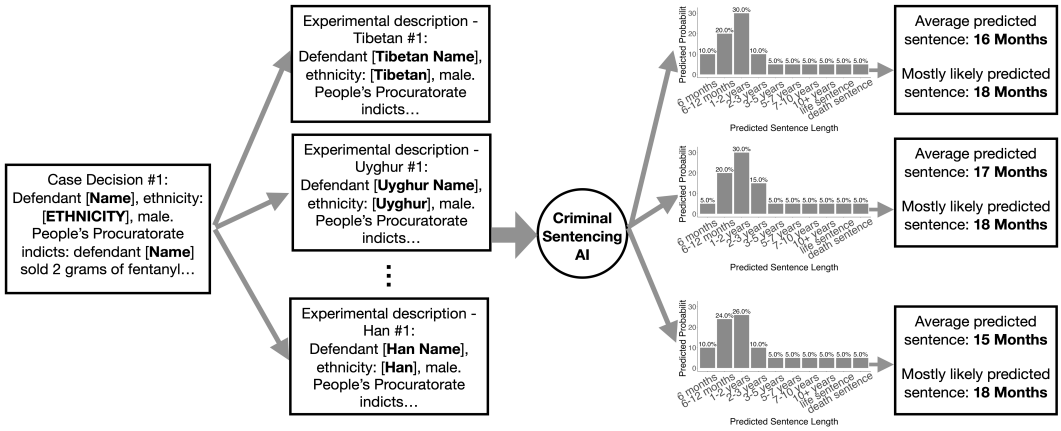
[3]Chen, Stephen, "Chinese scientists develop AI 'prosecutor' that can press its own charges." Dec 26, 2021. https://www.scmp.com/news/china/science/article/3160997/chinese-scientists-develop-ai-prosecutor-can-press-its-own

[4]Chen, Jiang & Guanqing Liu "Huaining: Three Steps to Efficient Sentencing Prediction" Aug 22, 2020. https://news.sina.com.cn/c/2020-08-22/doc-iivhuipp0037724.shtml

believe that the software is a strong contender for use in actual criminal trials and may be representative of such efforts to automate court decisions.[5]

The criminal sentencing software takes the text of a criminal case description as input and generates two outputs: 1) a probability distribution over possible sentence lengths and 2) a probability distribution over possible charges for the case. A case description includes the facts and evidence of the case as well as information on the defendants, including their names and ethnicity. In the main text, I focus on the software's predictions on sentence lengths and report results on charge predictions in Appendix A.3.3.

FIGURE 1. STYLIZED EXAMPLE OF EXPERIMENTAL SETUP




To test for ethnic bias, I conduct an audit experiment on the software. Specifically, I construct (fictitious but realistic) case descriptions for five kinds of crimes based on real case descriptions extracted from judicial decisions of local criminal courts. For each real case description, I replace the name and stated ethnicity of the original defendant with a name and ethnicity from seven (six minority and one majority) ethnic groups respectively. Thus, from each real case description, I generate seven experimental case descriptions for which the only difference is the name and ethnicity of the defendant. This is similar to a block randomized design for which the blocking variable is the real case description. Using the experimental case descriptions as input into the criminal sentencing software, I compare the

[5]See Section A.1. of the Online Appendix for details on the interface, algorithms, performance, and training data of the software. Public access to the software has since been revoked as of February 17, 2023.

predicted sentence lengths when the ethnic cues vary. A stylized example of the experimental setup is shown in Figure 1.

**Crime selection.** I choose five kinds of crimes based on the existing literature and to generate variations on the severity and political nature of the crimes.[6] First, the experiment replicates the settings in Hou and Truex (2022) by focusing on 1) drug crimes in Yunnan province and 2) drug crimes in all of China. In addition, three other crimes - 3) homicides and aggravated assaults, 4) frauds, and 5) crimes related to disturbing public order and state agencies (hereafter refered to as public disturbance) - are included. Similar to drug crimes, homicides and aggravated assaults are given long sentences on average. In contrast, frauds serve as an example of a minor crime with shorter sentences. Public disturbance crimes are an example of crimes that challenge the political authority of the state. For each crime, I randomly sample 1000 real case descriptions from local criminal courts published between October 2019 and September 2020 and construct the experimental case descriptions.

**Ethnic groups selection.** I focus on six ethnic minority groups in China, in addition to the Han majority group. Based on Hou and Truex (2022), I include the 1) Bai, 2) Yi, and 3) Zhuang ethnic groups. The authors found that the Bai and Yi received longer sentences than Han defendants whereas the Zhuang received shorter sentences. In addition, I select three ethnic groups that Hou and Truex (2022) found to receive sentences that are not statistically different from the Han. The 4) Tibetan and 5) Uyghur ethnic groups are included in the experiment as they have historically been subject to repression in China (Brox and Bellér-Hann, 2014). The 6) Hui ethnic group is also included as it is the most populous ethnic minority group in China.

**Estimation.** With five crimes (and 1000 judicial decisions for each crime) and seven ethnic groups, I construct $1000 \times 5 \times 7 = 35000$ experimental case descriptions to audit the criminal

[6]See Appendix A.2.3. for details on the specific charges each crime includes.

sentencing software. Given the block randomized design, I estimate the following difference in means in outcomes, while blocking on the real case descriptions, for each of the crimes:

$$Y_{ij} = \beta_i E_i + \tau_j + \varepsilon_{ij} \tag{1}$$

where $Y_{ij}$ is the outcome of interest. In particular, I study two outcomes (as shown in Figure 1):

1. average predicted sentence: sentence length averaged over the probability distribution of all possible sentence lengths.
2. mostly likely predicted sentence: sentence length with the highest predicted probability.

$E_i$ is a binary indicator for the $i$th ethnic minority group. The binary indicator for the Han majority group is omitted. $\tau_j \in \{1, 2, \ldots, 1000\}$ is the fixed effect for the $j$th real case description. Following the existing literature (Yin and Li, 2009; Hou and Truex, 2022), predictions of life sentences are converted to a sentence length of 264 months and predictions of death sentences are converted to 360 months. Outcomes are then log-transformed.[7]

# 4. Results

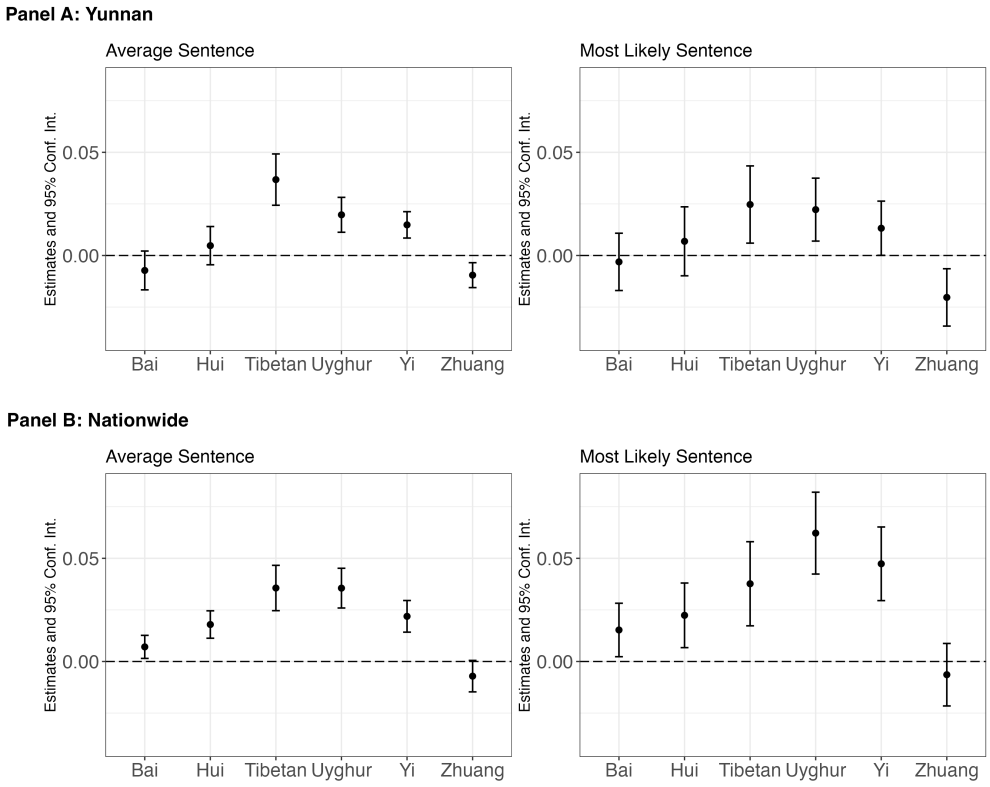
**Drug Crimes** Panel A and B in Figure 2 show the estimates for Yunnan and nationwide drug cases respectively. As shown in Panel A, Tibetan, Uyghur, and Yi defendants receive longer average predicted sentences, ranging from 1.5% to 3.7% longer than the baseline Han defendants.[8] This discriminating pattern also holds for the most likely predicted sentences, although the magnitudes are smaller. Similarly, Panel B shows that Tibetan, Uyghur, and Yi defendants also receive longer predicted sentences in drug cases across China, and for both average sentences and the most likely sentences. In addition, Hui defendants also receive

[7] In Appendix A.3.2, I also present results using negative binomial regression and outcomes without log-transformation. The substantive conclusions are largely unchanged.

