UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Automating Autocracy: Authoritarian Institutions and the Politics of Artificial Intelligence

Permalink

https://escholarship.org/uc/item/0h8054vz

Author

Yang, Eddie

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Automating Autocracy: Authoritarian Institutions and the Politics of Artificial Intelligence

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Political Science

by

Eddie Yang

Committee in charge:

Professor Margaret E. Roberts, Chair
Professor Claire Adida
Professor Kirk Bansak
Professor Ruixue Jia
Professor David Lake

2024

Copyright

Eddie Yang, 2024

All rights reserved.

The Dissertation of Eddie Yang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

Dissertation Approval Page iii

Table of Contents iv

List of Figures vi

List of Tables viii

Acknowledgements ix

Vita x

Abstract of the Dissertation xi

Introduction 1

Chapter 1 The Limits of AI for Authoritarian Control 5
1.1 Background: AI and Autocracy 11
1.2 Theory: The Digital Dictator's Dilemma 13
1.2.1 Data Deficiencies in Authoritarian Regimes 13
1.2.2 Automating Autocracy with Bad Data 15
1.2.3 Irony of the Free World 17
1.2.4 Summary 18
1.3 Data and Research Design 18
1.3.1 Repression-performance Trade-off 19
1.3.2 Data Augmentation 22
1.4 Results 23
1.4.1 More Repression, Worse Performance 23
1.4.2 Marginal Impact of More Training Data 26
1.4.3 Accuracy Improvement from Democratic Data Augmentation 26
1.5 Discussion 29

Chapter 2 Propaganda and AI 33
2.1 Government Manipulated Texts and Propaganda Bias 36
2.2 Propaganda in China as a Case 38
2.3 Propaganda in Open-Source Training Data 39
2.4 LLM Audit: Memorization 42
2.5 LLM Audit: Political Valence 45
2.6 Pre-Training Experiment 49
2.7 Discussion 52

Chapter 3 Ethnic Discrimination in AI-assisted Criminal Sentencing 56
3.1 Background 58

3.2 Data and Research Design . . . 59
3.3 Results . . . 62
3.4 Discussion . . . 64

Appendix A Supplementary Materials for Chapter 1 . . . 67
A.1 Further Details on Censorship AI . . . 67
A.1.1 Model Details . . . 67
A.1.2 Training Details . . . 68
A.1.3 Hyperparameters . . . 68
A.1.4 Hardware . . . 69
A.2 Further Details on the Weibo and the Twitter Data . . . 70
A.2.1 Political Sensitivity Service . . . 70
A.2.2 Details on the Weibo Data . . . 70
A.2.3 Details on the Twitter Data . . . 71
A.2.4 International Social Media Data Collection by Authoritarian Regimes . . 71
A.3 Additional Results . . . 73
A.3.1 Alternative Measure of Performance . . . 73
A.3.2 Error Rate Result for Models Trained on Larger Dataset . . . 75
A.3.3 Lower Cutoff, Data Leakage, Larger Model, and Alternative Model . . . 76

Appendix B Supplementary Materials for Chapter 3 . . . 78
B.1 Further Details on the Criminal Sentencing Software . . . 78
B.1.1 Use of AI in law in China . . . 78
B.1.2 User interface of the criminal sentencing software . . . 80
B.1.3 Prediction models of the criminal sentencing software . . . 81
B.1.4 Model performance . . . 82
B.1.5 Training data . . . 83
B.2 Further details on the audit experiment . . . 84
B.2.1 Graphical representation of the audit experiment . . . 84
B.2.2 Example of case description in English . . . 86
B.2.3 Details on collection of court rulings . . . 87
B.2.4 Minority names . . . 89
B.3 Additional results . . . 90
B.3.1 Effect by ethnic cues . . . 90
B.3.2 Additional results from alternative models . . . 91
B.3.3 Results on charges . . . 94
B.3.4 Causal mechanisms . . . 96

Bibliography . . . 100

LIST OF FIGURES

Figure 1.1. Stlyzed Example of Censorship AI Training and Testing 16

Figure 1.2. Experimental Design 21

Figure 1.3. Model Performance across Training Datasets 24

Figure 1.4. False Positive Rate vs. False Negative Rate 25

Figure 1.5. Performance Difference from Doubling Training Data 27

Figure 1.6. Effect of Twitter Data Augmentation 28

Figure 1.7. Principal Component Analysis of Content on Weibo and Twitter 32

Figure 2.1. Percent of Chinese CulturaX Matched to Propaganda - Political Topics . . . 41

Figure 2.2. Memorization Test Result 44

Figure 2.3. Audit Design 46

Figure 2.4. Human evaluation result 48

Figure 2.5. LLM-as-a-judge result 49

Figure 2.6. Pre-Training Experiment Result 51

Figure 3.1. Stylized Example of Experimental Setup 60

Figure 3.2. Estimates Across Ethnic Groups for Drug Crimes 63

Figure 3.3. Estimates Across Ethnic Groups for Other Crimes 66

Figure A.1. Model Performance across Training Datasets 74

Figure A.2. False Positive Rate vs. False Negative Rate 75

Figure A.3. Model Performance across Training Datasets 77

Figure B.1. Criminal Sentencing Software User Interface 80

Figure B.2. Data Collection in the Audit Experiment 85

Figure B.3. Example of Case Description 86

Figure B.4. Density Plot of Average Sentences for Homicides and Aggravated Assaults, Public Disturbance 92

Figure B.5. Directed acyclic graph of the causal relations between ethnicity and predicted sentences 97

LIST OF TABLES

Table 1.1. Summary Statistics of Training Datasets 22

Table 2.1. Percent of Chinese CulturaX Matched to Propaganda 41

Table 2.2. Memorization Examples 44

Table 2.3. Example templates used for political statements 45

Table 2.4. Responses to the question: Is China an autocracy? 53

Table 3.1. Comparison with Hou and Truex (2022) 64

Table A.1. Hyperparameters 68

Table A.2. Flagged Keywords 70

Table B.1. Comparison between actual and predicted sentence lengths 85

Table B.2. Summary Statistics on Judicial Decisions 88

Table B.3. Effect of Defendant Name 90

Table B.4. Effect of Defendant Ethnicity 92

Table B.5. Negative Binomial Results 93

Table B.6. OLS results on un-transformed outcomes 93

Table B.7. Results on Charges 97

Table B.8. Associations between Ethnicity and Education, gender, and employment status 99

Table B.9. Effects of education level and gender on predicted sentence 99

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of institutions, colleagues, friends, and family. I thank the Yankelovich Center, the Institute for Practical Ethics, and the UC Institute on Global Conflict and Cooperation at the University of California San Diego for financial support for my dissertation research. I thank the Center on Democracy, Development and the Rule of Law and the Institute for Human-Centered Artificial Intelligence at Stanford University for their generous funding during my time at Stanford through the CDDRL/HAI predoctoral fellowship. I thank the Mortara Center for International Studies at Georgetown University for its generous funding during my time at Georgetown through the Global Political Economy Project predoctoral fellowship.

I owe a great debt to many colleagues and friends in the profession who helped shape my dissertation and lent their generous support in various ways. In particular, thanks to Larry Diamond, Henry Farrell, Anita Gohdes, Genia Kostka, Kathleen McNamara, Marc Meredith, Salma Mousa, Gareth Nellis, Abraham Newman, Glen Weyl, and Yiqing Xu. My dissertation committee – Margaret Roberts, Claire Adida, Kirk Bansak, David Lake, and Ruixue Jia - have been the best committee I could wish for. Special thanks to my Chair, Molly, whose idea sparked my dissertation and from whom I learned a tremendous amount. I also thank the many other kind and generous people who made my graduate study an amazing experience.

Thanks to my girlfriend, Kerry, for her unwavering support and unconditional love. Thanks to my late cat, Brown, for his companionship. Thanks to my parents, who have always been my source of strength and inspiration. I dedicate this dissertation to them.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Waight, Hannah; Eddie Yang; Yuan, Yin; Messing, Solomon; Roberts, Margaret E.; Stewart, Brandon M., and Tucker, Joshua A. The dissertation author is the primary investigator and author of this material.

VITA

2016	Bachelor of Arts, University of California Los Angeles
2017	Master of Arts, University of Chicago
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

The Impact of U.S.-China Tensions on U.S. Science (with Ruixue Jia, Margaret E. Roberts, Ye Wang). *Proceedings of the National Academy of Sciences*, Forthcoming.

The Authoritarian Data Problem: How the Politics of Data Shapes AI (with Margaret E. Roberts). *Journal of Democracy*, 2023 (Invited Contribution).

Hierarchically Regularized Entropy Balancing (with Yiqing Xu). *Political Analysis*, 2023.

Censorship of Online Encyclopedias and its Political Effect on NLP Models (with Margaret E. Roberts). *ACM FAccT*, 2021.

ABSTRACT OF THE DISSERTATION

Automating Autocracy: Authoritarian Institutions and the Politics of Artificial Intelligence

by

Eddie Yang

Doctor of Philosophy in Political Science

University of California San Diego, 2024

Professor Margaret E. Roberts, Chair

Trained on massive amounts of data, Artificial Intelligence (AI) has achieved amazing results in generating realistic texts, images, videos, and more. However, the data that gives AI its capabilities - data produced by internet users, institutions, and organizations across the world - inevitably contains the biases and politics that many have hoped AI could overcome. This dissertation studies the way one kind of politics - authoritarian politics - gets embedded in AI's training data and its downstream implications. It seeks to answer the following questions: 1) How do authoritarian institutions shape the behavior and performance of AI? 2) What is the effect of AI on authoritarian politics? 3) how does the interaction between authoritarian politics and AI spill over into democracies?

Using novel experiments on commercial AI systems, open-source large language models, and open-source training data, this dissertation presents both theoretical and empirical evidence on how authoritarian politics serve to enhance but also constrain the utility of AI for repressive ends. Furthermore, it shows how authoritarian politics can spill over into democracies in subtle ways through AI.

Introduction

From generative Artificial Intelligence (AI) like ChatGPT to self-driving cars, AI has transformed the modern way of life. At the same time, authoritarian governments have increasingly used AI for repressive purposes, such as digital surveillance and automated censorship. This has generated much heated policy debate on AI and its potential for digital authoritarianism, despite a dearth of theory and empirical research.

I provide both theoretical and empirical contributions to this emerging topic. In particular, I seek to answer three questions in this dissertation:

1. How do authoritarian institutions shape the behavior and performance of AI?

2. What is the effect of AI on authoritarian politics?

3. How does the interaction between authoritarian politics and AI spill over into democracies?

I study these questions primarily in the contexts of China and the United States, through experiments on commercial AI applications and analysis of large-scale text data.

As AI is trained on human data, authoritarian institutions, such as censorship and propaganda, can affect the output of AI by changing what is in the data. Additionally, under the shadow of repression, people's strategic behavior, for example - preference falsification and self-censorship, can also skew the data generating process. I embed AI in authoritarian politics and propose a theory of how the politics of data can both strengthen and weaken AI's utility for repressive ends. The core insight of this dissertation is that AI's performance depends on

the quality of training data and authoritarian politics can affect data quality by either adding or suppressing politically relevant information in the data generating process.

The three chapters of this dissertation illustrate the theory by studying how 1) authoritarian institutions can hamper AI's ability to automate control and repression in authoritarian regimes, 2) propaganda and censorship can influence the output of AI in democracy, and 3) AI serves to automate discrimination through biased data.

The first chapter is motivated by a peculiar phenomenon I discovered during my fieldwork working in a Chinese technology company: instead of relying on AI, social media platforms often resort to blunt information control measures during times of crisis when there is an outpouring of public grievances. This is especially true for languages used by ethnic minorities. I propose a theory of *the digital dictator's dilemma* to explain this phenomenon. The dilemma describes a repression-information trade-off in which the repressive nature of authoritarian regimes distorts the production of data and reduces the amount of useful information in AI's training data. Faced with repression and censorship, citizens, including minority groups, falsify their preferences and self-censor, thus substantially reducing the amount of politically sensitive information in the training data for AI. Such strategic behavior degrades the accuracy of AI, especially during periods of crisis when newly-generated data differs significantly from the training data as people stop self-censorship and reveal their true preferences.

Leveraging on the information I gathered during fieldwork, I use a large-scale AI experiment to empirically test the digital dictator's dilemma. Specifically, I replicate how Chinese technology companies train censorship AI by using a similar AI algorithm and millions of real-world social media posts as training data. By changing the distribution of the training data to represent different degrees of preference falsification and self-censorship, I show that the accuracy of censorship AI decreases when there is more strategic behavior, with a larger drop in performance during times of crisis. Furthermore, simply using more training data does not solve the problem, However, augmenting the training data with data from democracies, in this case data from Chinese Twitter users, can partially improve the performance of censorship AI.

Finally, I use a text analysis method to show that the limits of democratic data augmentation are due to differences in the online discourse between domestic and international netizens.

This chapter draws on the classic literature of the dictator's dilemma[1] as well as citizens' strategic behavior in authoritarian states[2] but adapts them to the new domain of AI. By endogenizing AI in authoritarian politics, it challenges an implicit assumption of the existing literature on digital authoritarianism that more data means more accurate predictions from AI. In contrast, the chapter highlights how authoritarian institutions can degrade the performance of AI by distorting the data generating process. Additionally, the chapter adds to our understanding of the politics of information in authoritarian regimes. While the existing literature has focused on *domestic* mechanisms of information gathering, such as elections and strategic (non-)censorship, the paper demonstrates one way through which autocrats can integrate *international* sources of information (e.g., Twitter) to boost authoritarian control by using such information to train more accurate censorship AI, while excluding citizens from accessing such information.

While the first chapter examines how data produced under the constraints of authoritarian institutions limits the performance of AI in an autocracy, the second chapter studies how data from authoritarian regimes can contaminate AI in democracies. Specifically, I show the way censorship and propaganda can have a measurable impact on the output of AI. I use an audit experiment on large language models, including ChatGPT, and computational analysis on billions of text in AI's training data to show that the presence of propaganda in the training data skews the output of AI to be more in line with the views of the autocrats, resulting in AI generating favorable assessments of authoritarian political events and figures, such as for Mao Zedong and Joseph Stalin. This chapter contributes to the literature on the political bias of AI but is distinct from existing studies by focusing on sources of bias from non-democratic contexts.

Chapter 3 studies a commercial AI used to automatically sentence criminal defendants in China. The empirical setting represents a case in which the authoritarian government wants to

[1]Wintrobe, R. (2000). The political economy of dictatorship. Cambridge University Press.

[2]Kuran, T. (1997). Private truths, public lies: The social consequences of preference falsification. Harvard University Press.

use AI to reduce the workload of judges and ensure equality and justice in sentencing decisions. Using an algorithm audit experiment in which I change the name and ethnic status of the defendants and ask the criminal sentencing AI to make predictions on fictitious but realistic cases, I show that the goal of equality and justice in sentencing decisions is likely soiled by biased training data - defendants with ethnic minority name and status receive longer predicted sentences than Han majority defendants who are described to have committed the exact same crimes. Using the existing literature on biases from judges in China as a reference, I show that AI not only replicates human biases but also introduces new biases likely not seen in human data.

Together, the three chapters introduce a new way of thinking about the interplay between AI and authoritarian politics. The theory of the politics of data and how it affects AI endogenizes technology in authoritarian politics. It introduces nuances to the dominant view that AI is the silver bullet for autocrats and suggests that the seeming comparative advantage of authoritarian regimes in data collection for training AI may be a mirage, when the quality of data is also considered. Furthermore, one implication of this dissertation is that the flow of data, both from democracy to autocracy and vice versa, has important consequences for AI output and performance.

Chapter 1

The Limits of AI for Authoritarian Control

Abstract

An emerging literature suggests Artificial Intelligence (AI) can greatly enhance autocrats' repressive capability and strengthen authoritarian control. This paper argues that AI's ability to do so may be hampered by existing repressive institutions. In particular, I suggest that autocrats suffer from a "Digital Dictator's Dilemma," a repression-information trade-off in which citizens' strategic behavior in the face of repression diminishes the amount of useful information in the data for training AI. This trade-off poses a fundamental limitation in AI's usefulness for serving as a tool of authoritarian control - the more repression there is, the less information there will be in AI's training data, and the worse AI will perform. I illustrate this argument using an AI experiment and a unique dataset on censorship in China. I show that AI's accuracy in censorship decreases with more pre-existing censorship and repression. The drop in AI's performance is larger during times of crisis, when people reveal their true preferences. I further show that this problem cannot be easily fixed with more data. Ironically, however, the existence of the free world can help boost AI's ability to censor.

From facial recognition technologies used in digital surveillance systems to algorithms for automated censorship, Artificial Intelligence (AI) seems to be a powerful addition to the autocrat's toolkit that strengthens their control (Feldstein, 2019*b*; Xu, 2021). Given its repressive potential, large-scale uses of AI for authoritarian control have quickly become a reality. For instance, China has used AI-powered surveillance and censorship technologies to restrict citizens' freedom of expression. Similarly, leaked documents show that Russia is developing AI systems

to monitor and censor its internet.[1] Even for authoritarian regimes without the technological capacity to develop their own repressive AI systems, the export of such systems from countries such as China, the U.S., and Israel has expanded digital authoritarianism on a global scale (Feldstein, 2019*a*; Beraja et al., 2023). The dual effect of AI's repressive capability and the authoritarian regimes' seeming comparative advantage in collecting large amounts of data for training AI (Lee, 2018) has led to alarms that AI can enable digital dictators to achieve a level of control unimaginable to their counterparts of other times (Kendall-Taylor, Frantz and Wright, 2020; Beraja et al., 2021).

Yet despite this seeming dystopia of AI-powered digital authoritarianism, we still see the old, brute-force repression and blanket information control measures being used, especially during such critical times as popular protests and uprisings. For example, the Iranian government shut down the internet for the entire country multiple times in 2022, despite being equipped with the latest surveillance and censorship technologies from China.[2] Even for China, often touted as the most AI-savvy authoritarian regime, the outpouring of public grievances has at times proven to be too difficult a task for AI to handle and blanket censorship had to be used.[3]

Why does AI struggle in enforcing repression and information control while achieving much more impressive results in other tasks such as chat and question answering? Why do autocrats shift away from AI during times of crisis - times when autocrats are expected to rely on it the most?

