Automated Repression: Ethnic Discrimination in AI-assisted Criminal Sentencing

Eddie Yang*

April 2023

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, on average, defendants with ethnic minority status 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.

*Department of Political Science, University of California San Diego. Email: z5yang@ucsd.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 and surveillance (Beraja et al., 2021; 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 the predicted sentence lengths of 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 authoritarian regimes. 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 arrest, and how dangerous a suspect is considered to be to the public.[3] Local courts have also

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

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

[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/

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 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 choice of software is thus primarily determined by feasibility concerns. However, the company that develops the software in question holds partnerships with the Supreme People's Court as well as a top Chinese university. While I cannot 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 believe that the software is a strong contender for use in actual criminal trials and may be

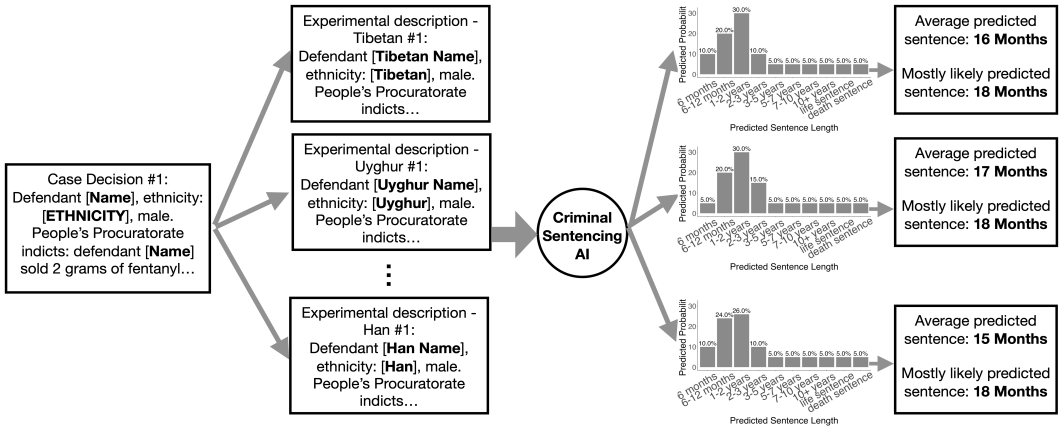
chinese-scientists-develop-ai-prosecutor-can-press-its-own

[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

representative of such efforts to automate court decisions.[5]

The criminal sentencing software takes the text of a 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 perform an audit experiment on the software. Specifically, I construct (fictitious but realistic) case descriptions for five categories of crimes based on real case descriptions extracted from judicial decisions from 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 predicted sentence lengths when the ethnic cues vary. A stylized example of the experimental

[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.

setup is shown in Figure 1.

Crime category selection. I choose five categories of crimes based on the existing literature and Chinese criminal laws, as well as 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 categories - 3) homicides and aggravated assaults, 4) frauds, and 5) crimes related to disturbing public order and state agencies (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 category, 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 crime categories (and 1000 judicial decisions for each category) and seven ethnic groups, I construct $1000 \times 5 \times 7 = 35000$ experimental case descriptions to

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

audit the criminal 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 crime categories:

$$Y_i = \sum_{j=1}^{6} \tau_j M_{ij} + \sum_{k=1}^{1000} \beta_k B_{ik} + \varepsilon_i \tag{1}$$

where Y_i 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.

M_j is a binary indicator for the jth ethnic minority group. The binary indicator for the Han majority group is omitted. B_k is a binary indicator for the kth 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

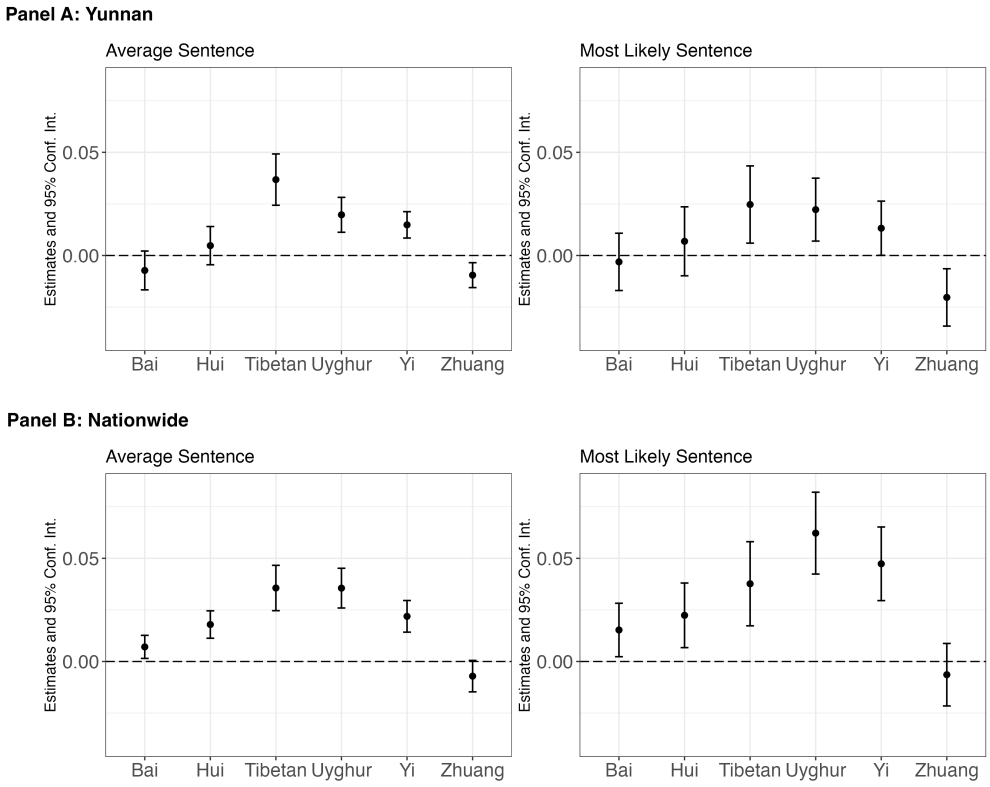
Drug Crimes Panel A and B in Figure 2 show the estimates for Yunnan and nationwide drug cases respectively. As shown in Panel A, Tibetan, Uyghur, and Yi defendants receive longer average predicted sentences, ranging from 1.5% to 3.7% longer than the baseline Han defendants.[8] This discriminating pattern 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)