[8] The mean of average sentence lengths for Han defendants is 44 months. 1.5% and 3.7% translate to 0.7 month and 1.6 months respectively.

longer sentences. The magnitudes of bias range from 1.8% to 3.6% for average sentences and 2.2% to 6.2% for most likely sentences. In Appendix A.3.1, I also show each of the two ethnic cues, namely the defendant's name and ethnicity, contributes to the biases.

FIGURE 2. ESTIMATES ACROSS ETHNIC GROUPS FOR DRUG CRIMES




Taken together, the findings from Yunnan and nationwide drug cases suggest systematic bias against several ethnic minority groups in predicted sentencing lengths. The results share similarities with ethnic discrimination by criminal court judges in China (Hou and Truex, 2022). Specifically, both this paper and Hou and Truex (2022) find evidence of ethnic bias of similar magnitudes against Yi defendants in Yunnan and nationwide drug cases. Both also find that Zhuang defendants receive shorter sentences in Yunnan cases but not in nationwide cases. Additionally, both studies find bias against Hui defendants for nationwide cases.

However, the findings based on the criminal sentencing software also point to AI's unique bias patterns not seen in Hou and Truex (2022). The most notable difference is

Table 1. Comparison with Hou and Truex (2022)

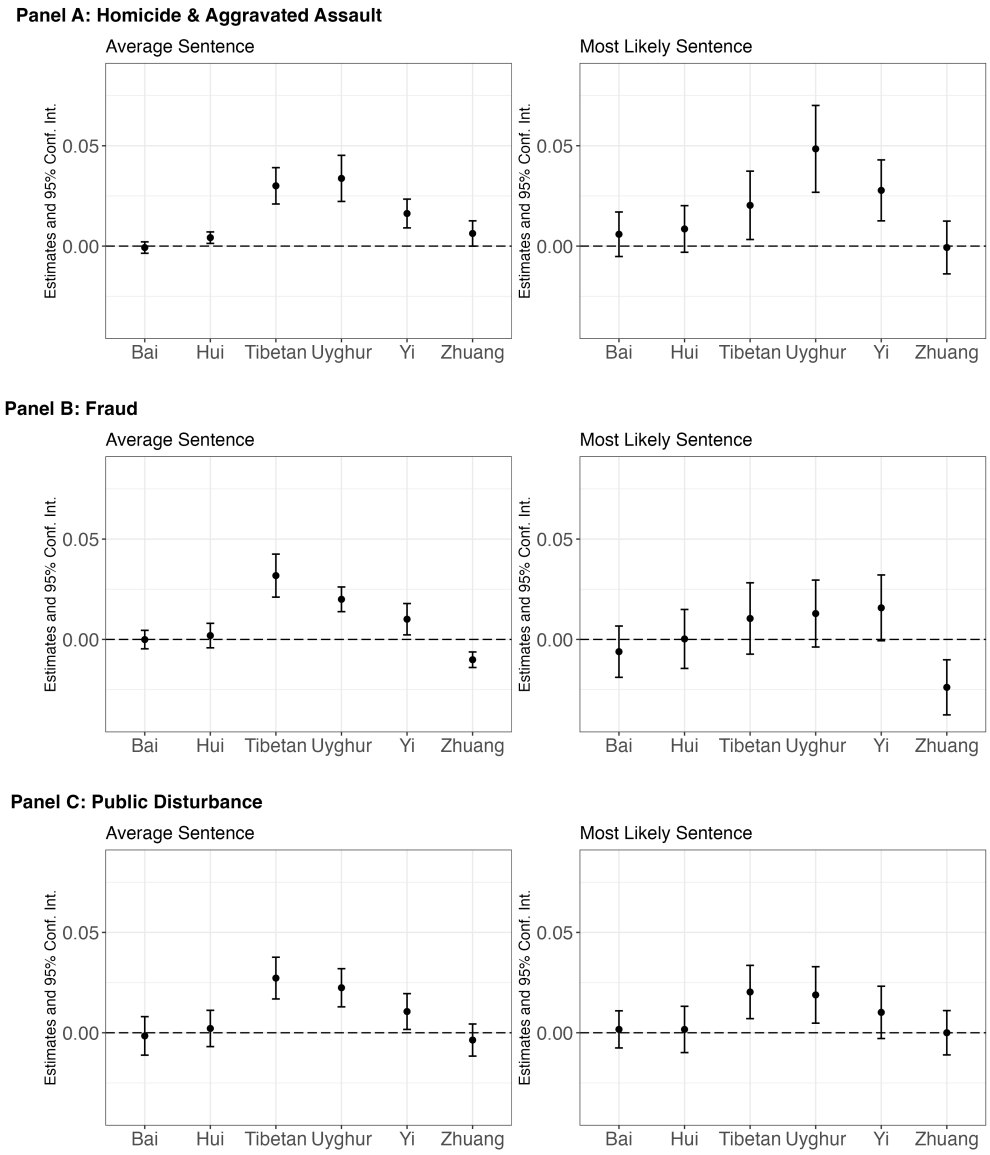
| | Yunnan | | Nationwide | |
|---|---|---|---|---|
| | Human Judges | AI | Human Judges | AI |
| Bai | + | . | + | . |
| Hui | . | . | + | + |
| Tibetan | . | + | . | + |
| Uyghur | . | + | . | + |
| Yi | + | + | + | + |
| Zhuang | - | - | . | . |

+ longer sentence than Han, - shorter sentence than Han, . no difference

the criminal sentencing software's bias against Tibetan and Uyghur defendants for both Yunnan and nationwide cases that is not seen in judge decisions. Additionally, Hou and Truex (2022) find bias against Bai defendants but there is no such bias from the criminal sentencing software. Table 1 summarizes the findings of the two studies. I discuss potential explanations for such differences in the conclusion.

**Other Crimes** Figure 3 presents estimates for homicide and aggravated assault, fraud, and public disturbance cases. Similar to the drug cases, I find ethnic bias in terms of sentence length against Tibetan, Uyghur, and Yi defendants, resulting in upward of 4.8% longer sentences. The ethnic bias holds across all three crimes and both measures of sentence length, although some estimates are less precise. The results suggest that ethnic bias is present and quite stable across variations in the type, severity, and political nature of the crime.

Figure 3. Estimates Across Ethnic Groups for Other Crimes




# 5. Discussion

Using an audit experiment on a commercial criminal sentencing software, I uncover biases against ethnic minorities in automated criminal sentencing decisions. These biases have both similarities with and differences from biases exhibited by criminal court judges in China. The findings highlight the potential of AI to be used as a tool to automate

repression by mass-producing biased sentencing decisions against ethnic minorities. However, the deviation of the biases of AI from those of human judges, especially non-discrimination of some ethnic minorities, illustrates the potential limitations and difficulties of using AI to automate repressive tasks previously carried out by human agents.

One direction of future research is to better understand how AI's discrimination patterns came to be. However, tracing AI's discrimination patterns to their root causes is extremely difficult. Although we can be fairly certain about the type (sentencing decisions) and source (Supreme People's Court's centralized website) of the data that is used to train the criminal sentencing software[9], the actual training data as well as information on the technical details of the software and the training procedure are not available. Even with such information, we currently lack methods that can precisely attribute the behavior of the AI to specific characteristics of the training data or the procedure (Barocas et al., 2018).