I argue in this paper that authoritarian institutions limit AI's repressive capabilities, making it less omnipotent than scholars have previously argued. The key insight is that AI's capabilities are contingent on the data it is trained on. To effectively enforce control and repression, AI requires enough politically relevant information in its training data. However,

[1] Dasha Litvinova. The Associated Press. "The cyber gulag: How Russia tracks, censors and controls its citizens." May 23, 2023. https://bit.ly/3DRObFl

[2] Nima Khorrami. The Diplomat. "How China Boosts Iran's Digital Crackdown." October 27, 2022. http://bit.ly/45kQnSd

[3] See e.g., Quartz, "China's highest profile #MeToo case shows limits of censorship." November 3, 2021. https://qz.com/2083922/chinas-highest-profile-metoo-case-shows-limits-of-censorship

institutions of control and repression inherently restrict both the quantity and quality of such information. For instance, under the shadow of authoritarian institutions, citizens' strategic behavior, such as preference falsification and self-censorship, can corrupt the data needed for AI to learn to automate repression. The inherent tension between the corruption of the data generating process by authoritarian institutions and AI's reliance on data for performance gives rise to what I call the *digital dictator's dilemma*: to enable AI to automate authoritarian control, autocrats need to collect better quality data by relaxing repression and information control but doing so can jeopardize the autocrats by allowing dissent and opposition and in the worst case defeat the very purpose of using AI.

To further develop the theory and derive testable hypotheses, I focus on the use of AI for censorship as a case study. In this setting, preference falsification and self-censorship create a missing data problem for AI - politically sensitive and censorable content is suppressed in the data generating process, thus reducing the amount of relevant data necessary for AI to learn to automate censorship. The negative impact on AI's performance is even more pronounced during times of crisis when citizens suddenly reveal their true preferences (Kuran, 1991). Furthermore, I argue that the missing data problem cannot be easily fixed by simply collecting more data or using technical solutions such as bigger models to boost AI's performance. Ironically, however, the theory implies that data leakages from the "free world" - data generated from democracies that is not subject to the same political constraints - can partially mitigate the missing data problem and help boost AI's ability to automate censorship.

To empirically test the theory, I use a large-scale AI experiment to compare the accuracy of censorship AI systems trained on data with different degrees of missingness as a result of preference falsification and self-censorship. To construct the training data, I first combine two datasets of Chinese social media posts, totaling more than 10 million in size, from previous studies (Fu and Zhu, 2020; Hu et al., 2020). I then use an automated censorship service from a Chinese technology company to label the political sensitivity of each post. The social media posts and their sensitivity scores allow me to model different degrees of missingness by

constructing training datasets with different distributions along the political sensitivity dimension. For example, to model the case in which the regime is highly repressive and there is a high degree of preference falsification and self-censorship, the training dataset will have few social media posts with high political sensitivity scores.

With each training dataset, I train a deep-learning model using the same AI technologies as technology companies to automatically predict censorship. The accuracy of the censorship AI models is then evaluated using two types of test datasets: *status quo* and *crisis*. The status quo test set is sampled from the same distribution as the models' respective training datasets. It represents the situation in which the test environment is the same as the training environment. In contrast, the crisis test set models times of crisis in which people stop self-censoring and reveal their true preferences. It is thus sampled from the full distribution of the unaltered data.

To test the effect of having more data on the performance of AI, I repeat the above procedure but double the size of the training datasets. To test the effect of data leakages from the free world, I construct a dataset of social media posts by Chinese users on Twitter on the same topics and posted during the same period as the domestic data. The political sensitivity of the Twitter data is also obtained using the automated censorship service. The Twitter data is then used to augment the original training datasets. Notably, the full distribution of the Twitter data is used for all training datasets without altering (assuming no self-censorship and preference falsification on Twitter). A new set of censorship AI models are trained using the augmented training datasets and their accuracy is compared with the original models.

The AI experiment establishes three sets of results. One, as the regime becomes more repressive and there is more preference falsification and self-censorship, the resultant increase in data missingness causes the accuracy of the censorship AI model to fall. The drop in AI's performance is larger during times of crisis than times of status quo, as true preference revelation during crisis causes a larger mismatch between the distributions of the training and test datasets. Two, doubling the amount of domestic training data has a marginal effect on improving the accuracy of censorship AI, as sampling more data points under the same data generating process

does not address the data missingness problem. Three, augmenting domestic training datasets with (international) data from Twitter helps improve the accuracy of censorship AI. However, the improvement from Twitter does not fully close the accuracy gap caused by the missing data problem. Through text analysis of the domestic and international data, I give suggestive evidence that this is due to differences in the discourse between the two data sources.

Taken together, the theory and the empirical results highlight that the classic data issues caused by citizens' strategic behavior in authoritarian regimes not only hamper traditional dictators but also the new generation of digital dictators who wish to use AI for authoritarian control. In the context of censorship, the paper demonstrates the limits to automated censorship with AI and how such limits are borne out of existing authoritarian institutions. The findings of the paper correspond with news reports of authoritarian regimes' struggle with data in training censorship AI[4] and offer potential explanations for why autocrats at times resort to more brute force information control measures in place of AI, especially during crisis.

Although focusing on AI, the paper builds on theories of authoritarian politics, particularly theories of citizens' strategic behavior in authoritarian states (Kuran, 1997; Wintrobe, 2000; Jiang and Yang, 2016; Roberts, 2018; Shen and Truex, 2021), and expands them to the new domain of AI. In doing so, the paper reveals new insights about how authoritarian institutions tasked with repression and censorship can influence the performance and utility of AI by changing the data generating process of AI's training data. The findings contribute to our understanding of the relationship between digital technology and autocratic rule. While the existing literature has shown that AI can be useful for autocrats, the paper highlights that the general equilibrium effect of AI may not be as favorable toward autocrats as the existing literature has argued. Specifically, the paper challenges an implicit assumption of the existing literature on technology and autocracy - that more data means more accurate predictions from AI (Diamond, 2019; Xu, 2021). In contrast, the paper shows that (distributionally) biased data can lead to less accurate

[4]For example, technology companies in China have struggled to collect enough data from ethnic minority groups as a result of heavy self-censorship and preference falsification. See e.g., Shen Lu. Protocol. "I helped build ByteDance's censorship machine." February 18, 2021. https://bit.ly/45pZQqL.

predictions and simply adding more biased data will not solve the problem. Ironically, however, the paper points out that biases created by preference falsification and self-censorship may be partially mitigated by the presence of a free and robust society outside of the authoritarian regime.

More broadly, the paper contributes to our understanding of autocrats' strategy for information control and regime survival. As modern autocrats move away from mass repression and rely more on the manipulation of the information environment (Guriev and Treisman, 2019, 2020), censorship and information gathering have become essential for authoritarian control. Traditionally, autocrats face a trade-off: in order to gather necessary information for regime survival, autocrats need to relax restrictions on the freedom of expression and the press, but doing so risks generating dissent and allowing citizens to learn about the regime's corruption or incompetence (Egorov, Guriev and Sonin, 2009; Egorov and Sonin, 2020). While existing studies have focused on how *domestic* mechanisms, such as elections (Cox, 2009; Rozenas, 2010; Miller, 2015) and strategic (non-)censorship (King, Pan and Roberts, 2013; Lorentzen, 2014; Chen and Xu, 2017), allow autocrats to strike a delicate balance in the trade-off, the role of *international* sources of information (e.g., diaspora and independent media) has been less explored.

The paper demonstrates one way through which autocrats can integrate international sources of information (e.g., Twitter) to boost authoritarian control by using such information to train more accurate censorship AI, while excluding citizens from accessing such information. Qualitative evidence suggests the use of international sources of information is already happening systematically, on a large scale, and across authoritarian regimes.[5] On the other hand, the paper also points out the limits of such an approach for autocrats, as the paper joins an emerging literature (Esberg and Siegel, 2021) in highlighting the difference in content between domestic and international sources of information.

Finally, the paper speaks to a larger policy debate on regulations on data privacy and AI. While existing studies have highlighted the essential role of data on AI innovation (Beraja,

[5] See Appendix A.2.4 for details.

Yang and Yuchtman, 2020) and how authoritarian regimes have a comparative advantage in collecting more data (Lee, 2018), this paper, echoing recent work (Farrell, Newman and Wallace, 2022), shows that such comparative advantage may be a mirage if the quality of data is taken into consideration. Furthermore, the findings of the paper imply that, to prevent strategic and adversarial behavior from citizens that degrades data quality, stronger data privacy regulations may be necessary to assure citizens that their rights and data are protected. Additionally, stronger data privacy regulations can help reduce data leakages and prevent autocrats from leveraging on such data.

1.1 Background: AI and Autocracy

In this paper, AI refers to machines and computer programs that are capable of performing tasks that typically require human intelligence. Examples of AI performing "intelligent" tasks include playing chess, recognizing faces, and automating censorship. Essentially, AI can be seen as a technology of prediction (Agrawal, Gans and Goldfarb, 2019): predicting the best next move in chess, the identity of a face, and whether content should be censored or not.

A key underlying technology that powers much of recent AI advancement is deep learning - complex algorithms that extract relationships between data. In one paradigm, training deep learning algorithms follows a two-stage process: 1) a pre-training stage where deep learning algorithms are trained to obtain general capabilities and 2) a fine-tuning stage where the pre-trained model is adapted to a specific task using customized datasets. For example, to train a censorship AI, we can first obtain a pre-trained model, which is usually trained on large corpora of general text. The pre-trained model is then fine-tuned on a censorship-specific dataset to improve its ability to predict censorship. This paper focuses on data and training in the second stage as fine-tuning has a large impact on AI's performance on specific tasks.

Just like OLS regression, in the fine-tuning stage, deep learning algorithms take as input some characteristics X with their corresponding outcomes or labels Y and fit a function $Y = f(X)$.

Unlike OLS regression, deep learning algorithms generally do not pre-specify the relationship between X and Y but rather use a data-driven approach to learn the functional form of $f(\cdot)$. Additionally, deep learning algorithms are usually much more complex, involving upward of billions of parameters.

Recent successes with AI have transformed the modern way of life. Nowadays, people use voice assistants like Amazon's Alexa and Apple's Siri to automate tasks such as sending text; students rely on AI chatbots like ChatGPT for answers to questions and assignments; and people increasingly use self-driving technologies to assist with their driving. A key contributing factor to these successes is the availability of large, high-quality data. This enables deep learning algorithms to extract complex relationships that are essential for complicated tasks such as playing chess and carrying out conversations. For example, chess-playing AI AlphaGo Zero was trained on 4.9 million chess games (Silver et al., 2017) and AI chatbots such as ChatGPT are trained on trillions of words scraped from books and the internet.

Like other areas of society, political institutions have also incorporated the use of AI in their decision-making process. For example, 11 U.S. states and 178 additional counties in other states are using algorithmic risk assessment tools to assist judges in making bail decisions.[6] The state of Telangana in India has used facial recognition systems to verify voter identity, with more Indian states following suit in the upcoming elections. Perhaps even more so than democracies, authoritarian regimes such as China, Iran, and Russia have embraced AI to automate tasks such as surveillance (Xu, 2023), censorship, and meting out criminal sentences in place of judges (Yang, 2023). Yet despite AI's growing importance in politics, evaluations of its impact are rare in political science, with a few notable exceptions (Imai et al., 2020; Xu, 2021; Allie, 2023).

[6]Mapping Pretrial Injustice, "Where are Risk Assessments Being Used?" https://bit.ly/3s9lktG

1.2 Theory: The Digital Dictator's Dilemma

AI relies on data to acquire its predictive capabilities. To be good at chess, AI needs to learn from matches played by high-caliber players in training. Similarly, to predict censorship, AI needs sufficient censorable content in the training data. On the other hand, if the chess data lacks expert-level matches or if there is little censorable content, the performance of AI can be subpar.

Yet data quality issues are a classic problem in authoritarian regimes. Under the shadow of censorship and repression, citizens self-censor and falsify their preferences to avoid punishment for voicing opinions that the regime finds unpalatable (Kuran, 1997; Shen and Truex, 2021). Such strategic behavior not only reduces the amount of political information in the observed data that AI is trained on but also increases the difficulty of the prediction problem by diminishing the differences between content that is censorable and non-censorable. Without additional sources of information that are not subject to the same political constraints, AI's performance can be handicapped by the quality of data in authoritarian regimes. This section leverages theories of citizens' strategic behavior in the face of authoritarian institutions to explain in detail why digital dictators struggle to realize the repressive potential that AI promises.

1.2.1 Data Deficiencies in Authoritarian Regimes

That data quality can affect the performance of AI has been established by a substantial literature in computer science. A well-known example is the racial bias embedded in facial recognition systems. Previous studies have found that facial recognition systems systematically mis-recognize faces with darker skin tones at higher rates (Cook et al., 2019; Grother, Ngan and Hanaoka, 2019; Vangara et al., 2019; Robinson et al., 2020). The disparity in performance across racial groups has been attributed to an under-representation of African-American faces in the training data (Buolamwini and Gebru, 2018) and subsequently a large body of work has focused on increasing the training data quality through more racially balanced samples (Kärkkäinen and

Joo, 2019; Wang, Zhang and Deng, 2021).

While computer science research has focused on data quality in the U.S. and other democracies, less studied is how data quality is affected in authoritarian regimes. In particular, two mechanisms serve to degrade data quality and subsequently affect AI performance in authoritarian regimes. One, citizens can falsify their public preferences to pander to the autocrat (Kuran, 1997). While citizens may hold grudges against the autocrat, the regime, or specific policy in private, the fear of censorship and repression can steer citizens away from voicing their private preferences but rather "toe the party line" in public (Shih, 2008; Wedeen, 2015). The suppression of such grudges in the data generating process and the congregation of publicly expressed preferences create a mismatch between citizens' private preferences and the public opinion data that the autocrat observes (Jiang and Yang, 2016). For example, a number of studies have shown that publicly expressed popular support for authoritarian regimes is often higher than the actual level of support (Kalinin, 2016; Robinson and Tannenberg, 2019; Hale, 2022; Nicholson and Huang, 2022).

The second mechanism that negatively affects data quality in authoritarian regimes is self-censorship (Berinsky, 1999; Shen and Truex, 2021). Rather than falsifying their preferences, citizens engaging in self-censorship simply refrain from voicing any public opinion at all. By self-censoring, citizens avoid the psychological cost of falsifying their preferences (Crabtree, Kern and Siegel, 2020) while still able to avoid punishment from the regime.

Both preference falsification and self-censorship corrupt data quality on political preferences and attitudes by creating a missing data problem in which politically valuable but sensitive information is missing in the observed data. This skews the distribution of the observed data, creating a mismatch with the true distribution of public opinions. The severity of the missing data problem is a function of the cost of voicing dissent - the higher the cost, the less such information will be in the data (Tannenberg, 2022). In the authoritarian context, such cost comes in the forms of censorship and physical repression and is thus determined by the existing authoritarian institutions. This leads to the classic dictator's dilemma where autocrats have to

trade off censorship and repression with information gathering (Wintrobe, 2000).

1.2.2 Automating Autocracy with Bad Data

Bad data is bad. A slew of problems in authoritarian regimes has been attributed to bad or missing data, such as inefficient governance and policy implementation (Wallace, 2022; Trinh, 2023), indiscriminate repression (Gohdes, 2020), and even surprise breakdown of authoritarian regimes (Kuran, 1991; Lohmann, 1994). Yet little studied is the fact that the use of AI in politics suffers just as much, if not more, from bad data. Just as facial recognition systems trained predominantly on faces of light skin tones fail to recognize faces of darker skin tones, AI that is trained to automate repression and censorship can be crippled by bad data caused by citizens' strategic behavior of preference falsification and self-censorship.

Specifically, AI suffers two related problems from preference falsification and self-censorship: they 1) reduce sensitive information in the training data, and 2) increase the difficulty of the prediction task. To see why this is the case, consider the stylized example of censorship AI in Figure 1.1. In this example, AI's training data is composed of censorable and non-censorable content, which are generated from the unobserved underlying distribution of political sensitivity. As an example, sensitivity scores above 0.5 map onto censorable content that AI should flag and scores below as non-censorable. Because of preference falsification and self-censorship, the data generating process is right-censored (shaded region), in that content with high sensitivity scores will not be generated. This reduces the amount of data with political sensitivity above 0.5, resulting in a smaller amount of censorable content in the training data (Problem 1). Additionally, because the shaded region is censored, it reduces the overall distance (and increases the similarity) between censorable and non-censorable content, making the prediction problem more difficult (Problem 2).

Both of the above-mentioned problems become more severe as the level of repression and censorship increases and people falsify their preferences and self-censor more (the shaded

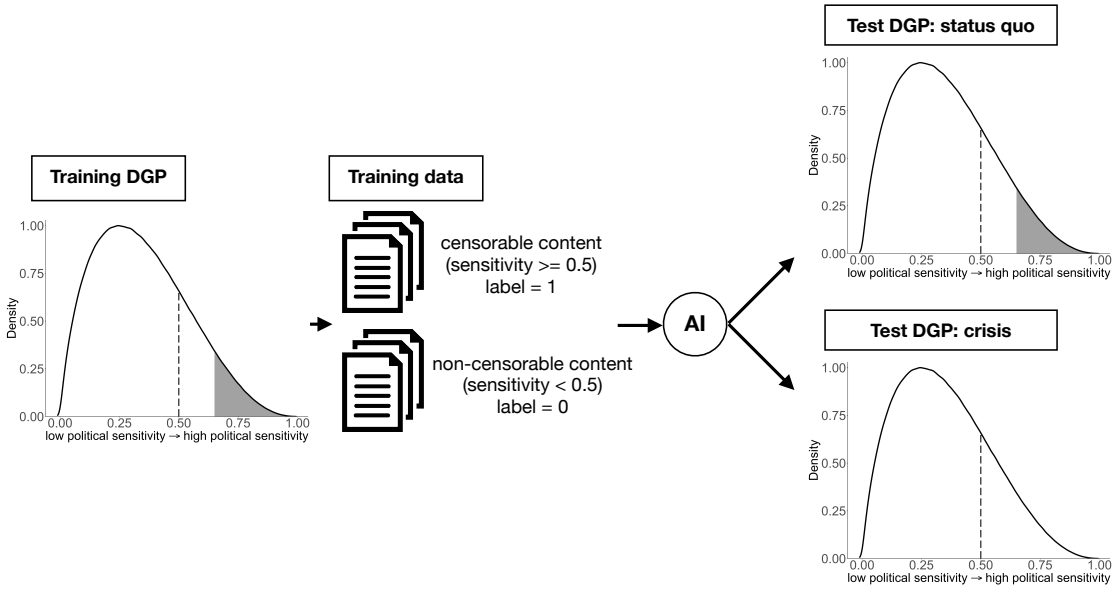

Figure 1.1. Stlyzed Example of Censorship AI Training and Testing

Note: The shaded regions in the data generating processes (DGP) of the training data as well as the status quo test data indicate right-censoring. This causes content with high sensitivity to be missing in the observed training and status quo data. Given the observed training data, censorship AI solves a binary classification problem of predicting whether content should be censored or not. Test data is used to evaluate the performance of censorship AI.

region becomes larger). Additionally, Problem 1 will cause a larger drop in the performance of AI during periods of crisis than during the status quo. Status quo refers to the situation in which the test data that is used to evaluate performance is drawn from the same distribution as the training data (Figure 1.1). It represents "business as usual" in authoritarian regimes, where the level of preference falsification and self-censorship is maintained. In contrast, crisis refers to the situation in which the test data is drawn from the full distribution without missing data. It represents times of political turmoil, such as protests and coups, when there is an information cascade and citizens do away with preference falsification and self-censorship (Lohmann, 1994). The mismatch between the distributions of the training data and the test data during crisis can further pull down the performance of AI.[7] This can be particularly bad for autocrats: autocrats need AI to perform the best during crisis but it is exactly in times of crisis that AI fumbles in performance. Therefore, not only does the digital dictator suffers from the trade-off between

[7]This is referred to as distribution shift in the computer science literature. See e.g., Storkey et al. (2009).

repression and AI performance, he faces the additional risk that the trade-off is exacerbated in the most dangerous times of his rule.