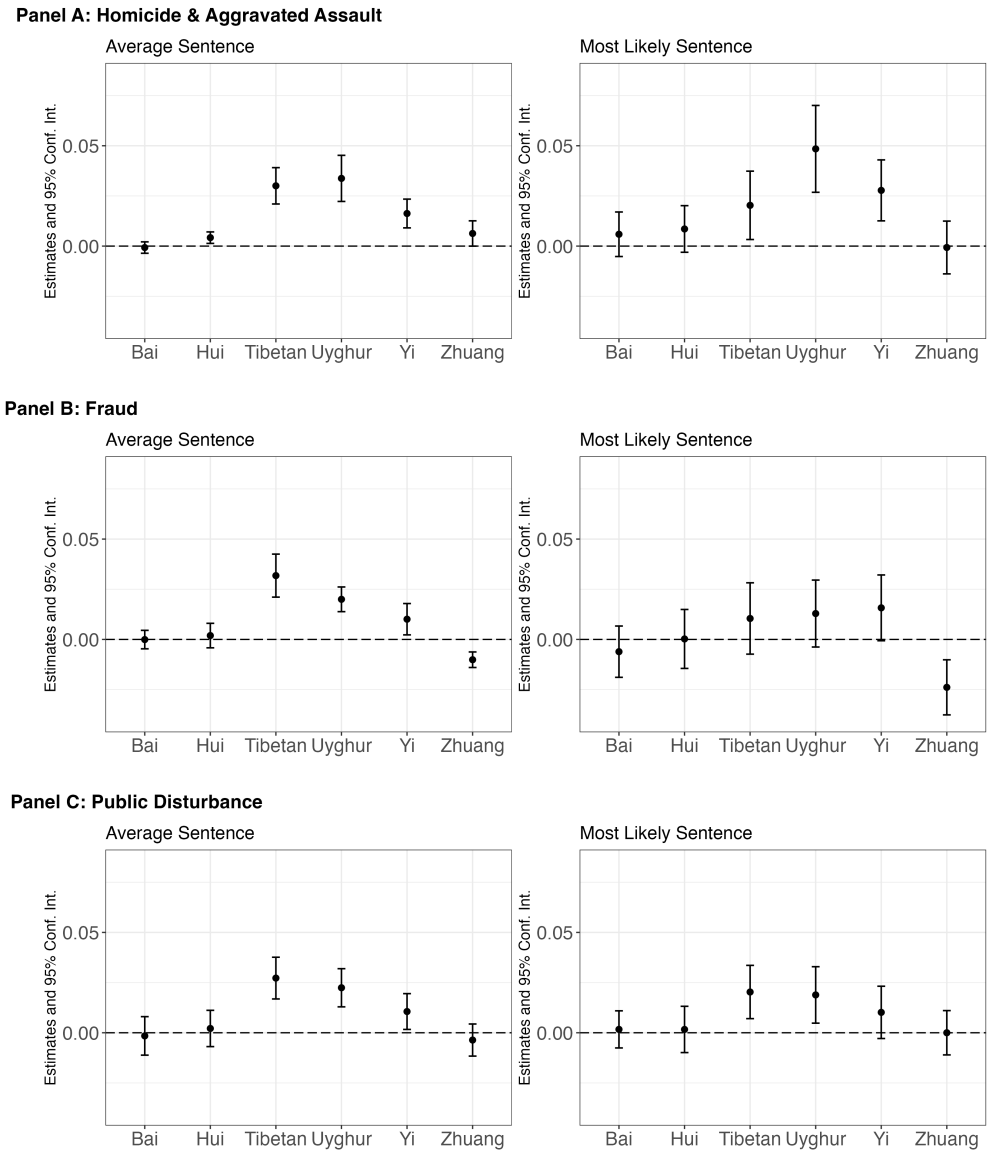
	Yunnan		Nationwide	
	Human Judges	AI	Human Judges	AI
Bai	+	.	+	.
Hui	.	.	+	+
Tibetan	.	+	.	+
Uyghur	.	+	.	+
Yi	+	+	+	+
Zhuang	-	-	.	.

+ 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 crime categories and both measures of sentence length, although some estimates are less precise. The results suggest that ethnic bias is present despite variations in the type, severity, and political nature of the crimes.

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 population in China, it is likely that there is relatively sparse information in the training data for

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

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

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*.