Here I offer several conjectures that plausibly explain the bias patterns of the criminal sentencing software based on the available data and insights from existing literature on bias in AI systems. With regards to AI's bias against Tibetan and Uyghur ethnic groups that is not found in judge decisions on drug cases in Hou and Truex (2022), it is possible that similar bias may be present in judge decisions on other criminal cases. Such bias can create spurious correlations between Tibetan/Uyghur status and sentence length. It is possible that AI picks up such correlations and generalizes them to drug cases. The fact that AI can pick up and generalize potentially spurious correlations has been well-documented in the computer science literature.[10] Whether similar discrimination is found in judge decisions should be a focus of future research.

With regards to AI's non-discrimination against Bai ethnic minorities, a possible explanation is that, in contrast to Tibetan, Uyghur, and Yi names that are mostly distinct from Han names and convey strong ethnic cues, Bai names are much more culturally assimilated (and thus similar) to Han names. Coupled with the fact that Bai has a relatively small popu-

[9]See Appendix A.1.1 and A.1.5 for more details.

[10]See e.g., Zhao et al. (2017).

lation in China, it is likely that there is relatively sparse information in the training data for AI to learn to discriminate against Bai defendants. This suggests an interesting phenomenon in which marginalized groups may undo existing discrimination through the "ignorance" of AI. Establishing such a phenomenon and exploring its manifestations in other domains can be an interesting direction of future research.

# References


Alesina, A. and E. La Ferrara (2014). A test of racial bias in capital sentencing. *American Economic Review 104*(11), 3397–3433.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias. propublica, may 23, 2016.

Barocas, S., M. Hardt, and A. Narayanan (2018). Fairness and machine learning. fairmlbook.org, 2019.

Ben-Michael, E., D. J. Greiner, K. Imai, and Z. Jiang (2021). Safe policy learning through extrapolation: Application to pre-trial risk assessment. *arXiv preprint arXiv:2109.11679*.

Beraja, M., A. Kao, D. Y. Yang, and N. Yuchtman (2021). Ai-tocracy. Technical report, National Bureau of Economic Research.

Brox, T. and I. Bellér-Hann (2014). *On the fringes of the harmonious society: Tibetans and Uyghurs in socialist China*. Nias Press.

Choi, D. D., J. A. Harris, and F. Shen-Bayh (2022). Ethnic bias in judicial decision making: Evidence from criminal appeals in kenya. *American Political Science Review*, 1–14.

Cohen, A. and C. S. Yang (2019). Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy 11*(1), 160–91.

Diamond, L. (2019). The road to digital unfreedom: The threat of postmodern totalitarianism. *Journal of Democracy 30*(1), 20–24.

Grossman, G., O. Gazal-Ayal, S. D. Pimentel, and J. M. Weinstein (2016). Descriptive representation and judicial outcomes in multiethnic societies. *American Journal of Political Science 60*(1), 44–69.

Hou, Y., C. Liu, and C. Crabtree (2020). Anti-muslim bias in the chinese labor market. *Journal of Comparative Economics 48*(2), 235–250.

Hou, Y. and R. Truex (2022). Ethnic discrimination in criminal sentencing in china. *The Journal of Politics 84*(4), 000–000.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The quarterly journal of economics 133*(1), 237–293.

McNamee, L. and A. Zhang (2019). Demographic engineering and international conflict: Evidence from china and the former ussr. *International Organization 73*(2), 291–327.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*(6), 1–35.

Shen-Bayh, F. F. (2022). *Undue Process*. Cambridge University Press.

Xu, X. (2021). To repress or to co-opt? authoritarian control in the age of digital surveillance. *American Journal of Political Science 65*(2), 309–325.

Yang, E. and M. E. Roberts (2021). Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 537–548.

Yin, M. and X. Li (2009). Empirical analysis of homicide: the perspective of 493 murder-cases ( 故意杀人罪实证研究: 以 493 例故意杀人罪例为视角). *Criminal Science (* 中国刑事法杂志*) 18*, 1.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

# Automated Repression: Ethnic Discrimination in AI-assisted Criminal Sentencing

Online Appendix

## Table of Contents

# A. Further Details on the Criminal Sentencing Software

## A.1. Use of AI in law in China

The Supreme People's Court of China has on multiple occasions advocated for the use of AI in the judicial system, with the stated goal of "digitizing and modernizing sentencing procedures and making citizens feel that every court case is handled with fairness and justice."[1] As a result, prosecutors and judges in local courts have been using AI to assist with various workload since 2016. More recently, courts across the country have initiated experiments to integrate AI into adjudication by introducing software that reviews evidence, suggests outcomes, checks the consistency of judgments, and makes recommendations on how to decide cases (Stern et al., 2020). For example, the city of Hangzhou has experimented with an artificial "judge assistant" that automatically recommends relevant laws and suggests judgements to judges.[2] According to one report, the use of AI in court has increased both the efficiency and accuracy with respect to criminal cases in the province of Hainan, with AI reducing the judges' time spent on sentencing decisions by 50%.[3] Yet, despite the wide adoption of AI in the Chinese courts, the fairness of such AI systems is largely untested (Stern et al., 2020).

The development of AI for the judicial system is in part facilitated by the Supreme People's Court's initiative to centralize the collection of judicial decisions. Beginning in 2013, all courts in China were asked to upload judicial decisions onto a central website administered by the Supreme People's Court.[4] As of March 2023, over 139 million court documents has

[1] Supreme People's Court, "Opinions by the Supreme Court on Accelerating the Establishment of Smart Court 最高人民法院关于加快建设智慧法院的建议" April 12, 2017. http://gongbao.court.gov.cn/Details/5dec527431cdc22b72163b49fc0284.html

[2] Mei Zhu, "Hangzhou Internet Court Experiments with artificial "judge assistant" 杭州互联网法院试点应用 "AI 助理法官"" June 14, 2019. https://zj.zjol.com.cn/news.html?id=1221164

[3] Ke Ji Ri Bao, "Avoid Different Judgments in Similar Cases, an AI Judge Came to Hainan 避免同案不同判，海南来了位 AI"法官"" April 15, 2019. http://scitech.people.com.cn/n1/2019/0415/c1057-31030409.html

[4] https://wenshu.court.gov.cn/

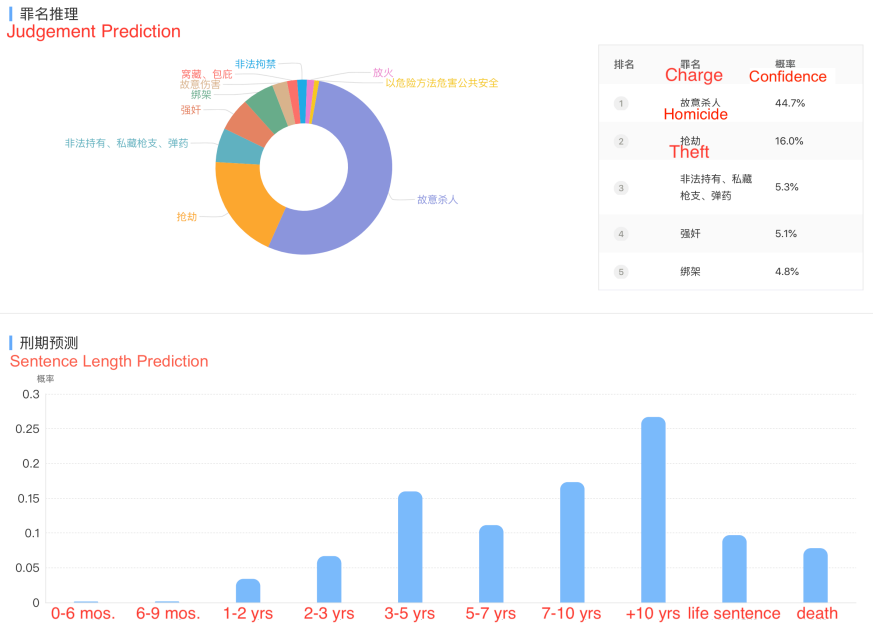
been uploaded, including over 9.9 million documents for criminal trials. These cases are written in a standardized format across provinces and emphasize outcomes and case facts (Liebman, 2015). Technology companies keen on developing AI for the judicial system have capitalized on this large collection of legal documents by turning them into training data for AI models to learn to make sentencing decisions. Part of this collection has also become the standard training data in an annual nationwide AI competition for companies and academic institutions to showcase their latest legal technologies.[5]

[5] http://cail.cipsc.org.cn/

## A.2. User interface of the criminal sentencing software

Figure A1 shows the user interface of the criminal sentencing software. The bottom panel gives the probability distribution over binned sentence lengths. It also gives the exact probabilities when the cursor hovers over the bins of the histogram. The top right panel gives the probability distribution over possible charges. The top left panel gives the same charge distribution but in a pie chart. Note that the analysis in the paper uses results returned by the software's application programming interface (API) rather than from the interface as shown here. The API results give more precise probabilities as well as charges that have been truncated by the user interface.