1.2.3 Irony of the Free World

Autocrats want to use AI to strengthen authoritarian control. What can they do in light of the repression-performance trade-off? A brute-force measure is to simply collect more data. However, the additional data will suffer from the same quality issues if it is collected from the same data generating process that is tainted by preference falsification and self-censorship. In other words, for prediction, sampling more from the biased distribution does not correct for the (distributional) bias. Once there is enough data for AI to learn about the biased distribution, simply collecting more data without changing the constraints under which citizens generate data should have a marginal impact on the performance of AI.

On the other hand, however, if autocrats can somehow collect data from the right-censored parts of the data generating process (i.e., content that is self-censored or that reflects citizens' private preferences), then the performance of AI can be improved. One way autocrats can do so is by collecting data that is not generated *domestically* but *internationally*, especially from democracies where citizens do not face the same political constraints. For instance, instead of solely relying on data from domestic social media to train a censorship AI, autocrats can augment such data with content from international social media, such as Twitter and Facebook. Doing so not only boosts the performance of AI but also keeps the level of repression and censorship unchanged domestically, thus potentially mitigating the digital dictator's dilemma. Qualitative evidence suggests that such practice is already used systematically, on a large scale, and across authoritarian regimes (Appendix A.2.4).

How well data augmentation from international sources works depends on how similar such data is to the right-censored parts of the data generating process. In particular, there needs to be sufficient overlap in the topics and semantics between international and domestic sources.

Furthermore, data from international sources needs to be diverse enough in terms of political sensitivity to cover the entire span of right-censoring in domestic sources. Existing evidence suggests that oversea sources of information are qualitatively different from domestic sources, both in terms of topical distribution as well as political sensitivity (Esberg and Siegel, 2021). Such differences will limit the effect of data augmentation on AI performance.

1.2.4 Summary

To summarize, I leverage theories of citizens' strategic behavior in authoritarian regimes to explain the performance of AI under different levels of repression and censorship. Specifically, the theory implies the following hypotheses.

Repression-performance trade-off:

1a. As repression and censorship increase and people engage in more preference falsification and self-censorship, the performance of AI will decrease.

1b. The drop in AI performance is larger during times of crisis than times of status quo.

Data augmentation:

2a. More data collection under the same data generating process has a marginal impact on performance.

2b. Data from international (especially democratic) sources can improve AI performance.

1.3 Data and Research Design

I choose AI that is used to automate censorship as the empirical setting. In this case, AI solves a binary classification problem: given a social media post, predict whether its label should be 0 (not censor) or 1 (censor). In practice, a censorship AI is trained by fine-tuning a pre-trained model with labeled censorship data. The pre-trained model is usually a general open-source

deep learning model and the labeled censorship data consists of social media posts with their corresponding censorship labels.

1.3.1 Repression-performance Trade-off

To test the theory's hypotheses on the repression-performance trade-off, the ideal empirical set-up would be to have multiple parallel worlds where the AI technologies are fixed but the data generating process is subject to varying degrees of preference falsification and self-censorship. The performance of censorship AI from these worlds can then be compared.

To approximate the ideal set-up, I use a large-scale AI experiment to replicate as close as possible the actual training of censorship AI models in practice. Specifically, the experiment uses 1) the same AI algorithm that technology companies use, 2) training data that consists of millions of real-world user-generated content, and 3) industry-level training procedures with state-of-the-art computing hardware. Using a unique dataset of Chinese social media posts for which the political sensitivity is known, the experiment compares the accuracy of censorship AI models trained with different versions of the dataset that vary on the distribution of political sensitivity.

To construct the training data, I combine two datasets of Chinese social media posts from previous studies (Fu and Zhu, 2020; Hu et al., 2020). The social media posts, totaling more than 10 million in size, are on the topics of COVID-19 and were posted on Weibo, a Chinese social media platform, during the early period of the COVID-19 pandemic (Dec. 2019 - Feb. 2020). I focus on the early period of the pandemic because this was when censorship of COVID-19 topics had not caught up[8] and therefore the social media posts have a relatively wide distribution of political sensitivity. The combined dataset serves as the basis from which I construct different versions of training dataset and use them to train censorship AI models specifically for COVID-19.

[8]According to one Chinese technology company, the technology to automatically censor COVID-19 topics was not put to use until around Feb. 27, 2020. bit.ly/3sbkYCZ.

To get the political sensitivity of the social media posts in the combined dataset, I use an automated censorship service from a Chinese technology company.[9] The service is sold to smaller social media companies to help conduct censorship. It takes the text of social media posts as input and outputs a political sensitivity score for each post that ranges from 0 to 1, with 1 being the most sensitive. The political sensitivity scores serve as the latent variable. I use the service default sensitivity score of 0.5 as the threshold to generate the censorship labels - social media posts with scores above 0.5 have a label of 1 (censorable) and posts with scores below 0.5 have a label of 0 (non-censorable).[10] The social media posts and their censorship labels can then be used as training data for censorship AI. Following standard industry practice, I down-sample social media posts with labels of 0 to partially account for the imbalance in the proportion of the two classes (0 and 1) of labels. Therefore, the main sample has a size of 1 million social media posts.

To model different degrees of data missingness due to preference falsification and self-censorship, I construct training datasets with different distributions of political sensitivity from the main sample. As Figure 1.2 shows, I construct five versions of training dataset with varying degrees of missingness. To model the case where there is no missingness, I use the entire sample as the training dataset. The other four versions use different thresholds (0.9, 0.8, 0.7, 0.6) above which the corresponding social media posts are missing from the training dataset.[11] A threshold of 0.6 means that only social media posts with sensitivity scores between 0 and 0.6 are in the training dataset. This models the most extreme case in which the regime is highly repressive and there is a high degree of preference falsification and self-censorship. Table 1.1 reports the summary statistics of the different versions of the training dataset.

For each version of the training dataset, I train a separate censorship AI model on it. Specifically, I use a Chinese version of BERT (Bidirectional Encoder Representations from

[9] See Appendix A.2.1 for more details.

[10] In the Appendix, I show results using a different threshold and the substantive conclusions remain unchanged.

[11] In the appendix, I allow imperfect preference falsification and self-censorship by allowing 10% of data from the missing part of the distribution to leak into the training datasets. The substantive conclusions remain unchanged.

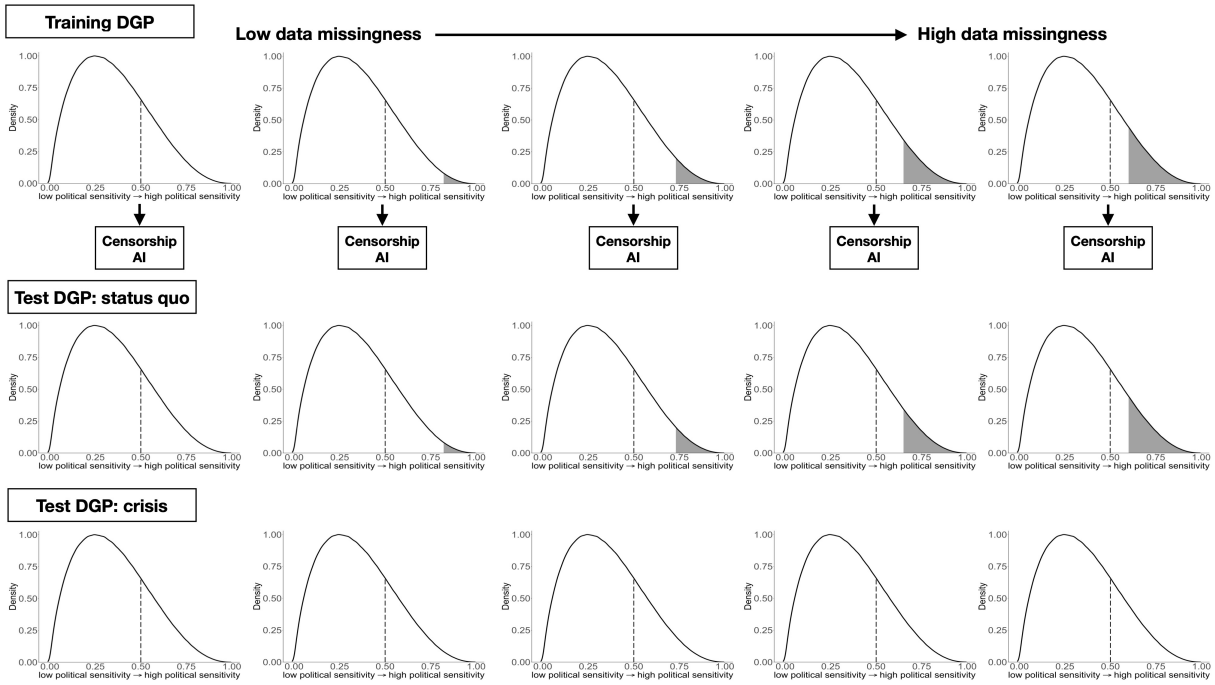

Figure 1.2. Experimental Design

Note: Graphical representation of the research design. There are five versions of training dataset corresponding to different degrees of missingness. The "status quo" test datasets are drawn from the same distributions of their corresponding training datasets. All "crisis" datasets are drawn from the full distribution. Note that this is a stylized representation. The shapes of the actual distributions are different from the graph.

Transformers; Devlin et al. 2018) as the pre-trained model and fine-tune it on the training datasets for censorship. BERT is a deep learning model with more than 100 million parameters and was first developed by Google. The Chinese version was developed by Chinese academic and industry research labs (Cui et al., 2020). Since its introduction, BERT has been one of the most popular deep learning models for prediction and is widely used in commercial applications.[12] To further address the imbalance in the two classes of labels in the training datasets, I weight each social media post by the inverse of the proportion of its label class in the specific version of the training dataset. Additionally, to account for the uncertainty from data sampling and the stochastic nature of the fine-tuning process, each version of the training dataset is used to train 20 models with the training data shuffled each time. This allows me to obtain uncertainty estimates

[12]In the Appendix, I provide details about the pre-trained model, specifically how it is used in practice for censorship, based on fieldwork in technology companies.

Table 1.1. Summary Statistics of Training Datasets

Training dataset	Version #1	Version #2	Version #3	Version #4	Version #5
Threshold	no missingness	0.9	0.8	0.7	0.6
No. of positive labels (censor)	256,188	191,904	148,465	101,782	57,585
No. of negative labels (not censor)	743,812	743,812	743,812	743,812	743,812

for model performance. More details about the training procedure are included in Appendix A.1.2.

To evaluate the performance of the different censorship AI models, I follow the theory and construct two sets of test data: status quo and crisis (Figure 1.2). The social media posts in the status quo test data are drawn with the same missingness as the corresponding version of the training dataset whereas the crisis test data is always drawn from the full distribution regardless of the version. To be able to compare performance evaluated on different test data, each test data is a balanced sample of 2500 positive labels and 2500 negative labels. To measure the performance of censorship AI, I use accuracy, defined as $\frac{\text{No. of correct predictions}}{\text{Total no. of predictions}}$, in the main text and report other measures of performance in the Appendix.

1.3.2 Data Augmentation

To test the effect of more data on AI's performance (hypothesis 2a), I follow the same experimental set-up as above but double the size of the initial sample from 1 million to 2 million while keeping the distribution of political sensitivity unchanged. A new set of censorship AI models are trained using the larger training datasets and their accuracy is compared with the original models.

To test the effect of data leakages from international sources (hypothesis 2b), I scraped all 558,322 social media posts by Chinese users on Twitter on the same COVID-19 topics and posted during the same period as the Weibo data. The political sensitivity of the Twitter data is also obtained through the automated censorship service. I then construct the Twitter dataset

of 270,000 tweets with political sensitivity scores above 0.5 and use it to augment the Weibo training datasets.[13] Notably, the full Twitter dataset is used for all training datasets without altering, assuming that there is no data missingness from international sources as a result of changing domestic repression levels. A new set of censorship AI models are trained using the augmented training datasets and their accuracy is compared with the original models.

1.4 Results

I first present evidence of the repression-performance trade-off and AI's performance gap on status quo and crisis test data. I then show evidence that adding more data to the training dataset has a marginal impact on AI's performance but augmenting with data from democratic sources results in a partial improvement in AI's censorship accuracy.

1.4.1 More Repression, Worse Performance

Figure 1.3 presents evidence of the repression-performance trade-off. It shows the accuracy of censorship AI models trained with datasets of varying degrees of missingness. The threshold (x-axis) indicates the political sensitivity score above which data is missing from the training datasets. The threshold of 1 means the training dataset has no missing data and the threshold of 0.6 has the most missing data. The model accuracy is evaluated on both the status quo and crisis test data.

Evaluations on the status quo data (blue) show that as the level of repression and data missingness increases, the accuracy of the censorship AI model decreases, with the worst-performing model being trained on the dataset with the most missingness. A similar downward trend is also observed for the crisis test data (red). Moreover, in line with the theory, the decrease in model accuracy is significantly larger in crisis, when people reveal their true preferences, than

[13] I exclude tweets with labels of 0 from the Twitter dataset as missingness in the Weibo datasets only comes from social media posts with positive labels.

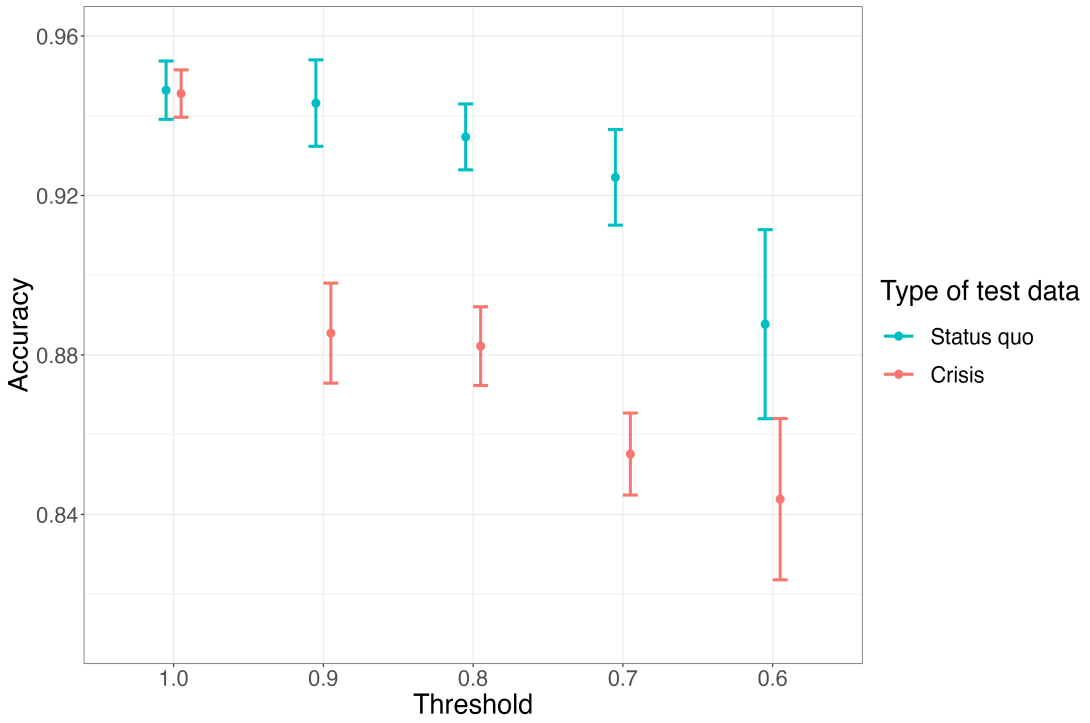

Figure 1.3. Model Performance across Training Datasets

Note: Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

in the status quo.

While accuracy serves as an indicator of the overall performance of censorship AI models, it does not reveal the type of error that the models make. Specifically, the models' errors can be false positives (where prediction is censorship but the actual label is non-censorship) or false negatives (where prediction is non-censorship but the actual label is censorship). The types of error have important implications for authoritarian rule, as false negatives (failing to censor) allow transmission of politically sensitive information among citizens and thus should be more costly for autocrats than false positives (censoring more than they should).

Figure 1.4 breaks down the models' errors by type. It reports the false positive rate, defined as $\frac{\text{No. of false positives}}{\text{Total no. of negatives}}$, and false negative rate, defined as $\frac{\text{No. of false negatives}}{\text{Total no. of positives}}$, for different censorship AI models. As Figure 1.4 shows, the false positive rate is low and stays relatively stable across different thresholds. However, as data missingness increases, the false

negative rate increases drastically, with the largest false negative rate more than seven times that of the smallest. This is true for both the status quo test data and the crisis test data, with a larger increase in false negative rate during crisis. Therefore, Figure 1.4 points to a particularly bad situation for autocrats as censorship AI models are more likely to not censor truly censorable content when data missingness increases.

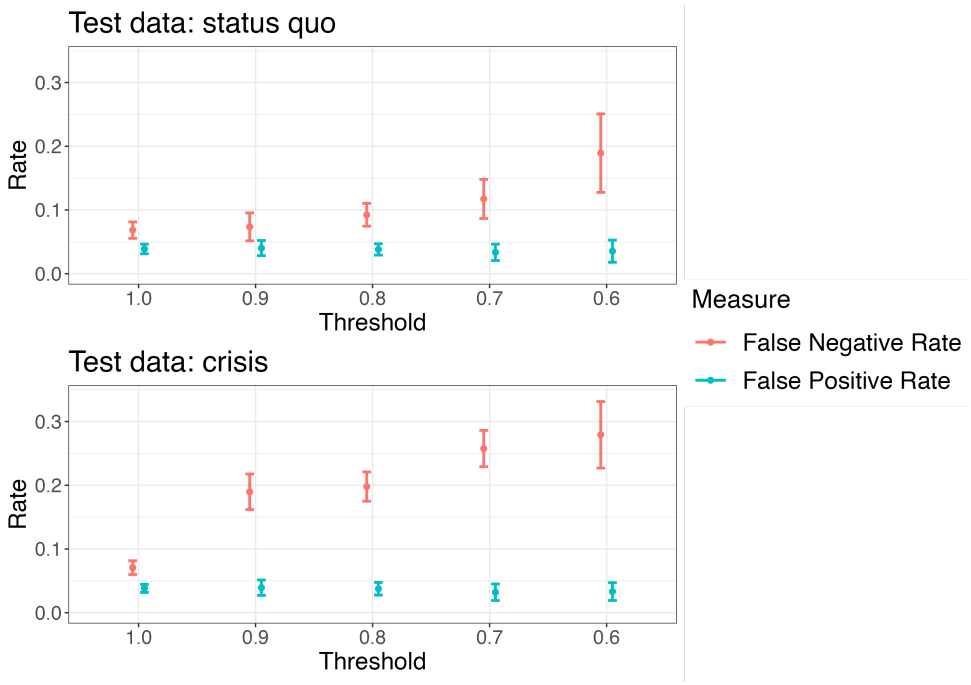

Figure 1.4. False Positive Rate vs. False Negative Rate

Together, Figure 1.3 and Figure 1.4 provide evidence for the repression-performance trade-off and show that the drop in model accuracy is larger during crisis and concentrated on errors of false negatives. In the Appendix, I provide evidence that the substantive conclusions are robust to various changes to the experimental set-up, such as using a larger pre-trained model, changing the censorship decision rule (e.g., from 0.5 to 0.4), allowing some leakage of the missing data into the training data, and using a different deep learning model architecture.

1.4.2 Marginal Impact of More Training Data

Given the previous results, one of the ways autocrats may choose to respond to the problems is to collect more data and train the model on a larger dataset. Figure 1.5 presents the result of doubling the size of the training dataset on model accuracy. Specifically, Figure 1.5 shows the difference in accuracy, on both test data, between models trained with the original training datasets and those trained with double the amount of training data. The difference in model accuracy as a result of larger training datasets is less than one percentage point and not statistically significant for all thresholds, except for the lowest threshold (0.6) with a marginal accuracy improvement of 0.87 percentage point for the status quo test data. Figure 1.5 thus provides evidence that additional data that is collected under the same informational environment where there is preference falsification and self-censorship has a marginal impact on the performance of censorship AI.

In the Appendix, I show that the breakdown of the errors by models trained with the larger training datasets follows a similar trend to the original models - the false positive rate is low and stays relatively stable across different thresholds but the false negative rate increases drastically as data missingness increases.

1.4.3 Accuracy Improvement from Democratic Data Augmentation

While additional data collected domestically provides little improvement to model accuracy, data from international sources that is generated without the same political constraints can help boost model performance. Figure 1.6 provides evidence that augmenting the original Weibo training dataset with data from Twitter improves model accuracy during crisis. Specifically, Figure 1.6 compares the accuracy of models trained on the original Weibo training datasets with models trained on datasets augmented by the Twitter data.

The augmentation provides no improvement during the status quo. This is because the status quo test data is sampled from the same distribution as the original training data. In this

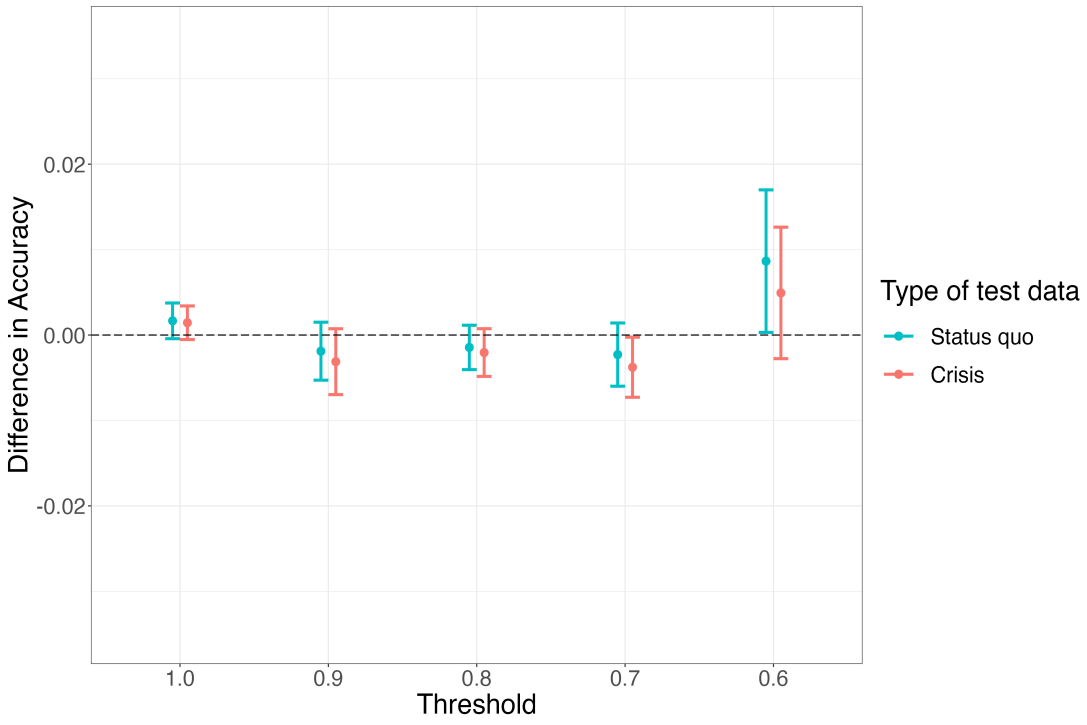

Figure 1.5. Performance Difference from Doubling Training Data

Note: Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

case, the decrease in performance for both sets of models (original and Twitter-augmented) is due to the increase in similarity between censorable and non-censorable content rather than a mismatch in distribution between the training and test data. As augmentation does not change the fact that the prediction problem for the status quo data becomes more difficult as the threshold decreases, the Twitter data thus provides no accuracy improvement during the status quo.

On the other hand, Twitter data augmentation improves the accuracy of censorship AI models that suffer from missing data during crisis. The right panel of Figure 1.6 shows that, when there is missing data (thresholds $0.9 - 0.6$), the accuracy of the models trained on the augmented datasets is higher than the models trained on the original Weibo data. This shows that the Twitter data can partially compensate for the missing data from Weibo and reduces the mismatch in distribution between the training data and crisis test data.

It is important to note that the accuracy improvement from Twitter data is limited, in

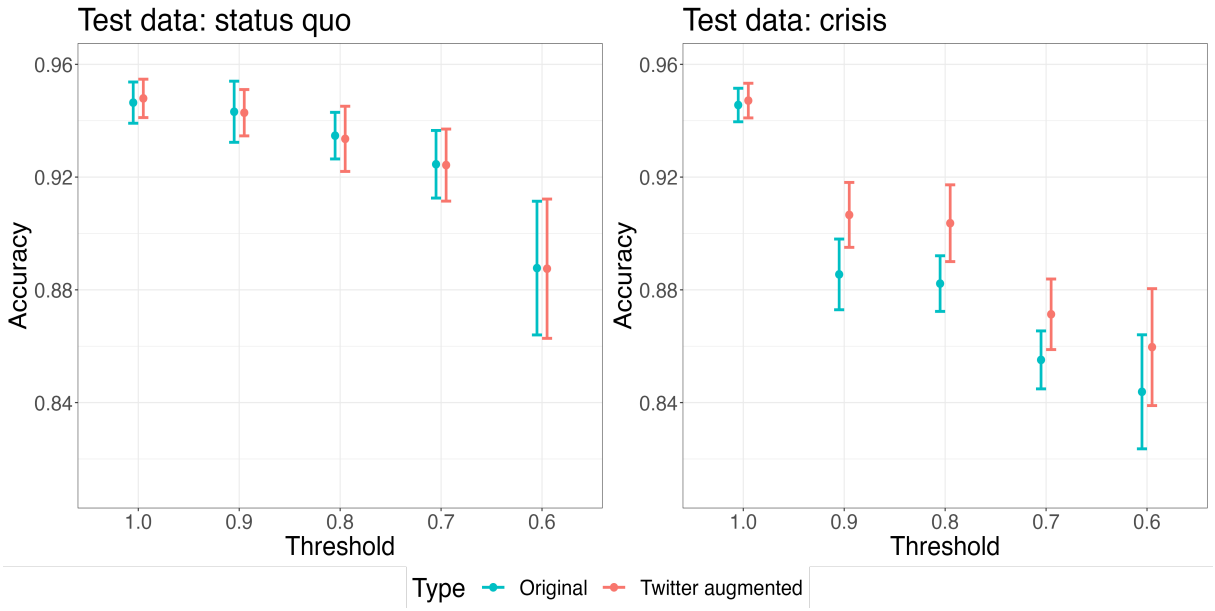

Figure 1.6. Effect of Twitter Data Augmentation

Note: Each threshold value represents a version of the original Weibo training dataset. The same Twitter data, without altering, is used to augment all versions of the original training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

that the models' accuracy is still lower than that of the models trained on the full Weibo data (threshold=1.0). One potential explanation for this is that the content on Twitter is different from the content on Weibo so that relying on Twitter data augmentation cannot fully compensate for the missing Weibo data.

Figure 1.7 provides suggestive evidence that the content in the Weibo data is indeed different from the content in the Twitter data. It shows how a censorship AI model trained on the Twitter-augmented dataset internally represents the Weibo and Twitter data. Specifically, I choose the censorship AI model that is trained on the entire Weibo and Twitter data (threshold $= 1.0$), so that I can obtain the internal representation of all training data. For presentational purposes, I randomly sample 5000 censorable social media posts each from the Weibo and Twitter data. I then use the model to obtain the embeddings (internal representation) of the combined 10,000 social media posts. As the embeddings are high-dimensional, I use principal component analysis (PCA) to reduce the dimensionality of the embeddings and plot the distributions of the first four

principal components in Figure 1.7. Essentially, PCA is a statistical technique that transforms the embeddings into a new set of uncorrelated variables called principal components, which are obtained by maximizing their ability to explain the variations in the original data.

As Figure 1.7 shows, the distributions of the model's internal representation of the Weibo data in the four principal components are quite different from the distributions of the Twitter data. In particular, the spread of the Weibo data is much wider whereas the Twitter data is more concentrated distributionally. T-tests show that the difference in distribution is statistically significant in all four principal components. Additionally, the distribution of political sensitivity is also different for Twitter and Weibo data, with Twitter data being more sensitive (p-value < 0.001) than Weibo data. The differences in distributions of the two datasets provide suggestive evidence of the limited ability of the Twitter data to compensate for the missing Weibo data.

Together, Figure 1.6 and Figure 1.7 provide evidence for the theory's data augmentation hypotheses: more data from domestic sources has a marginal impact on model accuracy but data from international sources helps improve model performance. Additionally, Figure 1.6 and Figure 1.7 show the limits to which such augmentation techniques can boost censorship AI's performance.

1.5 Discussion

Artificial Intelligence has become a key technology in the autocrats' toolkit and will be increasingly so in the foreseeable future. Its ability to ingest vast amounts of information and make predictions based on such information no doubt enables contemporary autocrats to sieve through information at a scale autocrats from other times could not have imagined. However, in this paper, I argue that there are inherent limits to the ability of AI to automate authoritarian control and that such limits are the result of existing authoritarian institutions. Just like their traditional counterparts, digital dictators face a dilemma between repression and information:

the more repression there is, the less political information there will be in the data, and the worse AI will perform. Regardless of how capable AI is, it cannot process nor aggregate information that is not observed.

The theory and the empirical findings of this paper provide some nuance to the ongoing debate on the effect of AI on authoritarian control. By problematizing the argument that more data means better prediction and better control and bringing to the forefront the issue of data quality, this paper argues that the general equilibrium effect of AI may not be as favorable toward autocrats as the existing literature has argued.

The theory of the paper relies on the assumption that in the face of increasing repression and censorship, people will falsify their preferences and self-censor more, causing greater data missingness. This is not a completely innocuous assumption. Although there is substantial empirical evidence supporting this assumption (Fu, Chan and Chau, 2013; Huang, 2015; Tanash et al., 2017) and it is in fact the premise of the dictator's dilemma in Wintrobe (2000), a few studies have shown that repression can generate both chilling and backlash effects (Huang, 2018; Pan and Siegel, 2020).[14] The scope conditions for the backlash effect identified in the literature are that repression and censorship are overt and visible to the public and that they are not strong enough to stifle most citizens' reactions (Pan and Siegel, 2020; Roberts, 2020). In the context of digital repression and censorship, which are more covert and all-encompassing by nature (Xu, 2021), the scope conditions may be too stringent and the backlash effect can be limited as a result. On the other hand, if there is indeed a substantial backlash effect, by the logic of the theory, this can have an unintended consequence of providing valuable information to the training data and boosting repressive AI's performance.

The theory also points to similar unintended consequences of political phenomena that work in the digital dictators' favor. One, polarization in authoritarian regimes can make the prediction problem easier. This is because, as the online discussion polarizes, the censorable content will be easier to identify by AI as their similarity with non-censorable content decreases.

[14] See Roberts (2020) for a survey of the debate.

This serves as an additional channel, on top of the ones the existing literature has identified (Svolik, 2018, 2019), through which (would-be) autocrats can use polarization to strengthen their rule.

Similarly, the theory suggests that if there are alternative, non-domestic platforms on which citizens can express dissent, then the repression-performance trade-off may be partially mitigated when autocrats also collect data from these platforms. Several recent studies have documented the migration of dissent from domestic to international platforms (Hobbs and Roberts, 2018; Esberg and Siegel, 2021; Esberg, 2022). In the context of AI, this can work in the autocrats' favor, as this allows them to collect uncensored information without changing the repressive environment domestically. However, as the paper demonstrates, this "irony of the free world" effect may be limited in its impact on AI performance, especially when discussions from international sources diverge from domestic sources.

While not explicitly spelled out, the paper points to the possibility that data from democracies boosting authoritarian AI is only half the story. By the same logic, data from authoritarian regimes can serve to contaminate AI from democracies. Given that major AI companies in the U.S. and Europe are relying on ever larger datasets to train their AI models, it is likely that data tainted by censorship and propaganda can influence the output of these models (Yang and Roberts, 2021). This can be especially concerning considering that such AI models are being deployed in important areas such as education and criminal justice. Documenting data leakages from authoritarian regimes and quantifying their effect on AI should be a focus of future research.

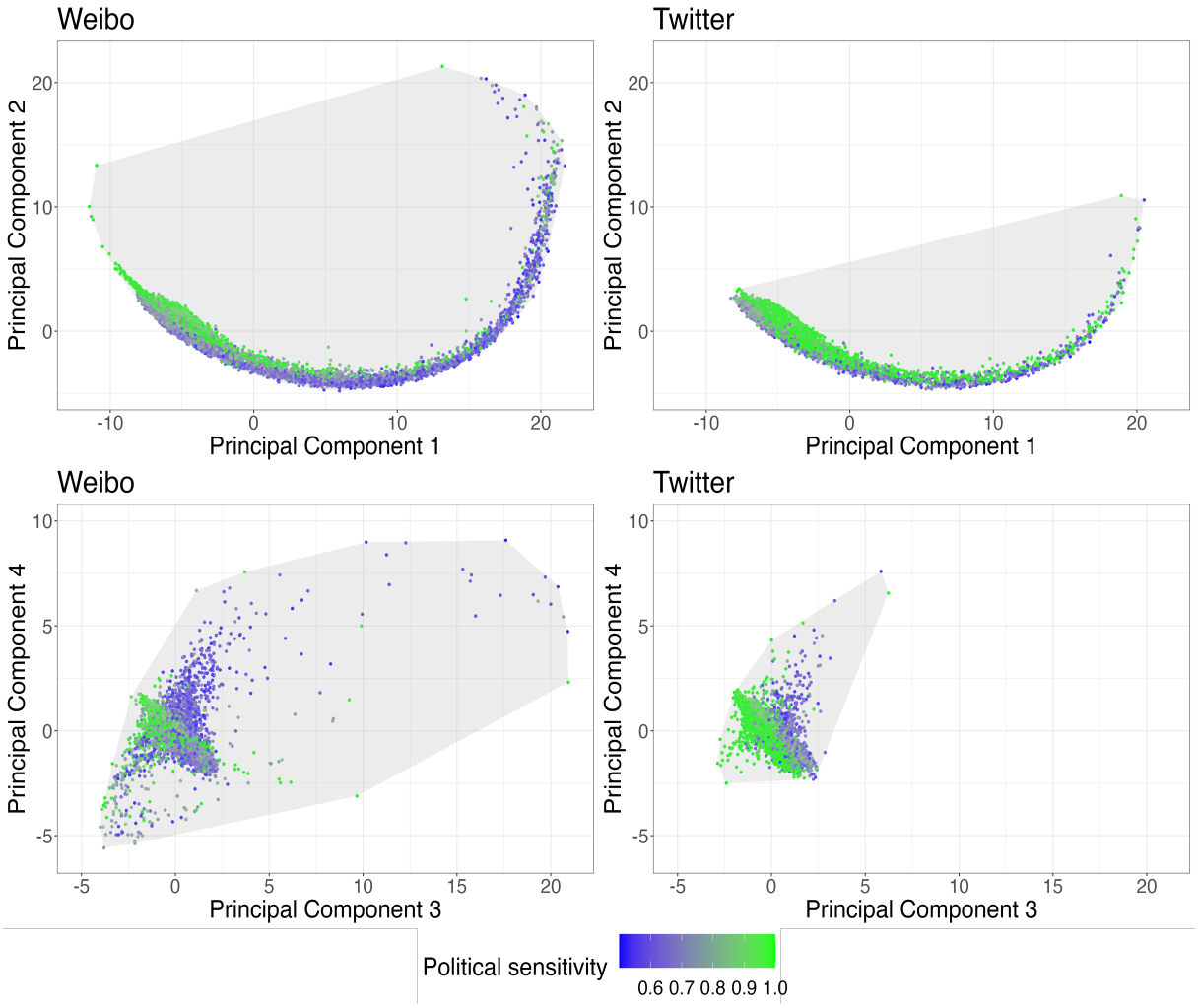

Figure 1.7. Principal Component Analysis of Content on Weibo and Twitter

Note: Each point represents a social media post and the color indicates its political sensitivity. Shaded regions represent the convex hulls of the points.