FIGURE A1. CRIMINAL SENTENCING SOFTWARE USER INTERFACE



## A.3. Prediction models of the criminal sentencing software

According to public information on the software, it uses a combination of transformer-based deep neural networks, such as BERT (Devlin et al., 2018) and ERNIE (Sun et al., 2019), fine-tuned on custom data and augmented with knowledge graphs. A deep neural network is a non-linear model that can have hundreds of millions of parameters and is capable of approximating complex functions, e.g., functions that map the text of case descriptions to sentence lengths and charges. These large pre-trained neural networks represent the state-of-the-art deep learning models for classification and have been used in a variety of other real life settings, such as in optimizing Google's search queries and medical imaging (Shamshad et al., 2022). In addition to deep neural networks, the software extracts information relevant for sentencing and charge from case descriptions (e.g., whether the crime resulted in deaths, whether the crime is intentional, and whether the defendant confessed to the crime) and uses this information in a knowledge graph (Hogan et al., 2021) to augment the software's ability at legal reasoning.

Although the exact prediction models used in the software are unknown, they are likely to be representative of similar efforts to automate court decisions. For example, among five other commercial criminal sentencing software, all state that they use deep neural networks and all but one state that they use knowledge graphs to embed legal knowledge in their software. While these other software cannot be publicly accessed, given the similarity in algorithms, the findings from the paper may be a more general phenomenon that applies to criminal sentencing software in China.

## A.4. Model performance

Table A1 reports the performance of the criminal sentencing software against the actual sentence lengths of the cases across crime categories. The correlation measure reports the correlation between the predicted average sentence lengths and the actual sentence lengths. R-squared is obtained by regressing the actual sentence lengths on the predicted average sentence lengths without intercept. Mis-classification reports the percentage of cases for which the predicted charge differs from the actual charge of the case. Except for public disturbance cases, the correlations are high ($> 0.7$) between the actual and the predicted sentence lengths. The high R-squared also suggests the predicted sentences explain substantial variations in the actual sentences. Mis-classification rates are low but non-negligible. Potentially due to the similarity among different fraud charges, the mis-classification rate is much higher for fraud than other crime category. Overall, the criminal sentencing software's predictions are much better than random. Considering that the input space (any raw text) to the software is vast, the relatively high performance of the criminal sentencing software is likely to be representative of the state-of-the-art efforts at automating criminal sentencing in China.

TABLE A1. COMPARISON BETWEEN ACTUAL AND PREDICTED SENTENCE LENGTHS

| | Drug: Yunnan | Drug: Nation | Homicide | Fraud | Public Disturbance |
|---|---|---|---|---|---|
| Correlation | 0.894 | 0.725 | 0.801 | 0.767 | 0.194 |
| R-squared | 0.883 | 0.672 | 0.750 | 0.746 | 0.674 |
| Mis-classification | 1.7% | 2.1% | 1.5% | 22.7% | 3.2% |

## A.5. Training data

Given the centralized, digital nature of the judicial decisions hosted on the Wenshu website by the Supreme People's Court, it has become the go-to training data for a variety of legal AI applications. Importantly, data missingness (judicial decisions not uploaded to the website) is primarily driven by resource constraints of the local courts (Liebman et al., 2020; Wu et al., 2022) and likely equally affects the training data for all legal AI applications as well. Similar to the collection of training data by legal AI companies, the judicial decisions used in the paper are also collected from the Wenshu website. The latest cases shown on the software's website are from 2018. Given the available information, I infer that the software is trained on data up until 2018. Therefore, to prevent testing the software with data it has encountered in training, I used cases from 2019 to 2020 to construct test examples in the audit experiment.

A publicly available dataset on Chinese criminal court cases (CAIL2018) covers a subset of cases from 2000 to 2017 (Xiao et al., 2018). Table A2 shows an example from the CAIL2018 dataset. The example includes a case description, used as input, as well as several outcomes, such as sentence length, charge, relevant law, and the fine amount.

TABLE A2. EXAMPLE FROM CAIL2018 DATASET

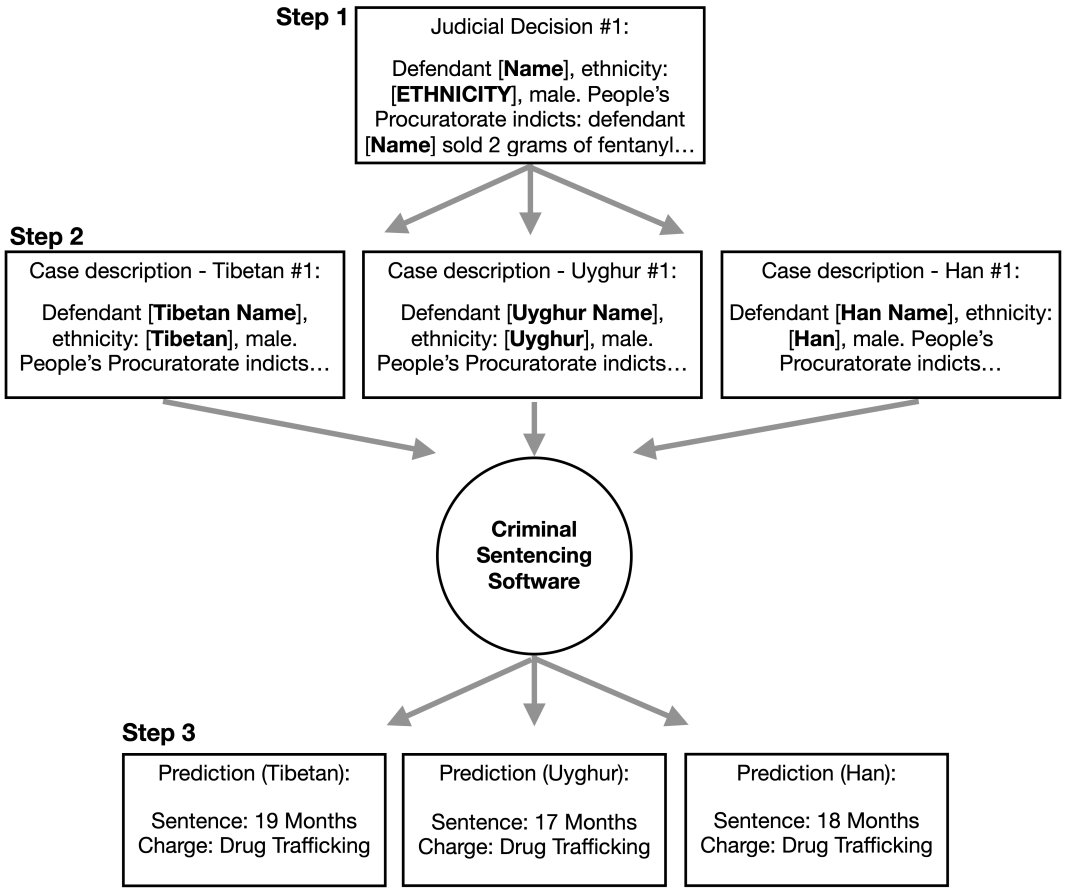
| Variable Name | Value |
|---|---|
| Case Description | 经审理查明：2014 年 9 月 28 日 14 时许，被告人杜某某与家人就餐，席间喝一瓶半啤酒。同日 16 时许，杜某某醉酒无证驾驶黑 LE0452 号铃木牌两轮摩托车返家，其驾车沿方正县莲新公路自南向北行驶至 3 公里处时，由于驾车操作不当，发生单方交通事故。过往群众随即报警并将杜某某送往方正县人民医院救治。经抽血检测，被告人杜某某血液中乙醇含量为 219.519mg / 100ml。经侦查，公安机关于 2014 年 9 月 28 日在方正县人民医院将被告人杜某某抓获。上述事实，由公诉机关向本院提交并经庭审质证、认证的下列证据予以证实：案件来源、到案经过及受案登记表：2014 年 9 月 28 日 16 时 20 分，杜某某酒后无证驾驶黑 LE0452 号铃木牌两轮摩托车沿方正县莲新路由南向北行至 3 公里处时，发生单方交通事故，同日在方正县人民医院因伤治疗期间被我局工作人员查获。经司法检验鉴定，杜某某血液中乙醇含量为 219.519mg / 100ml，属于醉酒状态。现场勘查笔录、平面图及照片：现场位于方正县莲新公路 3 公里处，东面是农田，南面是去往宝兴乡方向，西面是农田，北面是去往方正镇方向。黑 LE0452 号铃木牌两轮摩托车头南尾北，停放在莲新公路 3 公里东侧路边。驾驶人信息查询结果单、机动车信息查询结果单及车辆照片：被告人杜某某未办理机动车驾驶证，其所驾驶的黑色黑 LE0452 号金某铃木牌普通二轮摩托车登记所有人为王某某。哈尔滨市公安医院司法鉴定所乙醇检验报告：被告人杜某某血液中乙醇含量为 219.519mg / 100ml。公安交通管理行政处罚决定书、黑龙江省政府非税收入专用收据：被告人杜某某因未取得驾驶资格驾驶机动车辆于 2014 年 9 月 30 日被处以罚款人民币 500 元，同年 10 月 1 日缴纳罚款人民币 500 元。户籍证明及现实表现：被告人杜某某，男，汉族，黑龙江省方正县公安局宝兴派出所未发现杜某某在其辖区居住期间有犯罪行为，现实表现一般。7、被告人杜某某的供述：2014 年 9 月 28 日 16 时左右，我在宝兴乡王家村喝酒，喝一瓶半。喝完酒我驾摩托车回方正镇，当我驾车沿莲新公路由南向北行驶至 3 公里附近时，因车辆失控摔倒，我受伤了，事情经过就是这样。我对血液中酒精含量为 219.519mg / 100ml 的检验结果没有异议。被告人杜某某未提交和申请调取任何证据。 |
| Charge | 危险驾驶罪 |
| Relevant Law | 刑法第一百三十三条 |
| Life Imprisonment | False |
| Death Penalty | False |
| Sentence | 4 months |
| Fine | 0 |