Chapter 2

Propaganda and AI

Abstract

Social scientists have been critical of the potential for Artificial Intelligence (AI) systems to further entrench the human biases that these tools reflect. We examine a distinct and pernicious bias in generative AI systems: *propaganda bias*, wherein coordinated rhetoric from political institutions influences the training data and output of large language models. Leveraging two unique datasets of Chinese propaganda news articles, we provide evidence that propaganda does indeed affect large language models (LLMs). We find propaganda in the Chinese-language portions of open-source training datasets, particularly for political topics. We then present a series of audit studies showing that (1) widely used LLMs have memorized commonly employed propaganda phrases, and (2) Chinese-language prompts consistently generate more positive responses for Chinese political institutions and figures than the same queries in English. Finally, we conduct pre-training experiments on LLMs and show that adding propaganda to the training data can cause LLMs to generate responses that are more favorable to the Chinese government.

Trained on massive amounts of data, Artificial Intelligence (AI) systems can generate coherent and realistic content, including texts, images, videos, and more. However, they can also replicate and aggravate the social biases embedded in the training data (Kay, Matuszek and Munson, 2015; Noble, 2018; Broussard, 2023; Benjamin, 2019; Buolamwini and Gebru, 2018). For example, previous work has shown how AI systems can reproduce gender and racial stereotypes frequently expressed online (Barocas and Selbst, 2016; Sheng et al., 2019; Field

et al., 2021; Metaxa et al., 2021; Kotek, Dockum and Sun, 2023; Omiye et al., 2023). However, researchers have given less attention to how institutional strategies can influence training data and therefore induce biases in AI systems. We emphasize that state-sponsored propaganda (Guriev and Treisman, 2019; Carter and Carter, 2023; Guriev and Treisman, 2022; Sanovich, Stukal and Tucker, 2018; Mattingly and Yao, 2022), misinformation campaigns (Aral and Eckles, 2019; Grinberg et al., 2019; Lazer et al., 2018; Vosoughi, Roy and Aral, 2018; Tucker et al., 2018), and the coordinated censorship of information (Roberts, 2018; King, Pan and Roberts, 2013, 2014; Bamman, O'Connor and Smith, 2012) can all distort the composition of available content online, with potential downstream effects on AI systems.

We provide worrying evidence that political institutions affect the training data and output of AI systems. Focusing on large language models (LLMs) and open-source training data, we show that large quantities of text manipulated by political institutions are found in commonly used open-source training datasets, and appear to induce biases in state-of-the-art, widely used LLMs. We call this *propaganda bias*: The influence of strategic and coordinated rhetoric from political institutions on the training data and the output of LLMs.

One of the core challenges to measuring this type of bias is model opacity. Information about the training data and model weights for many of the most commonly used LLMs is not public. Our work collects several pieces of evidence that jointly demonstrate that propaganda is currently creating a non-trivial political bias in LLMs' output.

We start with observational evidence that content from news articles manipulated by the Chinese government appear with substantial frequency in common open-source multilingual training datasets. We audit CulturaX, a massive training dataset of multilingual texts that combines two other popular datasets, multilingual C4 (mC4) and OSCAR. We identify shared portions of exact text between the Chinese-language portion of CulturaX and two datasets of Chinese propaganda news articles. We find evidence of propaganda in CulturaX and for both subsets (mC4 and OSCAR): up to 1.17% of the Chinese language portion of CulturaX includes text snippets from the propaganda corpora. This percentage increases up to 21% for Chinese

language CulturaX documents discussing political topics, such as Chinese political leaders and institutions.

Second, we show that widely used LLMs have memorized common propaganda phrases–evidence of propaganda in the actual training data for production models. We prompt LLMs to complete partial text snippets from the propaganda documents and find LLMs reproduce substantial parts of common propaganda phrases with minimal deviations in the output.

We then use a series of audit and pre-training experiments to demonstrate the possible downstream effect of propaganda on the output of LLMs. We first show that popular commercial LLMs generate responses that are substantially more favorable toward China when the prompt is in Chinese than in English. We prompt ChatGPT and Claude 3 with a series of political questions about six countries - the United States, China, Russia, Germany, the United Kingdom, and North Korea. We find that both ChatGPT and Claude 3 exhibit a greater pro-China slant when prompted in Chinese about Chinese institutions and leaders than when prompted in English. We do *not* find language-based differences for prompts on non-China subjects.

To provide causal evidence that propaganda in the training data affects model output, we conduct a series of pre-training experiments. The introduction of a relatively small number of documents from the Chinese propaganda apparatus is sufficient to induce LLMs to generate responses that are more favorable to the Chinese government. Because this pattern is only substantially visible in Chinese, our pre-training experiments furthermore demonstrate the mechanism connecting training data poisoning (Shayegani et al., 2023) with Chinese propaganda and pro-China slant in LLM outputs in Chinese language.

Taken together, our results highlight the way political institutions can affect the production of training data for AI models and its downstream consequences on model output. The results have worrying implications, suggesting that the ability to influence the output of LLMs through propaganda may create strategic incentives for governments and other political actors to strengthen their propaganda apparatus and the spread of manipulated information online.

2.1 Government Manipulated Texts and Propaganda Bias

Governments have long used propaganda as a coordinated strategy to disseminate particular ideas and rhetoric, often in a deliberate attempt to control the information available to citizens and sway public opinions (Jowett and O'Donnell, 2018). For instance, by manipulating news and mass producing political texts and speeches, governments have used propaganda to influence domestic and foreign elections (Peisakhin and Rozenas, 2018; Selb and Munzert, 2018), boost the public image of government competence (Rozenas and Stukal, 2019), signal the strength of the government (Huang, 2015), and even stoke hatred toward marginalized ethnic groups (Voigtländer and Voth, 2015).

Modern government-organized campaigns can leave huge footprints on information available on the internet. For example, Waight et al. (n.d.) estimate that at least 5-7% of newspaper articles in China are directly scripted by the Propaganda Department, while King, Pan and Roberts (2017) estimates that hundreds of millions of online posts per year have been written at the direction of the Chinese government.

The fact that LLMs require (pre-)training on massive amounts of text - often scraped from the internet - means that propaganda is likely to get swept up when collecting training data. It is this "semi-supervised" learning of token sequences (Dai and Le, 2015; Kiros et al., 2015) - pretraining - that both necessitates the use of massive training corpora and underpins the success of modern LLMs. Furthermore, the reliance on massive amounts of text data creates incentives for those training LLMs to use freely available training data source like non-paywalled digital media content (Radford et al., 2019), and those creating propaganda tend to make their content as open and accessible as possible.

Thus our expectations from the outset were that the presence of propaganda in the training data would - perhaps unintentionally - skew the output of LLMs. A similar phenomenon has been observed when comparing word embeddings trained on Chinese Wikipedia with those trained on Baidu Baike, a Chinese online encyclopedia that is subject to propaganda and censorship.

Yang and Roberts (2021) show that, compare to the Chineses Wikipedia embeddings, Baidu Baike embeddings represent political concepts like democracy and election closer to word with negative connotations, such as chaos and poverty, while concepts about social control and Chinese political leaders are represented closer to positive words like prosperity and stability. A similar effect on widely used instruction-following LLMs would be particularly concerning, given their ability to launder manipulated content and opinions into seemingly objective texts (Benjamin, 2019).

We call this hypothesized effect *propaganda bias*. More precisely, we define propaganda bias as the influence of strategic and coordinated rhetoric from political institutions on the training data and output of large language models. In contrast to social biases in LLMs, such as gender and racial biases, that have garnered considerable attention, propaganda bias, and more generally the effect of political institutions on LLMs have received far less attention.

Measuring bias, whether social or institutional, implicitly relies on a normative counterfactual: what an unbiased system would look like (Johnson and Zhang, 2022). For example, Wan et al. (2023) attempt to estimate gender bias using a counterfactual of equal treatment. The authors prompt a large language model with male and female names and ask the model to write a reference letter. They find that the model systematically writes different letters for women and men, both with respect to lexical content and style, enabling them to reject the counterfactual of equal treatment.

Unlike the previous example, articulating what the output of an unbiased system would look like without propaganda is more difficult. For example, if a large language model is more likely to label an autocrat as a good leader, it could be because the model's training data includes state-manipulated content (i.e. propaganda bias), but it could also be because the training data includes pro-government authentic expression by non-state actors or because of human feedback in the reinforcement learning process of training LLMs (RLHF). Propaganda bias implies a particular mechanism through which a certain output is produced: that propaganda documents in training data systematically affect the model output.

2.2 Propaganda in China as a Case

To empirically measure propaganda bias, we connect instances of known state media manipulation with LLM training data. In particular, we focus on two sources of known Chinese state propaganda: government-authored propaganda scripts reprinted in Chinese newspapers and digital news articles promoted by the Chinese state propaganda apparatus on their digital news application - Xuexi Qiangguo.

One key mechanism by which the Chinese state intervenes in the media environment is a coordination mechanism that Waight et al. (n.d.) call *scripted propaganda* - that is, when newspapers are required to reprint government-authored script scripts rather than pursue their own reporting. The Chinese Communist Party Central Publicity Department and other government and party organs instruct newspapers on a daily basis to use government-authored scripts when reporting on a range of news stories, from meetings of top party leadership, to natural disasters and sudden public interest stories. Waight et al. (n.d.) demonstrate that when such directives are issued, they result in an observable burst of print newspapers reprinting the same newspaper articles on the same day. Waight et al. (n.d.) reverse engineer this process to identify newspaper articles reprinting government authored scripts at scale. Their measure captures these bursts of newspapers all printing the same text on the same day. We will make use of their dataset of 530,694 predicted propaganda articles as a source of state-manipulated media.

While legacy media remains an important mechanism for propaganda dissemination, the propaganda apparatus has adapted to the proliferation of digital media platforms and technologies with new strategies and methods to disseminate its propaganda content (Lu and Pan, 2021; Repnikova and Fang, 2019; Creemers, 2017; Han, 2018). One new platform is Xuexi Qiangguo, an app developed by Alibaba and the Publicity Department of the Chinese Communist Party. Xuexi Qiangguo was developed to teach users Xi Jinping thought and expose users to approved content from official sources, including state and party offices and state controlled media (Liang, Chen and Zhao, 2021). We draw on 198,872 news articles disseminated on Xuexi Qiangguo as a

second source of Chinese state propaganda content in our analysis.

We emphasize that there are other mechanisms by which the Chinese state intervenes in the media environment (Brady, 2009; Stockmann, 2013; Shambaugh, 2017; Qin, Strömberg and Wu, 2018; Pan, Shao and Xu, 2022), including standing orders, journalist training and socialization, and state efforts to promote propaganda messages on commercially owned digital platforms (Lu and Pan, 2021; King, Pan and Roberts, 2017; Repnikova and Fang, 2019; Esarey, 2015). Our results demonstrating the effect of these particular mechanisms on LLM bias are likely a lower bound on the overall magnitude of LLM propaganda bias attributable to the Chinese state.

2.3 Propaganda in Open-Source Training Data

Using our two datasets of state-manipulated media, we first investigate if propaganda is present in open-source multilingual training datasets for large language models. For open-source training data, we use CulturaX (Nguyen et al., 2023). CulturaX is a cleaned and de-duplicated multilingual dataset that combines two other popular multilingual datasets - mC4 (Raffel et al., 2019) and OSCAR (Ortiz Su'arez, Sagot and Romary, 2019; Ortiz Su'arez, Romary and Sagot, 2020; Abadji et al., 2021, 2022).[1] CulturaX contains 6.3 trillion tokens in 167 languages. All of the OSCAR and MC4 subsets of CulturaX are smaller subsets of the Common Crawl,[2] a massive dataset of daily web crawls across the Internet. C4, OSCAR, and more broadly Common Crawl based-corpora are common training data source for language and other machine learning models (Raffel et al., 2023; Scheible et al., 2020; Shalumov and Haskey, 2023; Serrano et al., 2022; Shliazhko et al., 2022; Mandal and Mahto, 2022).

For our purposes, we focus on the Chinese-language portion of CulturaX, which contains more than 200 million documents with more than 227 billion tokens. Using methods from

[1] CulturaX merges version 3.1.0 of mC4 with the following distributions of OSCAR: 20.19, 21.09, 22.01, and 23.01 (Nguyen et al., 2023).

[2] https://commoncrawl.org/

text reuse (Boumans et al., 2018; Cagé, Hervé and Viaud, 2020; Nicholls, 2019), we identify CulturaX documents which share many long sequences of text with the propaganda documents. We do not require CulturaX documents to exactly copy the propaganda documents because many Common Crawl documents are composites of content on web pages and we would thus expect wholesale reprinting to be rare. To measure these patterns of textual overlap between documents in CulturaX and those in the propaganda datasets, we use 5-word gram cosine similarity.[3] Intuitively, a high cosine similarity indicates that two documents use similar sequences of five words. We consider a document in CulturaX to be matched to a document in the propaganda datasets if the two documents have a cosine similarity higher than 0.2. In a validation exercise included in the Appendix we show that human coders coded at least 84% of CulturaX-propaganda pairs with greater than 0.2 5-word gram cosine similarity as engaging in text copying rather than independent writing. We provide further details on this validation exercise and these methods in the Appendix.

Overall, 1.17% (or 2,211,890) of documents from the Chinese-language portion of CulturaX are matched to the scripted propaganda news articles and 0.79% (or 1,492,593) of Chinese CulturaX documents are matched to the news articles from Xuexi Qiangguo. Furthermore, we find propaganda in both the mC4 and OSCAR subsets of CulturaX.[4] Table 2.1 breaks the estimates down by source for scripted propaganda articles (left) and Xuexi Qiangguo articles (right). In Figure 2.1, we condition the matching rate by whether the Chinese CulturaX document includes one of a series of political keywords. At the top of the plot we include keywords related to Chinese political institution (Central Committee Plenum, Party Congress, Chinese Communist Party, National People's Congress, the Ministry of Foreign Affairs, and the economy). At the bottom of the plot we include leader name keywords (Xi Jinping, Deng Xiaoping, Mao Zedong).[5]

[3]We use the **pkuseg** package (Luo et al., 2019) to segment Chinese documents in CulturaX and the propaganda datasets into Chinese words.

[4]We limit the CulturaX documents to those with at least 10 characters and are written in simplified Chinese. In total, there are 189 million CulturaX documents in our audit. Estimates for mC4, OSCAR (overall), and various distributions of OSCAR are potentially influenced by CulturaX's pre-processing procedure and may not reflect the standalone mC4 and OSCAR datasets.

[5]See Appendix for more details on this keyword matching.

Table 2.1. Percent of Chinese CulturaX Matched to Propaganda

Source	Percent Matched to Scripted Propaganda	Percent Matched to Xuexi Qiangguo
CulturaX	1.17%	0.79%
- mC4	1.55%	0.98%
- OSCAR	0.67%	0.53%
* OSCAR-2019	0.63%	0.13%
* OSCAR-2109	0.93%	0.72%
* OSCAR-2201	0.60%	0.84%
* OSCAR-2301	0.39%	0.49%

Note: Estimates for mC4, OSCAR (overall), and various distributions of OSCAR are potentially influenced by the pre-processing procedure used to construct CulturaX and may not reflect the standalone mC4 and OSCAR datasets.

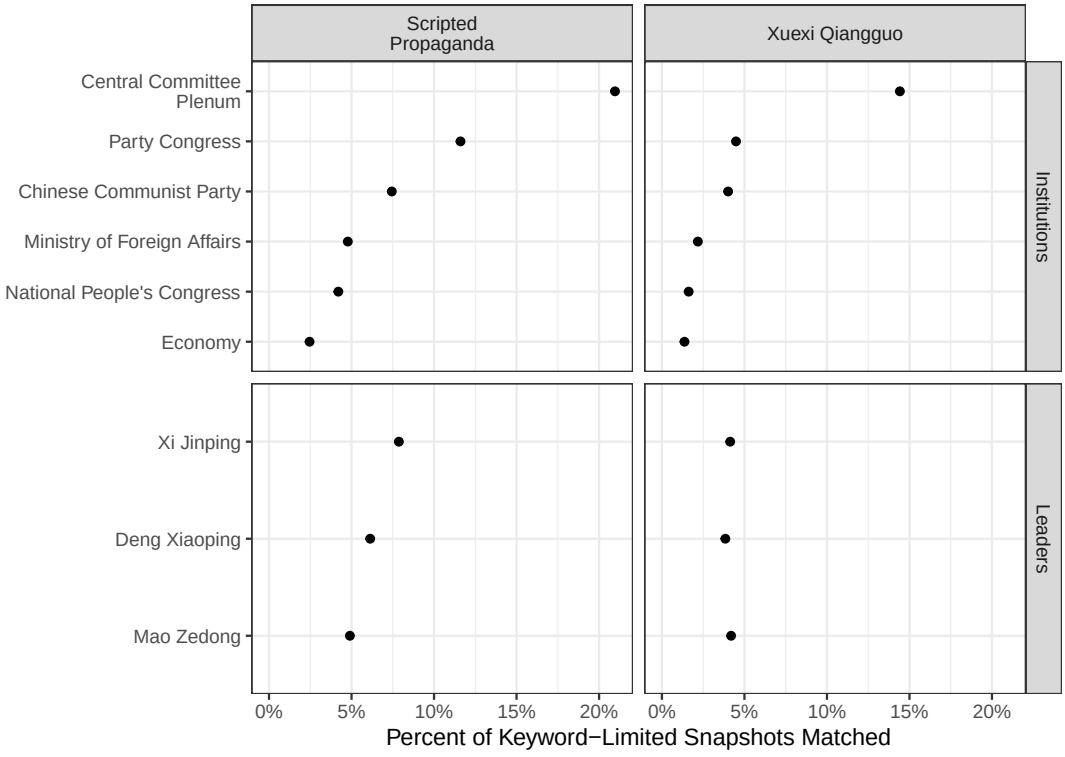

Figure 2.1. Percent of Chinese CulturaX Matched to Propaganda - Political Topics

Note: For both CulturaX and the propaganda datasets, an article is classified as being on a given topic if it includes a keyword of that topic. See Appendix for the keywords used for each topic.

The overall matching estimates are modest, but it is important to contextualize that there are many more Chinese CulturaX documents than propaganda documents in this analysis. Furthermore, the majority of Chinese CulturaX snapshots are unrelated to politics. In the Appendix we demonstrate that the most common web domains in the Chinese CulturaX corpus are online shopping, information technology, blogs, and travel websites. This finding helps explain why, when we condition on political keywords, the overall matching rate increases substantially. Chinese CulturaX documents that include a reference to political institutions, especially the Central Committee Plenums and national Party Congresses had the highest rate of being matched to a propaganda article. 21% of Chinese CulturaX documents that include a reference to central committee plenums are matched to scripted propaganda articles and 14.4% are matched to a Xuexi Qiangguo articles. These estimates reflect patterns in propaganda writ large, as information about political meetings are amongst the most heavily controlled and sensitive in the Chinese media (Waight et al., n.d.; Truex, 2019; Carter and Carter, 2021).

2.4 LLM Audit: Memorization

Although we have shown that propaganda is present in *open-source* training datasets, the *actual* training data for many of the popular large language models are unknown.[6] To more directly demonstrate that Chinese propaganda is present in the training corpora of LLMs and that it has an effect on the output these models, we measure the extent to which commonly used LLMs *memorize* Chinese propaganda word sequences.

Memorization refers to when LLMs can reproduce sequences from their training data (Biderman et al., 2023). For example, using snippets of training data as prompts, Carlini et al. (2022) estimate that the 6 billion parameter GPT-J model memorizes about 1% of its training data and that the rate of memorization increases with the repetition of tokens in the training data

[6]For example, training data for GPT-3, GPT-4, Claude 2, Claude 3, and Llama-2 models are not opne-source. There is even less information about the Chinese language training data for these models.

and the size of the model. In our context, if large language models can be prompted to regurgitate propaganda text sequences, it would provide further evidence that the models have indeed been trained on propaganda data.

To test for propaganda memorization, we prompt GPT-3.5, GPT-4, Claude Sonnet, and Claude Opus to complete common twenty-word phrases from the two propaganda datasets. We first extract 155 most frequent 20-word propaganda phrases across the two propaganda datasets.[7] For each phrase, we prompt the models with the first half of the phrase and, if the model has been fine-tuned for chat, an instruction to complete the phrase. We then measure the extent to which the completions match the actual second halves of the propaganda phrases. We label a phrase as memorized if the model's completion has a normalized edit distance less than 0.4 with the phrase. In the Appendix we validate this threshold with hand labeling and provide more details on the analysis.

Figure 2.2 shows the memorization rates for all four models. The memorization rates range from 7.8% to 19.8%, indicating a substantial number of propaganda phrases are memorized by popular LLMs. Table 2.2 shows two examples of the model completion with the actual phrases and the prompts used. Although we do not know the training data for any of the four models. The memorization test result shows that all four models are able reproduce partial propaganda phrases with minimal deviations.

These results demonstrate that all four models can be prompted to reproduce partial propaganda phrases with minimal deviations. Although we do not know what data was actually used to train these four models, these findings demonstrate that the training data very likely includes Chinese state propaganda. One interpretation of the results of the experiment is that we are not picking up the memorization of propaganda phrases but instead capturing the memorization of common Chinese language text sequences. To test for this possibility, in the

[7]To identify these phrases, we first extracted the 1000 most frequent 20-word grams from each of the two propaganda datasets. We then took the intersection of the two sets of 1000 20-word grams, resulting in 172 unique 20-word grams. We removed 17 word phrases which only include a list of government officials' names, for an analytic sample of 155 propaganda phrases.

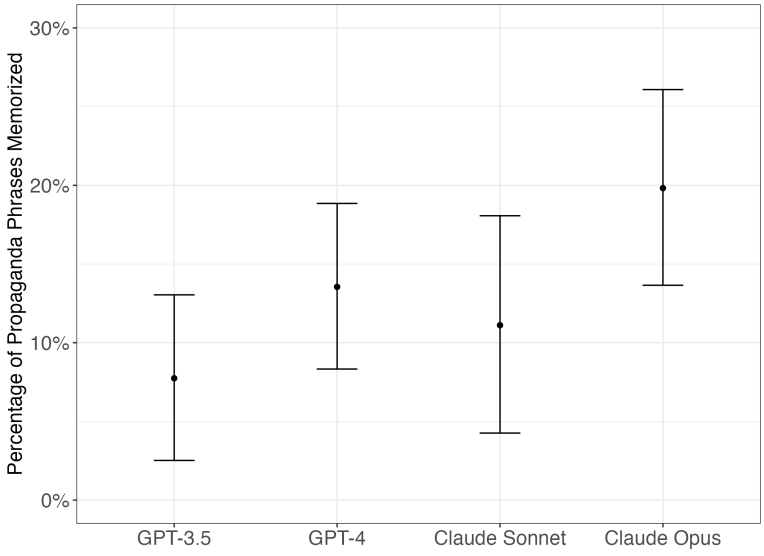

Figure 2.2. Memorization Test Result

Note: Claude Sonnet and Claude Opus refused to answer for 65 and 44 of the phrases respectively. The memorization precentages for Sonnet and Opus are calculated after removing the refusals.

Table 2.2. Memorization Examples

Model	Prompt	Completion	Actual Phrase	Norm. Edit Distance
Opus	中国马克思主义、二十一世纪马克思主义，是中华文化 (Chinese Marxism, Marxism of the 21st century, are Chinese culture)	和中国精神的时代精华，是实现中华民族伟 (and Chinese spirit's essence of the times, realizing the Chinese nation)	和中国精神的时代精华，实现了马克思主义 (and Chinese spirit's essence of the times, realizing Marxism)	0.37
GPT-4	全面建设社会主义现代化国家新征程、向第二 (A new journey to comprehensively build a socialist modernized country, toward the second)	个百年奋斗目标进军的第一个五年 (centenary goal by marching through the first five-year plan)	个百年奋斗目标进军的号角已经吹 (centenary goal by sounding the clarion call)	0.33

Appendix we estimate the memorization rate for text sequences which are highly characteristic of scripted propaganda documents as compared with newspaper articles not predicted to be propaganda. We also replicate this analysis using a random sample of text sequences. Our results jointly demonstrate that the Claude 3 and ChatGPT models have been trained on and learned from Chinese propaganda documents.

2.5 LLM Audit: Political Valence

Having demonstrated the presence of propaganda in training data, we now test whether the political valence of large language models is substantially different when prompted in different languages. If the inclusion of Chinese propaganda in training data has an effect on LLMs, our hypothesis is that it will influence LLMs to respond more positively when prompted in Chinese on topics heavily coordinated by the propaganda apparatus than when prompted in English on those same topics. We use an audit experiment to show that popular LLMs - GPT-3.5, GPT-4, Claude Sonnet, and Claude Opus - indeed show substantially different political valence when responding to Chinese and English prompts on political topics related to China.

We first construct three sets of political questions about political leaders, institutions, and political systems. We then query LLMs with the political questions in both Chinese and English and ask LLMs to generate open-ended responses. Table 2.3 shows the template we used to generate the prompts.

Table 2.3. Example templates used for political statements

Category	Example template
Political leader	Is [LEADER] a dictator/autocrat/good leader/bad leader?
Institution	Is [COUNTRY's] [INSTITUTION] corrupt/trustworthy?
Political system	Is [COUNTRY] a democratic/authoritarian/dictatorial country?

After generating the completions, we use both human evaluation and LLM-as-a-judge to

compare the LLMs' responses in terms of favorability toward the subject of the prompt. To ensure comparability, we translate the LLMs' English and Chinese responses into the other language. Therefore, for each prompt, we generate two pairs of comparison, one in English (English completion and Chinese completion translated into English) and one in Chinese (Chinese completion and English completion translated into Chinese). To evaluate each pair, we ask our human coders and LLM-as-a-judge the following forced choice question, substituting all upper case words with the subject of the prompt: " Which response reflects more positively on INSTITUTION / LEADER / COUNTRY?" Figure 2.3 provides an illustration of the experiment design. This design is analogous to past search engine audit studies that prompt the system with queries that vary along the dimension of interest (e.g., Metaxa et al., 2021).

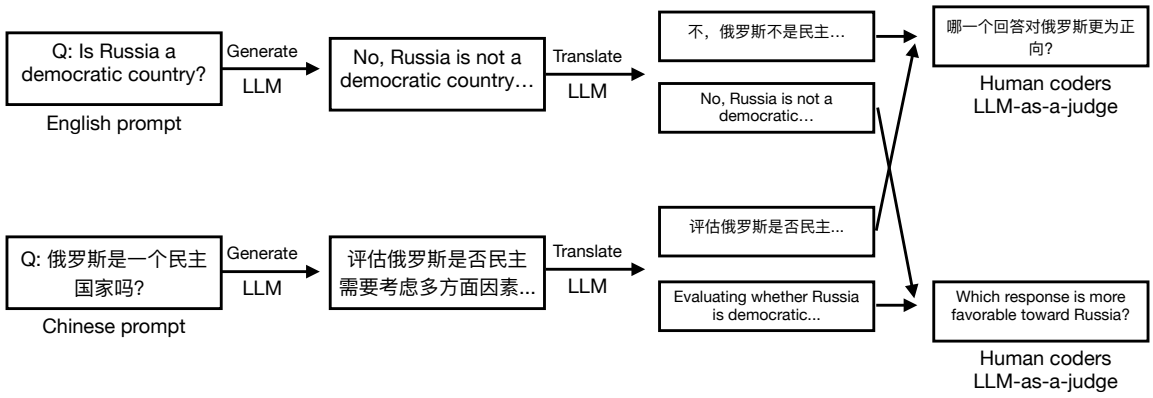

Figure 2.3. Audit Design

For human evaluation,[8] each pair of completions is rated by nine different human coders, who were provided detailed instructions on the rating task. We randomize whether human coders viewed the completion pairs in Chinese or English, thereby blinding our coders to the original language of the completions. Because human evaluation is labor intensive, we employ a limited set of prompts and focus on the responses from a single model (GPT-3.5). 70 of the prompts concern Chinese systems of government, institutions, or leaders, while 191 of the prompts concern non-Chinese systems of government, institutions, or leaders. Our main outcome variable

[8]Pre-registration: https://osf.io/knf67/?view_only=6b948d57e35240dd889a996e591b3073.

is a binary indicator of which response (either the completion from the original Chinese or original English prompt) is more favorable toward the subject of the prompt. We average the ratings across human coders for each comparison pair. In total, we have 261 comparisons.

For LLM-as-a-judge, we broaden the set of political questions and models. Specifically, we construct questions about political leaders, systems, and institutions for six countries: the United States, Germany, the United Kingdom, China, North Korea, and Russia. We then use LLM-as-a-judge to rate the responses to the Chinese and English versions of the questions. Consistent with the human evaluation task, our main outcome variable is which completion is more favorable to the subject of the prompt. For LLM-as-a-judge, we query each comparison ten times and average the ratings. In total, we have 1656 comparisons.[9] In future versions, we will use a method from (Egami et al., 2024) to debias the LLM ratings with a set of gold standard ratings from human coders. In the Appendix we report results on the refusal and equivocation rates in the LLMs' responses for both experiments as well as the full text of all prompts from the two tasks.

Our main outcome variable is a binary indicator of which response (from the Chinese or English prompt) is more favorable toward the country in question. For human evaluation, we average the ratings across human coders for each pair of comparison. For LLM-as-a-judge, we query each comparison ten times and average the ratings. In the appendix, we also report results on the refusal rate and equivocation of the LLMs' responses.

Figure 2.4 shows the human evaluation result for responses about Chinese and non-Chinese institutions, leaders, and political systems. Figure 2.4 shows, for 75.9% of the time, GPT-3.5's responses to Chinese prompts are rated more favorable toward China than its responses to English prompt. For prompts not about China, the favorability rating is close to random guessing. We report the breakdown of the results on institutions, leaders, and political systems and additional results on refusal and equivocation in the Appendix.

[9] In total, there are 828 prompts. Responses are translated into both English and Chinese, resulting in $828 \times 2 = 1656$ pairs of comparison. We cluster standard errors at the prompt-entity level in order to take into account within prompt and entity correlations.

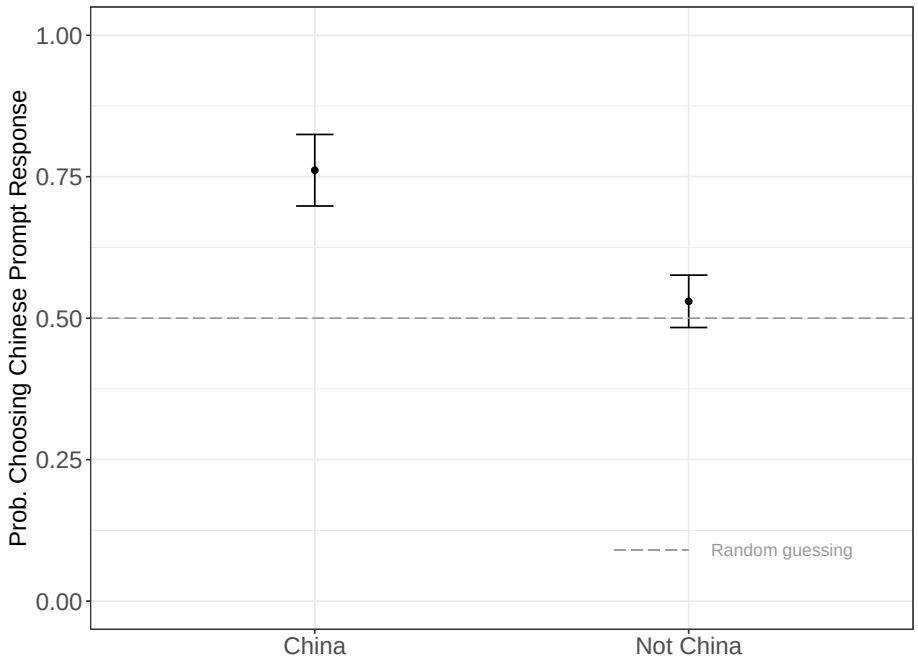

Figure 2.4. Human evaluation result

Note: The human coders were blinded to which response was from the Chinese versus English prompt. For each prompt, we average across human coder evaluations, where 1 indicates the human coder chose the Chinese prompt completion and 0 indicates the human coder chose the English prompt completion. We group the prompts by whether they are about Chinese leaders, institutions, or political system (n = 70), or non-China ones (n = 191). Because each pair of comparison is translated in both Chinese and English, we have $(70 + 191) \times 2 = 522$ comparisons in total.

Figure 2.5 shows the result of using LLM-as-a-judge to evaluate the favorability of the responses across the three groups of political questions from GPT-3.5, GPT-4, Claude Sonnet, and Claude Opus. We use GPT-4 to rate the responses from GPT-3.5 and GPT-4 and Claude Opus for responses from Sonnet and Opus. Figure 2.5 shows that, across all four models, the responses are more favorable toward China when the prompt is in Chinese than in English. This is true for all three groups of political questions. The magnitude of the differences in favorability is significant: responses to Chinese prompts are rated as more favorable between 68% and 88% of the time across models.

We do not observe the same consistent pattern for questions about the United States, the United Kingdom, or Germany. Interestingly, however, we observe a similar pattern for questions

about North Korea and Russia, where responses to Chinese prompts are more likely to be rated as more favorable.

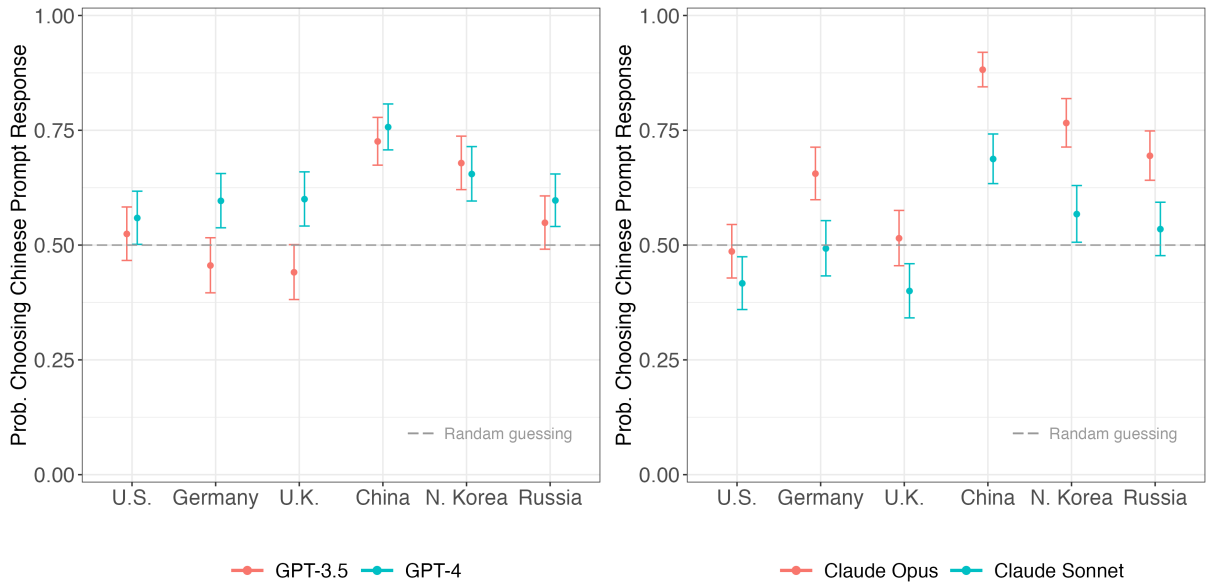

Figure 2.5. LLM-as-a-judge result

2.6 Pre-Training Experiment

To show a causal relationship between propaganda in the training data and model output, the ideal experiment would be to train multiple LLMs from scratch while varying the amount of propaganda that each LLM is trained on. To approximate this ideal experiment, we conduct experiments involving additional pre-training on the Llama-2 13B model. We have three experimental arms corresponding to the following training corpora: 1) scripted propaganda news articles, 2) non-scripted propaganda news articles that are matched to the topic and date distribution of 1), and 3) a random sample from the Chinese-language portion of CulturaX, after the documents matched to propaganda are removed.[10]

Our hypothesis is that continued pre-training on propaganda should skew the model

[10]We provide further details on the corpus construction in the Appendix.

output to be more favorable toward Chinese leaders, institutions, and political systems compared to Llama-2 model without additional pre-training. We expect this effect to be particularly pronounced if we prompt the propaganda model in Chinese rather than English. We compare additional pre-training on the propaganda corpus with pre-training on the other two control corpora in order to disentangle the effect of additional propaganda pre-training from additional Chinese language pre-training.

To see how training data affects model output throughout the training process, we save a model checkpoint every 100 training steps (for a total of 900 training steps), using a batch size of sixty-four training examples for each step. To give the models the ability to chat and answer questions, we fine tune all checkpoints on the same set of instructions (Chen et al., 2023) after pre-training. We then prompt the instruction-fine tuned models with the same set of political prompts we use for the LLM-as-a-judge audit.[11] To gauge the effect of further pre-training on the three corpora, we also apply the same instruction fine-tuning to a base Llama-2 13B model (without further pretraining) and compare the responses from the checkpoints of the three models to the base model. Similar to the LLM-as-a-judge audit, we use GPT-4 to rate the favorability of the responses.