# B. Further details on the audit experiment

## B.1. Graphical representation of the audit experiment

Figure A2 shows that step-by-step data collection process of the audit experiment. First, judicial decisions from local courts in China are collected from the central website administered by the Supreme People's Court. In step 2, case descriptions are constructed by replacing the names and ethnicity of the original defendants from the judicial decisions with those that convey specific ethnic cues from a pre-determined list of names and ethnicity (See Section B.4 for details on the names). In step 3, the experimental case descriptions are used as input in the criminal sentencing software to generate sentence and charge predictions. The predictions are then used as outcomes to compare discrepancies among defendants of different ethnicity.

FIGURE A2. DATA COLLECTION IN THE AUDIT EXPERIMENT



## B.2. Example of case description in English

FIGURE A3. EXAMPLE OF CASE DESCRIPTION

Defendant [NAME], female, date of birth: 1995/10/02, [ETHNICITY], household registration location: [PROVINCE], city, prefecture. [PROVINCE] people's procuratorate indicts: On 2020/03/20 at 1AM, Defendant [NAME] was arrested for selling 0.98 gram of methamphetamine to drug addict XXX for ¥300 around [PROVINCE], XXX district. public prosecutors read and showed in court: 1) proof of household registration, 2) recount of arrest, 3) administrative penalty decision, 4) defendant testimony and photos from the crime scene, 4) record for sealing, weighting and sampling of evidence, 5) witness testimonies. Public prosecutors claim that the action of selling methamphetamine by Defendant [NAME] constitutes drug trafficking. Defendant [NAME] did not dispute in court. Defendant [NAME] claims that the drug he sold did not go into circulation and thus the action should be considered less detrimental to society. Defendant [NAME] confessed the facts of the crime after the arrest and showed remorse and thus should be considered for a lighter sentence.

***Note:*** Translation of an actual, truncated case description in Chinese. Words in brackets are replaced to construct test examples.

## B.3. Details on collection of court rulings

TABLE A3. SUMMARY STATISTICS ON JUDICIAL DECISIONS

| Crime | Number of Decisions | Category |
|---|---|---|
| Smuggling, Trafficking, Transporting, & Manufacturing Drugs | 30 402 | Drug |
| Illegal Possession of Drugs | 2277 | Drug |
| Manufacturing, Dealing, Transporting & Smuggling Drug Raw Materials | 67 | Drug |
| Aggravated Assault | 42 733 | Homicide |
| Homicide | 2957 | Homicide |
| Credit Card Fraud | 1899 | Fraud |
| Illegal Fund-raising | 650 | Fraud |
| Insurance Fraud | 169 | Fraud |
| Loan Fraud | 162 | Fraud |
| Fraud Involving Financial Bills | 83 | Fraud |
| Fraud Involving Financial Certificates | 6 | Fraud |
| Fraud Involving Letters of Credit | 1 | Fraud |
| Securities Fraud | 0 | Fraud |
| Obstruction of Officer in Discharge of Duties | 9812 | Public Disturbance |
| Obstruction of State Agencies | 81 | Public Disturbance |
| Disturbance of Public Order or State Agencies | 26 | Public Disturbance |
| Assault on Police Officers | 0 | Public Disturbance |
| Instigating Violent Resistance to Law Implementation | 0 | Public Disturbance |

***Note:*** Judicial decisions are collected from wenshu.court.gov.cn website from Oct. 2019 to Sept. 2020. for all provinces. The collected crime categories are: drug-related crimes, aggravated assault and homicide, fraud, and public disturbance. The specific crimes that are included in each category are according to the Criminal Law Of The People's Republic Of China. A value of 0 means that no judicial decision was available for the particular crime during this period.

## B.4. Minority names

To ensure that the names used in the experiment accurately reflect ethnic minority names in China, minority names used to construct case descriptions are extracted from the Administrative Lawsuits dataset (Baik and Dai, 2022), which covers administrative rulings published between 2014 and 2018 in China. Getting names from a dataset that is different from the criminal court casess also helps minimize the chances that the software falsely associates a name with a case it has encountered during the training of the software.

From the Administrative Lawsuits dataset, I extract information on the names and ethnicity of defendants for each administrative case. I remove all names that have been anonymized (e.g. 某某, 某 XX, 某 **). Then from the list of remaining names, I extract all names for which the defendants' ethnicity is one of Han, Bai, Hui, Tibetan, Uyghur, Yi, and Zhuang. When constructing a case description, I randomly sample a name from the list that corresponds to the ethnicity of the defendant in that case description.

# C. Additional results

## C.1. Effect by ethnic cues

Table A4 and Table A5 show the effects of defendant name and defendant ethnicity respectively. Results in Table A4 are based on case descriptions that exclude the ethnic identity of the defendants. Results in Table A5 are based on case descriptions that anonymize the names of the defendants (by Chinese convention, all names are converted to 某某某).

Overall, both name and ethnicity effects are consistent with the main effects - defendants with Tibetan, Uyghur, and Yi names and defendants with Tibetan, Uyghur, and Yi ethnic identities on average receive longer predicted sentences across all crime categories and both measures of sentence lengths (average and most likely sentence). Hui defendants also receive longer sentences for nationwide drug crimes. The magnitudes of the ethnic bias from name or ethnicity alone are generally smaller than the main effects, which include ethnic cues from both names and ethnicity.