Figure 2.6 shows the results of the pre-training experiment. Specifically, Figure 2.6 shows the proportion of responses from the further pre-trained models, compared to the responses of the base model (0.0k), that GPT-4 rates as more favorable toward the institutions, leaders, or political systems of country in question. The results are further broken down into the overall proportion, when the prompts are in Chinese, and when the prompts are in English.

The pre-training experiment results show that continued pre-training on Chinese propaganda indeed skews the model output to be more favorable toward China. Although all three models show more favorable responses when compared to the base model, the magnitude of the effect is the largest for pre-training on the propaganda news articles—with roughly 80% of

[11]To further reduce the resources required for the experiment, we use LORA (Hu et al., 2021) for both pre-training and fine-tuning, where we update all linear layers with a rank of 8.

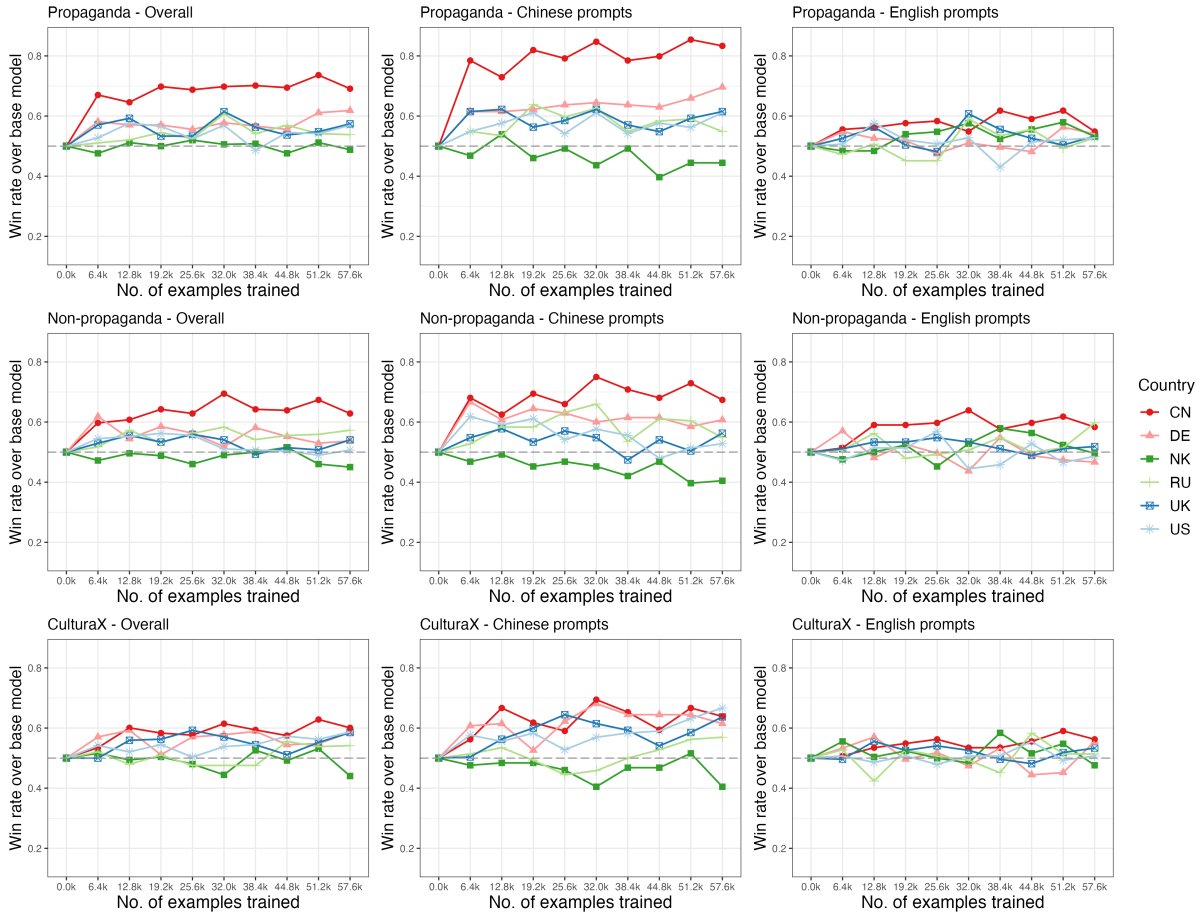

Figure 2.6. Pre-Training Experiment Result

responses rated as more favorable toward China for prompts in Chinese after additional pre-training. Comparing the results for the Chinese and English prompts, it is clear that pre-training on Chinese propaganda primarily affected responses to Chinese prompts, whereas responses to English prompts show little difference either to the base model or among questions about different countries. Additionally, the effect of propaganda is picked up very quickly (at 100 training steps with only 6400 training examples). This suggests that a relatively small amount of propaganda can skew a model's output.

To contextualize our findings, we provide an illustrative example of the different models' responses to the question "Is China an autocracy?" in Table 2.4. The responses show a stark contrast: the base model and the model further pre-trained on CulturaX provide definitive and affirmative responses to the question; the model trained on non-propaganda news articles equivocates; and the model trained on propaganda news articles refutes the claim, citing China's constitutional stipulations protecting people's freedom. We provide more examples in the Appendix.

2.7 Discussion

We document in this paper a phenomenon we call propaganda bias, the effect of strategic and coordinated rhetoric from political institutions on the training data and output of large language models. Using the case of Chinese propaganda, we show that (1) text sequences from Chinese propaganda are found in open-source datasets commonly used to train LLMs, (2) LLMs can be prompted to regurgitate propaganda text sequences, (3) LLMs exhibit a greater pro-China slant when prompted in Chinese about Chinese institutions and leaders than when prompted in English, and (4) adding additional propaganda to the pre-training data of a large language model can induce greater political slant.

Although propaganda is often undestood as a political strategy to influence the information environment, we show that a byproduct of propaganda is that, by influencing online

Table 2.4. Responses to the question: Is China an autocracy?

Model	Response
Base model (0.0k)	China is an autocratic country where the government has extremely high levels of control.
Propaganda model (57.6k)	China is not an autocratic country. China is a socialist country and its political system is a multi-party system with separate party and state organs. China's constitution stipulates that the government must respect and protect citizens' basic rights, including freedom of speech, freedom of religion, freedom of the press, freedom of assembly, and freedom of association.
Non-propaganda model (57.6k)	China is a centralized country, but it also has a certain degree of democracy.
CulturaX model (57.6k)	China is an autocratic country with government power concentrated in the central government. China's government is a unitary system, that is, the ruling party is the Communist Party of China. China's ruling party has strong control over the media, law and society.

content and making its way into the training data, it can influence the output of large language models. This fact has two concerning implications. One, it suggests that LLMs can serve as an intermediary that launders manipulated content into seemingly objective texts. This can further muddy the online information environment by increasing the persuasiveness and subtlety of propaganda. Two, the ability to skew LLM output may further incentivize political actors to expand their effort at manipulating available content. A growing literature suggests that LLMs can be susceptible to data poisoning and adversarial attacks.[12] We highlight propaganda as a channel through which political actors can leverage similar techniques.

Furthermore, existing AI safety efforts—fine-tuning with humans in the loop (often

[12]See e.g., Shayegani et al. (2023) for a survey on this topic.

called Reinforcement Learning with Human Feedback or RLHF), persistent testing, evaluation, and red-teaming—designed to align and moderate the output of LLMs appear not to have yet curbed this tendency. We emphasize that Chinese state propaganda is an important case for LLM development because there are no easy solutions for companies that would want to exclude these kinds of articles from the training data. While these articles are parroting the state's voice, the vast majority of especially the scripted propaganda articles are not being printed in known propaganda outlets, meaning that they cannot be easily filtered out by reference to their source. Our memorization and audit studies show that these measures do not or at least have not yet "fixed" propaganda bias in the most widely used models to date.

Our paper more broadly contributes to understanding the social and political forces shaping training data and AI systems. While recent research has highlighted the need for curated, high quality datasets for large language model training (Budach et al., 2022), these models have largely been trained on open web data. The reliance on massive crawls of web data like the Common Crawl indeed may be a feature, not a bug, of large language model training given these models' reliance on massive amounts of language text. Drawing on previous work in digital communication and digital information flows, we highlight three sets of social, technical, and political forces which shape the production, amplification, and selection of digital texts into these large open web-based training data corpora. First, commercial incentives to increase internet traffic and subscriber revenue shape both the types of content media organizations put online (Christin, 2018) as well as their decisions to paywall or keep that content on the open web. Second, the infrastructure of the web which lubricates the search for and production of information online is dominated by large for-profit companies, whose decisions create inequalities and injustices regarding which content is amplified and which is suppressed (O'neil, 2017; Fourcade and Healy, 2024; Gillespie, 2018; Noble, 2018). Finally, national legal regimes condition the flows of data within and across borders (Kokas, 2023), shaping access and representation.

Authoritarian governments may be particularly well-positioned in this social, political, and technical contestation within open web training corpora. First, control over the media

gives authoritarian governments a (semi-)direct mouthpiece through which to coordinate their messages and crowd out dissenting voices (Waight et al., n.d.). This degree of coordination makes it more likely such messages end up in the dragnet of the Common Crawl and other large scale web scraping enterprises. Second, state-owned media have not faced the same financial constraints which have hollowed out the news media in democratic contexts (Stockmann, 2013; Brady, 2009; Wang and Sparks, 2019). This may help explain why state owned news outlets like *Xinhua* and the *People's Daily* do not paywall their content. Maintaining open content in turns makes it more likely for state controlled media outlets to have their content pulled into web scrapped training data sets.

An open question to which we encourage greater focus in future research is to what extent this authoritarian dividend is particularly true for China. The Chinese government's hardline approach to strict control over the media environment (Waight et al., n.d.) stands out in an era "spin dictators" and other authoritarian governments' more subtle efforts at media manipulation and persuasion (Guriev and Treisman, 2022). When combined with the importance of repeated phrasing in LLM memorization and learning (Carlini et al., 2022), one consequence of strict media control in China is that the Chinese propaganda apparatus may be particularly effective at influencing AI systems. Propaganda regimes like Russia's which place limits over the *semantics* of speech (Yablokov and Chatterje-Doody, 2021) but not its *lexis* may generate less propaganda bias even if such propaganda ends up in LLM training data. We encourage future research on the relationship between propaganda regimes, data regimes, and LLM bias.

Acknowledgement: Chapter 2, in full, is currently being prepared for submission for publication of the material. Waight, Hannah; Eddie Yang; Yuan, Yin; Messing, Solomon; Roberts, Margaret E.; Stewart, Brandon M., and Tucker, Joshua A. The dissertation author is the primary investigator and author of this material.

Chapter 3

Ethnic Discrimination in AI-assisted Criminal Sentencing

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.

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, Harris and Shen-Bayh, 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, Liu and Crabtree, 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).

3.1 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 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

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

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).

3.2 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 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

[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

[5]See Appendix B.1 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.

main text, I focus on the software's predictions on sentence lengths and report results on charge predictions in Appendix B.3.3.

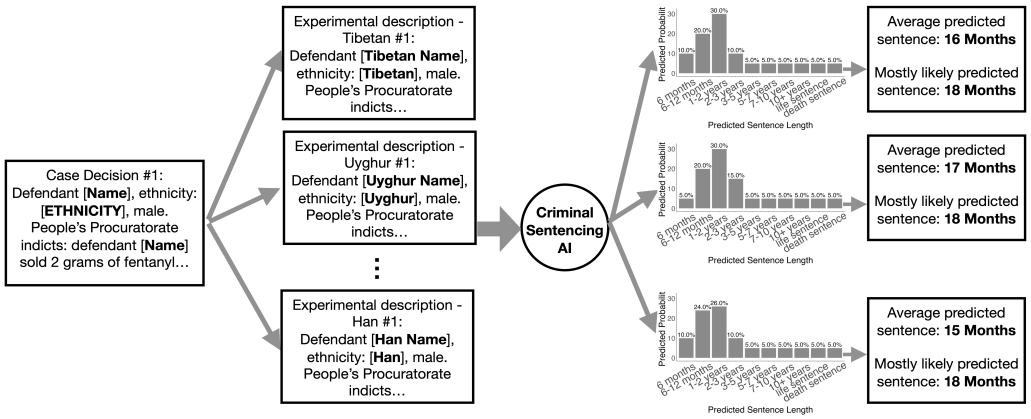

Figure 3.1. Stylized Example of Experimental Setup

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 setup is shown in Figure 3.1.

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

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

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 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:

$$Y_i = \sum_{j=1}^{6} \tau_j M_{ij} + \sum_{k=1}^{1000} \beta_k B_{ik} + \varepsilon_i \tag{3.1}$$

where Y_i is the outcome of interest. In particular, I study two outcomes (as shown in Figure 3.1):

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]

3.3 Results

Drug Crimes Panel A and B in Figure 3.2 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 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 B.3.1, I also show each of the two ethnic cues, namely the defendant's name and ethnicity, contributes to the biases.

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).

[7]In Appendix B.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.

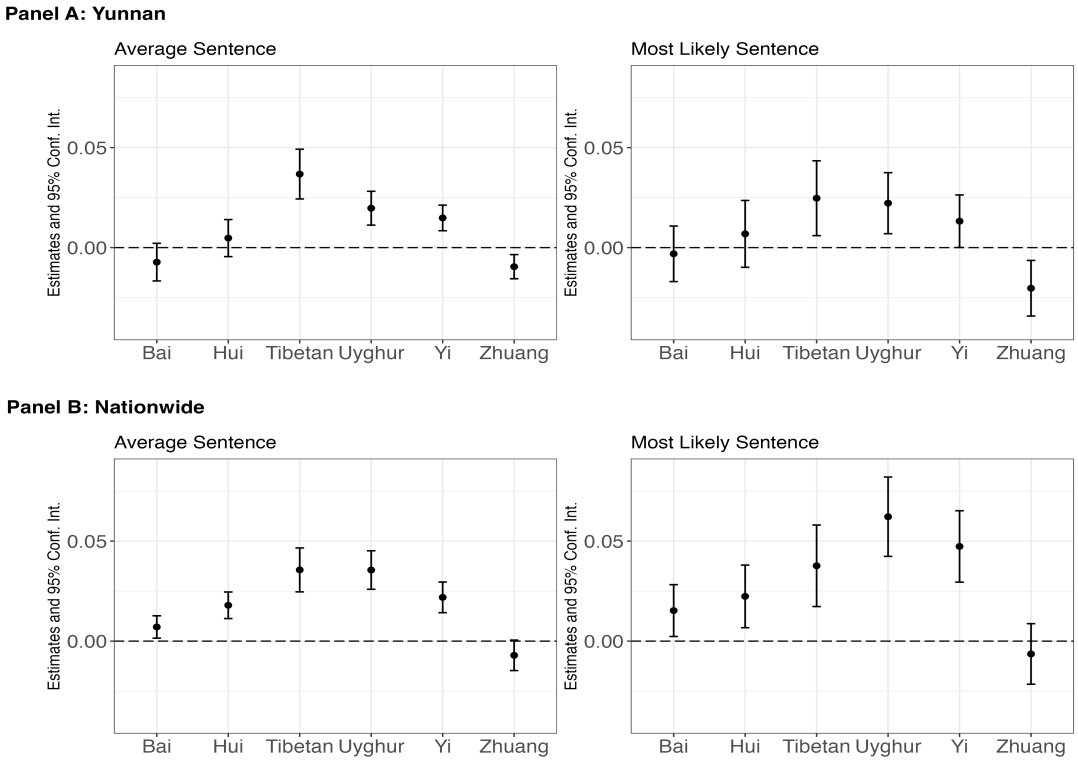

Figure 3.2. Estimates Across Ethnic Groups for Drug Crimes

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 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 3.1 summarizes the findings of the two studies. I discuss potential explanations for such differences in the conclusion.

Table 3.1. 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

Other Crimes Figure 3.3 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.

3.4 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, Hardt and Narayanan, 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 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.

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

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

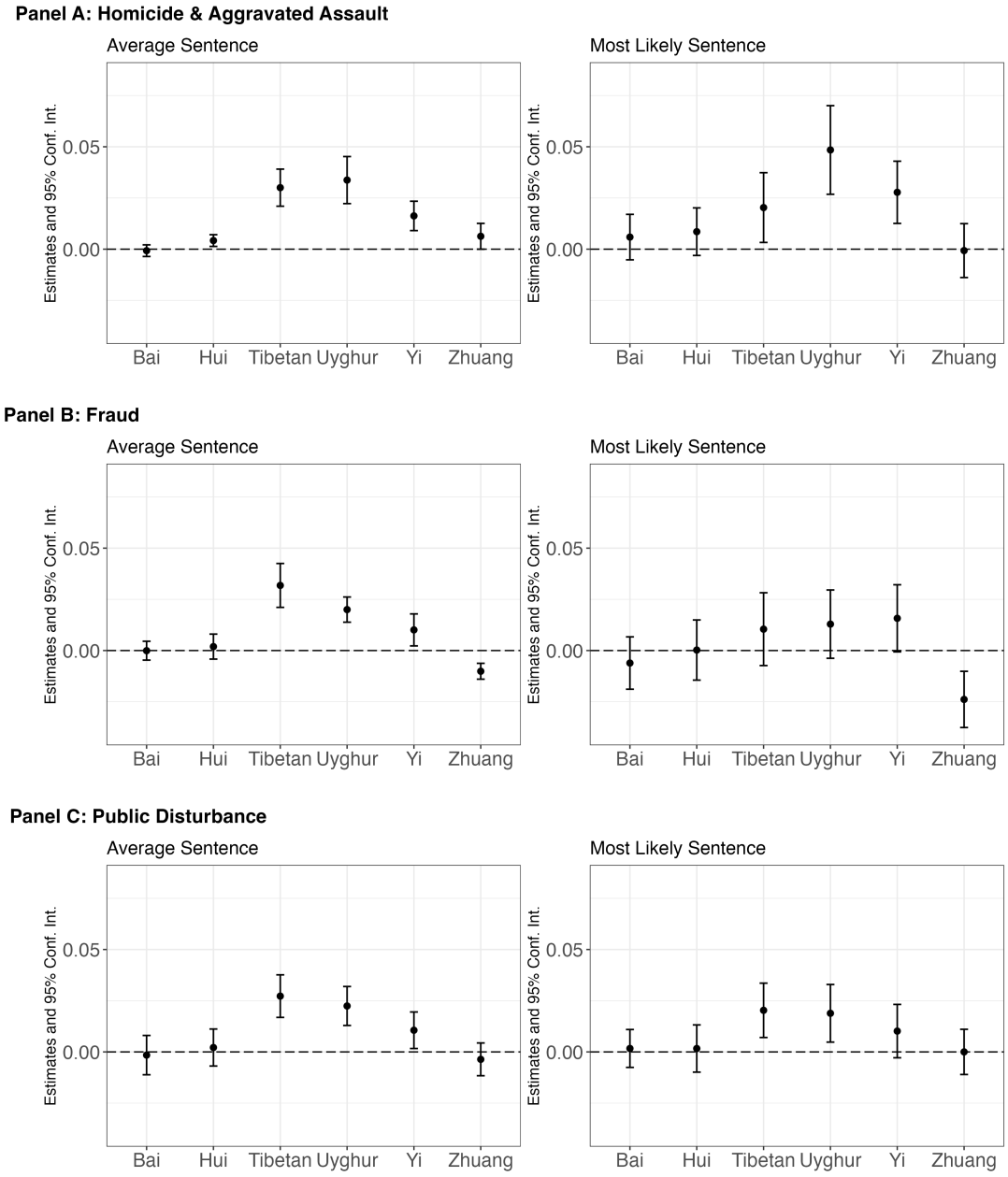

Figure 3.3. Estimates Across Ethnic Groups for Other Crimes

Appendix A

Supplementary Materials for Chapter 1

A.1 Further Details on Censorship AI

A.1.1 Model Details

Except for the result on alternative model architecture in Section A.3.3, the pre-trained BERT model used in the paper is the Chinese BERT with Whole Word Masking trained on extended Chinese text data (RoBERTa-wwm-ext).[1] The model has 102,269,186 trainable parameters and has been shown to perform well on a variety of Chinese prediction tasks (See https://github.com/ymcui/Chinese-BERT-wwm/blob/master/README_EN.md).