TABLE A4. EFFECT OF DEFENDANT NAME

| | Drug: Yunnan | | Drug: Nation | | Homicide | | Fraud | | Public Disturbance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely |
| Bai | 0.003 | 0.007 | 0.008** | 0.015** | 0.002 | −0.005 | 0.000 | 0.000 | −0.001 | 0.000 |
| | (0.003) | (0.005) | (0.004) | (0.006) | (0.001) | (0.006) | (0.003) | (0.007) | (0.002) | (0.007) |
| Hui | 0.002 | 0.006 | 0.015*** | 0.017** | 0.002 | −0.005 | 0.004 | −0.001 | 0.001 | −0.001 |
| | (0.002) | (0.006) | (0.004) | (0.007) | (0.002) | (0.005) | (0.003) | (0.006) | (0.002) | (0.006) |
| Tibetan | 0.035*** | 0.026*** | 0.036*** | 0.029*** | 0.030*** | 0.020** | 0.031*** | 0.016 | 0.028*** | 0.016 |
| | (0.005) | (0.009) | (0.006) | (0.010) | (0.005) | (0.008) | (0.005) | (0.010) | (0.003) | (0.010) |
| Uyghur | 0.026*** | 0.039*** | 0.030*** | 0.044*** | 0.023*** | 0.039*** | 0.017*** | 0.010 | 0.015*** | 0.010 |
| | (0.004) | (0.009) | (0.005) | (0.010) | (0.003) | (0.008) | (0.004) | (0.008) | (0.003) | (0.008) |
| Yi | 0.006** | 0.011* | 0.009** | 0.014** | 0.003 | −0.006 | 0.011** | 0.017* | 0.003 | 0.017* |
| | (0.003) | (0.006) | (0.004) | (0.006) | (0.002) | (0.006) | (0.005) | (0.009) | (0.002) | (0.009) |
| Zhuang | −0.001 | 0.003 | 0.003 | 0.008 | 0.002 | −0.003 | 0.002 | −0.005 | −0.002 | −0.005 |
| | (0.003) | (0.005) | (0.004) | (0.007) | (0.002) | (0.005) | (0.003) | (0.006) | (0.002) | (0.006) |
| FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

***Note:*** * p < 0.1, ** p < 0.05, *** p < 0.01. All outcomes are log-transformed. All models control for judicial decision fixed effects. Columns with header "Average" are based on models for which the average sentence lengths are the outcome variable. Columns with header "Most Likely" are based on models for which the most likely sentence lengths are the outcome variable.

## TABLE A5. EFFECT OF DEFENDANT ETHNICITY

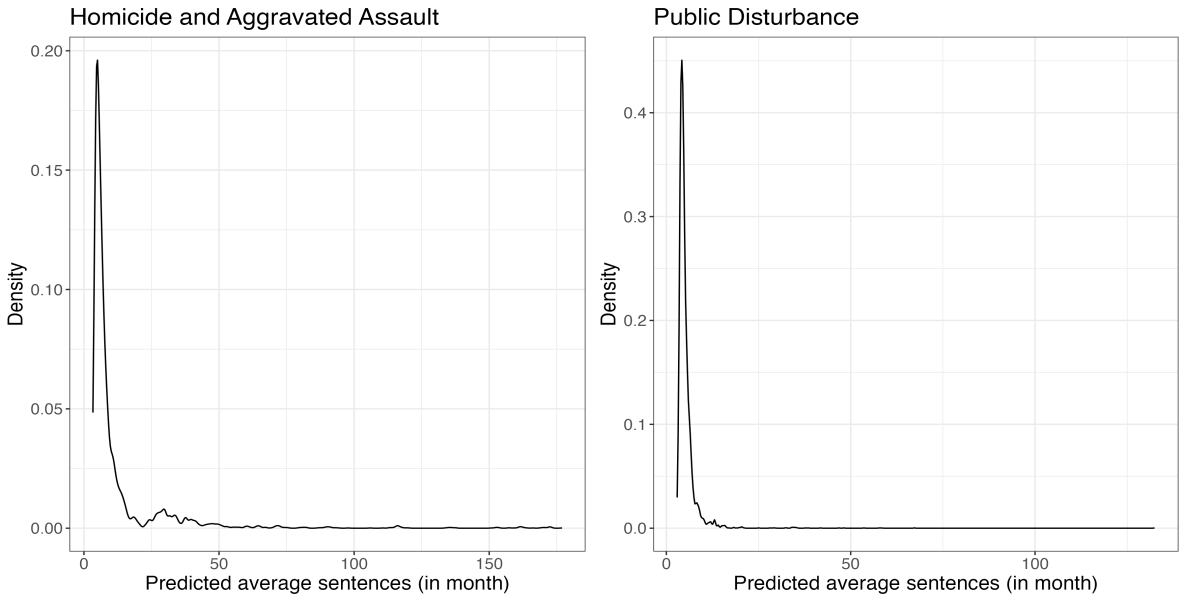
| | Drug: Yunnan | | Drug: Nation | | Homicide | | Fraud | | Public Disturbance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely |
| Bai | −0.001 | 0.013*** | 0.000 | 0.012** | −0.003*** | 0.010** | 0.001*** | 0.006* | 0.000** | 0.006* |
| | (0.001) | (0.005) | (0.001) | (0.005) | (0.000) | (0.004) | (0.000) | (0.003) | (0.000) | (0.003) |
| Hui | 0.005*** | 0.010* | 0.008*** | 0.012* | 0.004*** | 0.005* | −0.001 | 0.010* | 0.004*** | 0.010* |
| | (0.001) | (0.005) | (0.001) | (0.006) | (0.000) | (0.003) | (0.001) | (0.005) | (0.000) | (0.005) |
| Tibetan | −0.002** | 0.005 | 0.000 | 0.017** | −0.005*** | 0.009** | 0.001*** | 0.008** | −0.001*** | 0.008** |
| | (0.001) | (0.005) | (0.001) | (0.007) | (0.000) | (0.004) | (0.000) | (0.003) | (0.000) | (0.003) |
| Uyghur | 0.008*** | 0.018*** | 0.007*** | 0.023*** | 0.005*** | 0.012*** | 0.004*** | 0.012*** | 0.003*** | 0.012*** |
| | (0.001) | (0.005) | (0.001) | (0.007) | (0.000) | (0.004) | (0.000) | (0.004) | (0.000) | (0.004) |
| Yi | 0.011*** | 0.025*** | 0.013*** | 0.038*** | 0.005*** | 0.023*** | 0.010*** | 0.022*** | 0.008*** | 0.022*** |
| | (0.001) | (0.005) | (0.001) | (0.008) | (0.000) | (0.005) | (0.000) | (0.006) | (0.000) | (0.006) |
| Zhuang | −0.006*** | −0.010** | −0.013*** | −0.019*** | 0.006*** | −0.002 | −0.010*** | 0.006 | 0.003*** | 0.006 |
| | (0.001) | (0.005) | (0.001) | (0.006) | (0.001) | (0.004) | (0.001) | (0.004) | (0.000) | (0.004) |
| Fixed Effect | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

***Note:*** * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. All outcomes are log-transformed. All models control for judicial decision fixed effects. Columns with header "Average" are based on models for which the average sentence lengths are the outcome variable. Columns with header "Most Likely" are based on models for which the most likely sentence lengths are the outcome variable.

## C.2. Additional results from alternative models

In the main specification, the outcomes are log-transformed to mitigate the influence of extreme values from the right tail of the sentence distributions (See e.g., Figure A4). Here I report results from two alternative models that do not transform the outcomes. Table A6 reports results using negative binomial regression on the un-transformed outcomes and Table A7 reports results using OLS regression on the un-transformed outcomes. Although some estimates are less precise than those from the main specification, both sets of results suggest that the substantive conclusions are largely unchanged - Tibetan, Uyghur, and Yi defendants on average receive longer predicted sentences across most, if not all, crime categories and both measures of sentence lengths (average and most likely sentence) and Hui defendants also receive longer sentences for nationwide drug crimes.

FIGURE A4. DENSITY PLOT OF AVERAGE SENTENCES FOR HOMICIDES AND AGGRAVATED ASSAULTS, PUBLIC DISTURBANCE



## TABLE A6. NEGATIVE BINOMIAL RESULTS

| | Drug: Yunnan | | Drug: Nation | | Homicide | | Fraud | | Public Disturbance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely |
| Bai | −0.007* | −0.006 | 0.005 | 0.009 | 0.000 | 0.003 | 0.000 | −0.007 | 0.006 | −0.007 |
| | (0.004) | (0.005) | (0.004) | (0.008) | (0.002) | (0.005) | (0.002) | (0.007) | (0.020) | (0.007) |
| Hui | 0.000 | 0.010 | 0.018*** | 0.024** | 0.005** | 0.003 | −0.001 | −0.001 | 0.001 | −0.001 |
| | (0.006) | (0.010) | (0.004) | (0.010) | (0.002) | (0.005) | (0.003) | (0.006) | (0.017) | (0.006) |
| Tibetan | 0.024*** | 0.019 | 0.035*** | 0.032** | 0.035*** | 0.013* | 0.028*** | 0.019* | 0.054** | 0.019* |
| | (0.005) | (0.012) | (0.009) | (0.015) | (0.008) | (0.007) | (0.007) | (0.011) | (0.024) | (0.011) |
| Uyghur | 0.007* | 0.003 | 0.024*** | 0.052*** | 0.036** | 0.035 | 0.012*** | 0.008 | 0.027* | 0.008 |
| | (0.004) | (0.006) | (0.007) | (0.015) | (0.015) | (0.027) | (0.003) | (0.011) | (0.015) | (0.011) |
| Yi | 0.008*** | −0.001 | 0.021*** | 0.033*** | 0.019*** | 0.020*** | 0.004 | 0.018 | 0.011 | 0.018 |
| | (0.003) | (0.005) | (0.005) | (0.009) | (0.007) | (0.007) | (0.006) | (0.013) | (0.016) | (0.013) |
| Zhuang | −0.004 | −0.012** | −0.003 | −0.001 | 0.014** | 0.003 | −0.004* | −0.013 | −0.012 | −0.013 |
| | (0.002) | (0.006) | (0.004) | (0.012) | (0.006) | (0.009) | (0.002) | (0.011) | (0.014) | (0.011) |
| Fixed Effect | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

***Note:*** * p < 0.1, ** p < 0.05, *** p < 0.01. All models control for judicial decision fixed effects. Columns with header "Average" are based on models for which the average sentence lengths are the outcome variable. Columns with header "Most Likely" are based on models for which the most likely sentence lengths are the outcome variable.

## TABLE A7. OLS RESULTS ON UN-TRANSFORMED OUTCOMES

| | Drug: Yunnan | | Drug: Nation | | Homicide | | Fraud | | Public Disturbance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely |
| Bai | −0.329* | −0.287 | 0.127 | 0.073 | 0.001 | 0.040 | −0.010 | −0.173 | 0.031 | −0.173 |
| | (0.191) | (0.260) | (0.087) | (0.176) | (0.021) | (0.060) | (0.049) | (0.189) | (0.106) | (0.189) |
| Hui | 0.005 | 0.323 | 0.438*** | 0.447 | 0.065** | 0.040 | −0.030 | −0.043 | 0.006 | −0.043 |
| | (0.253) | (0.369) | (0.098) | (0.288) | (0.030) | (0.064) | (0.073) | (0.118) | (0.086) | (0.118) |
| Tibetan | 1.052*** | 0.603 | 0.851*** | 0.408 | 0.418*** | 0.168* | 0.679*** | 0.389* | 0.290** | 0.389* |
| | (0.233) | (0.451) | (0.227) | (0.362) | (0.098) | (0.088) | (0.156) | (0.233) | (0.129) | (0.233) |
| Uyghur | 0.329* | 0.041 | 0.575*** | 0.894*** | 0.436** | 0.433 | 0.280*** | 0.105 | 0.144* | 0.105 |
| | (0.170) | (0.220) | (0.163) | (0.286) | (0.182) | (0.295) | (0.075) | (0.266) | (0.080) | (0.266) |
| Yi | 0.371*** | −0.092 | 0.507*** | 0.568*** | 0.233*** | 0.249*** | 0.095 | 0.335 | 0.060 | 0.335 |
| | (0.115) | (0.184) | (0.110) | (0.168) | (0.085) | (0.086) | (0.149) | (0.309) | (0.081) | (0.309) |
| Zhuang | −0.158 | −0.433** | −0.065 | −0.061 | 0.169** | 0.043 | −0.096* | −0.250 | −0.061 | −0.250 |
| | (0.098) | (0.213) | (0.096) | (0.312) | (0.072) | (0.115) | (0.053) | (0.270) | (0.071) | (0.270) |
| Fixed Effect | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

***Note:*** * p < 0.1, ** p < 0.05, *** p < 0.01. All models control for judicial decision fixed effects. Columns with header "Average" are based on models for which the average sentence lengths are the outcome variable. Columns with header "Most Likely" are based on models for which the most likely sentence lengths are the outcome variable.

## C.3. Results on charges

In addition to sentence length, the criminal sentencing software also predicts charges for each case. Specifically, for each case description, the software outputs a list of possible charges, together with the predicted probability (confidence) of each charge. Similar to sentence length, I estimate the following difference in means in outcomes, while blocking on judicial decisions, for each of the crime categories:

$$Y_{ij} = \beta_i E_i + \tau_j + \varepsilon_{ij} \tag{A1}$$

where $Y_{ij}$ is the outcome of interest. In particular, I study two outcomes for charges:

1. *charge confidence*: probability associated with the actual charge of the case, as determined by the judicial decision for that case.

2. *mis-classification*: a binary indicator of whether the predicted charge with the highest probability is the same as the actual charge, with 0 being the same and 1 being different.

The outcome *charge confidence* studies whether the software becomes less certain about the charges when defendants are of ethnic minority backgrounds. *Mis-classification* studies a more consequential outcome in which the software mis-predicts the actual charge of the case. Both types of outcome are referred to as quality of service harms of AI systems in the computer science literature - harms that occur when different social groups experience differences in the quality of service of AI systems (Bird et al., 2020).

Table A8 shows the results on charges across different ethnic minority groups and crime categories. For charge confidence, ethnic bias is less pronounced than sentence length. Ethnic cues signaling Tibetan and Uyghur identities reduce the charge confidence of the criminal sentencing software. The effects are significant for homicide and aggravated assault, fraud, and public disturbance cases, although the magnitudes are small (less than one percentage point difference). Other ethnic groups also receive less confident predictions for some cases

but the effects are generally not significant and consistent.

For mis-classification, Uyghur, Yi, and Zhuang defendants have higher mis-classification rates than Han defendants for both homicide and aggravated assault cases as well as fraud cases. Additionally, Tibetan defendants also have higher mis-classification rates for fraud cases. Uyghur and Zhuang defendants have lower mis-classification rates for public disturbance cases. Considering the overall mis-classification rates are very small for most crime categories ($1.5\% - 3.2\%$ except for fraud, which is 22.7%), the magnitudes of the bias are substantial (upward of 47% higher than Han defendants).

TABLE A8. RESULTS ON CHARGES

| | Drug: Yunnan | | Drug: Nation | | Homicide | | Fraud | | Public Disturbance | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Confidence | Misclassify | Confidence | Misclassify | Confidence | Misclassify | Confidence | Misclassify | Confidence | Misclassify |
| Bai | 0.001 | 0.001 | −0.001 | 0.003 | −0.001** | 0.000 | 0.000 | 0.002 | 0.001 | −0.004 |
| | (0.001) | (0.003) | (0.001) | (0.003) | (0.000) | (0.001) | (0.001) | (0.003) | (0.001) | (0.003) |
| Hui | −0.002 | 0.001 | −0.003*** | 0.000 | 0.000 | 0.000 | −0.004*** | 0.004 | 0.000 | −0.004 |
| | (0.001) | (0.003) | (0.001) | (0.003) | (0.000) | (0.001) | (0.001) | (0.004) | (0.001) | (0.003) |
| Tibetan | −0.001 | −0.001 | 0.000 | 0.000 | −0.003*** | 0.002 | −0.003** | 0.011*** | −0.005*** | −0.003 |
| | (0.001) | (0.003) | (0.001) | (0.002) | (0.001) | (0.001) | (0.001) | (0.004) | (0.002) | (0.003) |
| Uyghur | 0.000 | −0.002 | −0.003 | 0.000 | −0.006*** | 0.007*** | −0.002** | 0.005* | −0.003** | −0.006** |
| | (0.001) | (0.003) | (0.002) | (0.003) | (0.002) | (0.003) | (0.001) | (0.003) | (0.002) | (0.003) |
| Yi | 0.001 | 0.000 | −0.001 | 0.003 | −0.002** | 0.004** | −0.001 | 0.008** | −0.001 | −0.002 |
| | (0.001) | (0.002) | (0.001) | (0.003) | (0.001) | (0.002) | (0.001) | (0.004) | (0.001) | (0.003) |
| Zhuang | 0.000 | −0.002 | −0.003* | 0.005 | 0.001 | 0.004** | −0.005*** | 0.006* | 0.002* | −0.005* |
| | (0.001) | (0.002) | (0.001) | (0.003) | (0.001) | (0.002) | (0.002) | (0.004) | (0.001) | (0.003) |
| Fixed Effect | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

***Note:*** * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. All models control for judicial decision fixed effects. Columns with header "Confidence" are based on models for which the probability of the actual charge is the outcome variable. Columns with header "Misclassify" are based on models for which the binary indicator of misclassification is the outcome variable.

## C.4. Causal mechanisms

While the audit experiment establishes that changing the ethnic identity of defendants can affect the predicted sentences and charges, the potential causal mechanisms can be two-fold: 1) exposure to ethnic cues directly affects AI's predictions and 2) ethnic cues affect predictions through its correlations with other attributes such as education, gender, age, and employment status (Sen and Wasow, 2016; Davenport et al., 2023). Figure A5 provides a visual representation of the two causal mechanisms.

FIGURE A5. DIRECTED ACYCLIC GRAPH OF THE CAUSAL RELATIONS BETWEEN ETHNICITY AND PREDICTED SENTENCES

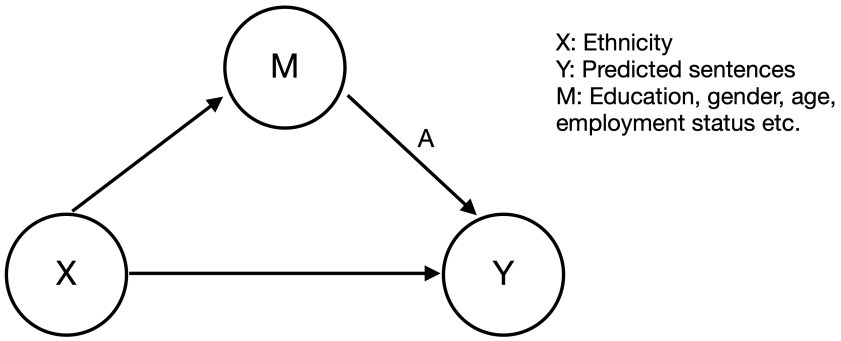



Table A9 shows that ethnicity is a significant predictor of defendants' education level, gender, age, and employment status, based on data on the criminal cases between 2019 and 2020. Ethnic minority defendants overall tend to have fewer years of education, be more employed, and younger than Han defendants. The correlation between ethnicity and age varies among different ethnic groups, with the Hui and Yi having a lower proportion of male defendants and the Bai, Tibetan, and Zhuang having a higher proportion of male defendants than the Han. Given these correlations, it is possible that ethnic cues also convey information about other attributes of the defendants.

To disentangle these two causal mechanisms, I performance an additional audit experiment to test if attributes such as defendant's education level, gender, age, and employment status have an effect on the predicted sentences (causal path A in Figure A5). Specifically,

I sample 1000 additional judicial decisions for each crime category. Similar to the main experiment, I construct case descriptions for which the only source of variation is one of education level, gender, age, and employment status. Unfortunately, the criminal sentencing software's public API was shut down while running the audit experiment, resulting in data being collected only on education level and gender and for crime categories related to nationwide drug cases and homicides and aggravated assaults.

TABLE A9. ASSOCIATIONS BETWEEN ETHNICITY AND EDUCATION, GENDER, AND EMPLOYMENT STATUS

| | Attributes | | | |
|---|---|---|---|---|
| | Education | Gender | Unemployed | Age |
| Bai | −0.094 | 0.051** | −0.097*** | −3.341* |
| | (0.216) | (0.022) | (0.030) | (1.757) |
| Hui | −0.406*** | −0.023* | −0.007 | −0.863 |
| | (0.121) | (0.014) | (0.017) | (0.802) |
| Tibetan | −0.261 | 0.046** | −0.113*** | −8.462*** |
| | (0.411) | (0.023) | (0.030) | (1.335) |
| Yi | −1.191*** | −0.112*** | −0.117*** | −7.026*** |
| | (0.093) | (0.015) | (0.012) | (0.713) |
| Zhuang | −0.585*** | 0.030*** | 0.015 | −2.737*** |
| | (0.062) | (0.008) | (0.012) | (0.392) |
| Num.Obs. | 57 800 | 71 484 | 72 720 | 30 983 |
| R2 Adj. | 0.003 | 0.002 | 0.001 | 0.003 |

***Note:*** * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors. Variable *Education* is coded in years. *Gender* is a binary variable with 1 being male and 0 being female. *Unemployed* is also a binary variable with 1 being unemployed. *Age* is coded in increment of 1.

Table A10 presents results of the effects of education level and gender on predicted sentence. For both nationwide drug cases and homicides and aggravated assaults, the effects of education level on predicted sentences are small (1.2 years change in education, the largest estimate from Table A9, results in $0.24\% - 0.76\%$ change in predicted sentence lengths) and not significant. The effects of gender on predicted sentences are not consistent: for drug cases, male defendants receive longer predicted average sentences but there is no effect for most likely sentences; for homicides and aggravated assaults, male defendants receive longer most likely sentences but no different average sentences. The magnitudes associated with

gender are small. Table A10 thus suggests that there is no strong evidence that ethnic cues result in longer predicted sentences either through education or gender.

TABLE A10. EFFECTS OF EDUCATION LEVEL AND GENDER ON PREDICTED SENTENCE

| | Drug: Nation | | Homicide & Assault | | Drug: Nation | | Homicide & Assault | |
|---|---|---|---|---|---|---|---|---|
| | Average | Most Likely | Average | Most Likely | Average | Most Likely | Average | Most Likely |
| Education | 0.006 | 0.004 | 0.002 | 0.002 | | | | |
| | (0.004) | (0.005) | (0.004) | (0.004) | | | | |
| Gender | | | | | 0.017*** | 0.000 | 0.002 | 0.038*** |
| | | | | | (0.002) | (0.006) | (0.002) | (0.007) |

***Note:*** * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. All models control for judicial decision fixed effects. Columns with header "Average" are based on models for which the average sentence lengths are the outcome variable. Columns with header "Most Likely" are based on models for which the most likely sentence lengths are the outcome variable.

# References


Baik, J. and L. Dai (2022). Administrative lawsuits dataset. *Working Paper*.

Bird, S., M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Davenport, L., H. Jefferson, and H. Rendleman (2023). Deconstructing race. *Working Paper*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hogan, A., E. Blomqvist, M. Cochez, C. d' Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. (2021). Knowledge graphs. *ACM Computing Surveys (CSUR) 54*(4), 1–37.

Liebman, B. L. (2015). Leniency in chinese criminal law: everyday justice in henan. *Berkeley J. Int'l L. 33*, 153.

Liebman, B. L., M. E. Roberts, R. E. Stern, and A. Z. Wang (2020). Mass digitization of chinese court decisions: How to use text as data in the field of chinese law. *Journal of Law and Courts 8*(2), 177–201.

Sen, M. and O. Wasow (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science 19*(1), 499–522.

Shamshad, F., S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu (2022). Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*.

Stern, R. E., B. L. Liebman, M. E. Roberts, and A. Z. Wang (2020). Automating fairness? artificial intelligence in the chinese courts. *Colum. J. Transnat'l L. 59*, 515.

Sun, Y., S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu (2019). Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Wu, X., M. E. Roberts, R. E. Stern, B. L. Liebman, A. Gupta, and L. Sanford (2022). Augmenting serialized bureaucratic data: The case of chinese courts. *Available at SSRN 4124433*.

Xiao, C., H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, et al. (2018). Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.