Based on information gathered in fieldwork, the model has been used extensively for commercial applications by technology companies. Among other applications, variants of the model have been used to predict censorship and more generally for content moderation (e.g., detecting pornography and spam). In contrast to more recent generative AI models such as GPT, BERT is better suited for prediction tasks and is in general much cheaper and faster in inference/prediction.

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext

A.1.2 Training Details

The BERT models are fine-tuned using the **transformers** library provided by Hugging Face.[2] To fine-tune the models, the social media posts in the training dataset need to be converted into strings of tokens (tokenization) that correspond to the internal dictionary of the BERT model. Tokenization is also provided as part of the **transformers** library.

During training, the F1 score is used as the evaluation metric to track model performance. The F1 score is defined as $F1 = 2 * (\text{precision} * \text{recall})/(\text{precision} + \text{recall})$, where precision is given by $(\text{no. of true positives})/(\text{no. of true positives} + \text{no. of false positives})$ and recall is given by $(\text{no. of true positives})/(\text{no. of true positives} + \text{no. of false negatives})$.

Early stopping was used to prevent over-fitting. Specifically, training was stopped if the F1 score on the validation set did not improve for two epochs. Deterministic training was used via the **enable_full_determinism()** function in the **transformers** library to ensure replicability.

To speed up training, I used mixed precision training with the TensorFloat-32 (TF32) precision format and the Apex fused Adam optimizer.

A.1.3 Hyperparameters

Table A.1 reports the hyperparameter values used in training.

Table A.1. Hyperparameters

Hyperparameter	Value
maximum token length	152
batch size	256
learning rate	4e-5
warmup steps	3000
No. of epochs	8

[2] https://huggingface.co/docs/transformers/index

A.1.4 Hardware

Models in the paper were fine-tuned using 8 NVIDIA A100 GPUs with 40GB of memory each (for about two weeks).

A.2 Further Details on the Weibo and the Twitter Data

A.2.1 Political Sensitivity Service

The political sensitivity service is provided by Baidu, a major Chinese technology company, and is publicly available. The service uses a combination of banned keywords collected by Baidu and deep learning models to assign political sensitivity to text. The sensitivity score ranges from 0 to 1, with 1 being the most sensitive. Table A.2 reports several keywords (and keyword combinations) that were flagged by the service. All of them seem to be sensible keywords that could be considered sensitive, especially during the early COVID-19 pandemic.

Table A.2. Flagged Keywords

Keywords	Translation
政府, 蛀虫	government, parasite
武汉, 问责	Wuhan, accountability
湖北, 瞒报	Hubei, withhold information
颜色革命	color revolution
中国经济, 衰退	Chinese economy, slowdown

A.2.2 Details on the Weibo Data

Weibo data from Fu and Zhu (2020) were collected by the authors based on a list of 40 COVID-19 related keywords. The data contains 1,230,353 posts that were posted between December 1, 2019 and February 27, 2020 on Weibo.

Weibo data from Hu et al. (2020) were collected for a longer time span (December 1, 2019 - December 31, 2020) and were based on a more extensive list of keywords. To ensure compatibility, I use a subset of the data that includes posts that were posted between December 1, 2019 and February 27, 2020 and contain at least one of the 40 keywords in Fu and Zhu (2020). The subset contains 8,518,113 Weibo posts.

All Weibo posts were anonymized to remove tags and other user information.

A.2.3 Details on the Twitter Data

Twitter data was collected using the Twitter research API with the following restrictions: 1) the tweet was posted between December 1, 2019 and February 27, 2020; 2) the tweet contains at least one of the 40 keywords in Fu and Zhu (2020); and 3) the language of the tweet is identified as Chinese by Twitter. The restrictions were used to ensure compatibility with the Weibo data. Similar to the Weibo data, the Twitter data were anonymized to remove tags and other user information.

A.2.4 International Social Media Data Collection by Authoritarian Regimes

Here I present some qualitativa evidence that social media data from international platforms such as Twitter, Facebook, YouTube and TikTok is being collected en mass by authoritarian regimes, most notably China and Russia.

Publicly available information suggests that large sclae Chinese Twitter data has been collected and used for AI training in China. The Natural Language Processing and Information Retrieval sharing platform hosted by the Beijing Institute of Technology shows that at least a hundred million Chinese tweets have been collected and from which five million is made publicly available.[3] The Peacock Chinese Twitter Corpus (PCTC) is another dataset of 4.9 millionn Chinese tweets.[4] Information gathered in fieldwork also confirms that international social media data is being used to augment AI training data by Chinese technology companies.

Similarly, leaked documents from Russia suggest that Russia is monitoring and collecting massive amount of social media data from platforms like Twitter, Facebook, YouTube, and

[3] http://www.nlpir.org/wordpress/2018/02/01/nlpir-500%E4%B8%87%E6%9D%A1twitter%E5%86%85%E5%AE%B9%E8%AF%AD%E6%96%99%E5%BA%93/

[4] https://figshare.com/articles/dataset/Peacock_Chinese_Twitter_Corpus_PCTC_/13489239/1

Tiktok and is in the process of using such data to develop automated censorship systems.[5] In particular, documents show that one Russian company has been collecting data on the scale of 140 million messages in Russian and other languages spoken in the former Soviet Union and 40 million images per day from Facebook, Instagram, TikTok, Twitter, and other social media platforms since 2014.[6]

[5] See Алеся Мароховская, Ирина Долинина, Соня Савина, Редакция, Полина Ужвак, Катя Бончосмоловская "Внутри машины цензуры." Feburary 8, 2023. https://istories.media/stories/2023/02/08/vnutri-mashini-tsenzuri/

[6] https://static.istories.media/uploaded/documents/0b809ea16feb42c7b8c91b022e45bd6b.pdf

A.3 Additional Results

A.3.1 Alternative Measure of Performance

In addition to accuracy, another commonly used metric to evaluate the performance of deep learning models is the F1 score. The F1 score is defined as

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where precision is given by $(\text{no. of true positives})/(\text{no. of true positives} + \text{no. of false positives})$ and recall is given by $(\text{no. of true positives})/(\text{no. of true positives} + \text{no. of false negatives})$.

In simple terms, precision is the ability of a model to correctly identify positive instances (true positives) out of the total instances it predicts as positive. It focuses on minimizing false positives, meaning the instances that are wrongly classified as positive (censor). Recall is the ability of a model to correctly identify all the positive instances (true positives) out of the total actual positive instances. It focuses on minimizing false negatives, meaning the instances that are wrongly classified as negative (not censor).

The F1 score combines precision and recall into a single metric by taking their harmonic mean. The harmonic mean gives more weight to lower values, so the F1 score will be high only if both precision and recall are high. It ranges between 0 and 1, with 1 indicating perfect performance and 0 indicating poor performance.

Figure A.1 reports the F1 scores for the censorship AI models trained on different training datasets. Similar to the main results, the result based on the F1 score shows that as data missingness increases, the performance of the censorship AI models becomes worse and the drop in performance is significantly larger for the crisis test data than for the status quo test data. The result is in accordance with the error rate result in Figure 1.4, as the increasing false negative rate will pull down the F1 score.

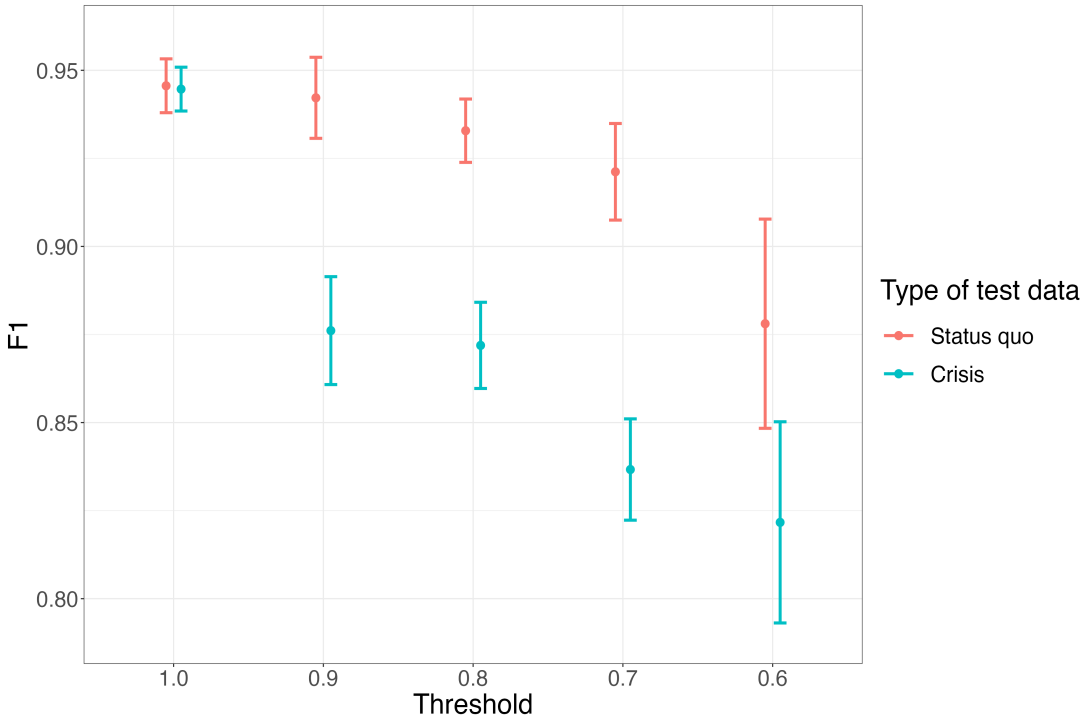

Figure A.1. Model Performance across Training Datasets

Note: Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

A.3.2 Error Rate Result for Models Trained on Larger Dataset

For models trained on the larger training datasets, the breakdown of the models' errors follows similar trends to the original models (Figure 1.4). Figure A.2 shows the false positive rate is low and stays relatively stable across different thresholds. However, as data missingness increases, the false negative rate increases drastically, with the largest false negative rate more than nine times that of the smallest. This is true for both the status quo test data and the crisis test data, with a larger increase in false negative rate during crisis.

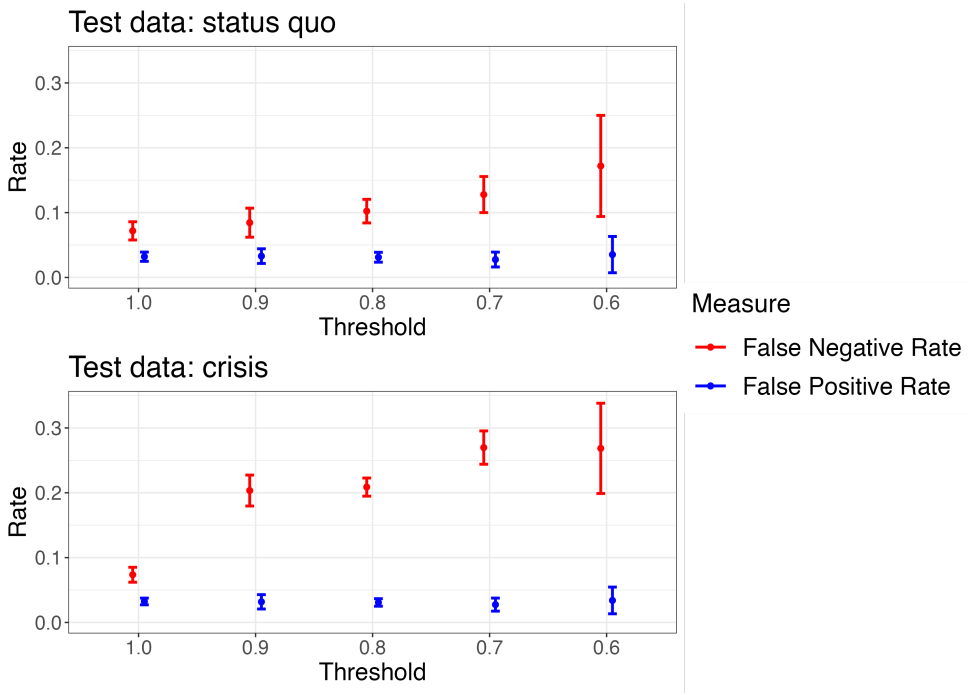

Figure A.2. False Positive Rate vs. False Negative Rate

A.3.3 Lower Cutoff, Data Leakage, Larger Model, and Alternative Model

Figure A.3 reports the results of several robustness checks. the substantive conclusions from the main results hold for all of the following alternatives. Specifically:

- Lower cutoff: Uses 0.4 instead of 0.5 to generate positive censorship labels, i.e., social media posts with sensitivity scores above 0.4 have a censorship label of 1 and those with scores below 0.4 have a censorship label of 0.
- Data leakage: Allows imperfect preference falsification and self-censorship by allowing 10% of data from the missing part of the distribution to leak into the training datasets.
- Larger BERT model: Uses a larger BERT model with 325,524,482 trainable parameters instead of the BERT model with 102,269,186 parameters.
- ELECTRA model: Uses the ELECTRA deep learning model[7] (Clark et al., 2020) instead of BERT.

[7] https://huggingface.co/hfl/chinese-electra-base-discriminator

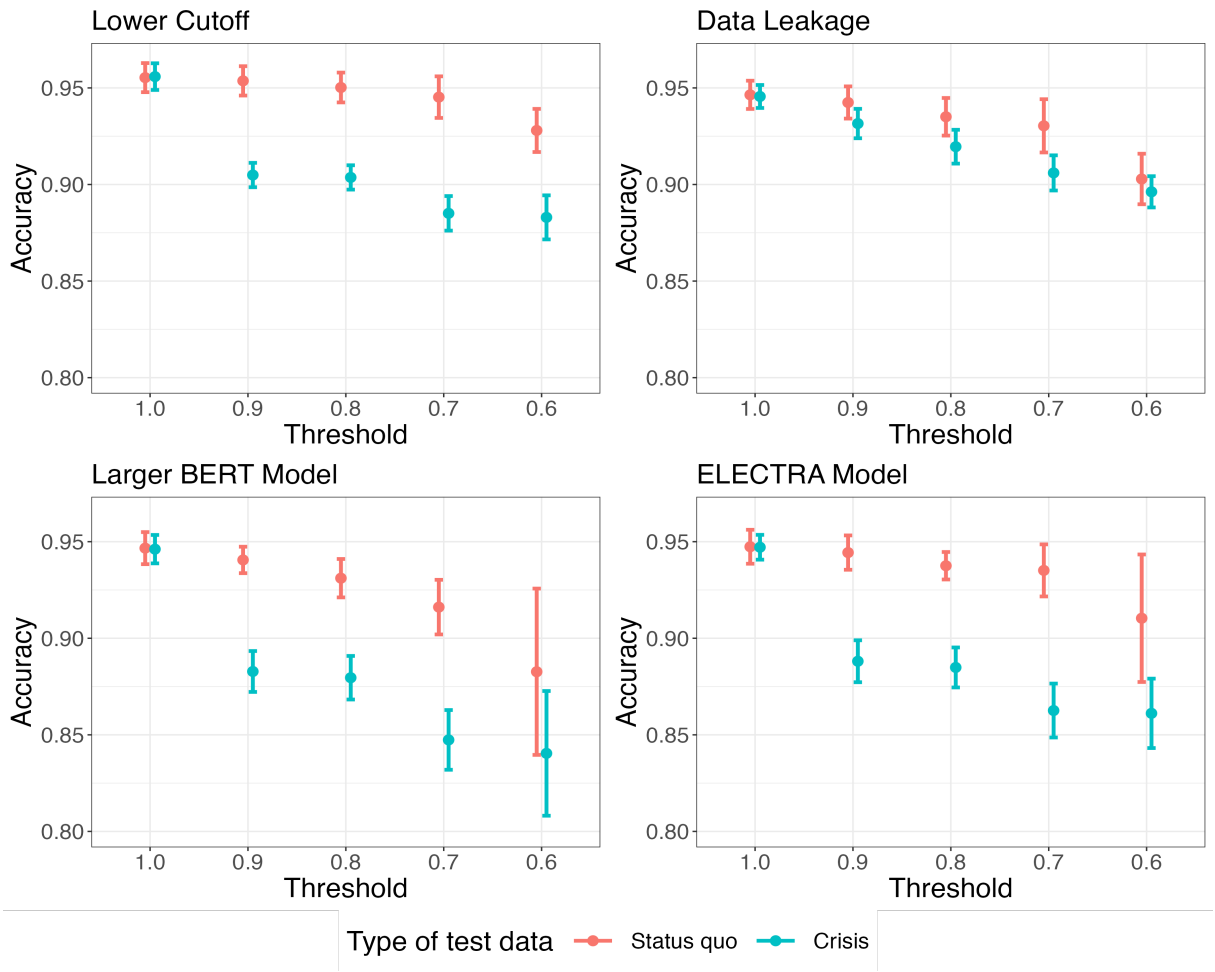

Figure A.3. Model Performance across Training Datasets

Note: Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

Appendix B

Supplementary Materials for Chapter 3

B.1 Further Details on the Criminal Sentencing Software

B.1.1 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

[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

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 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]

[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/

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

B.1.2 User interface of the criminal sentencing software

Figure B.1 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.

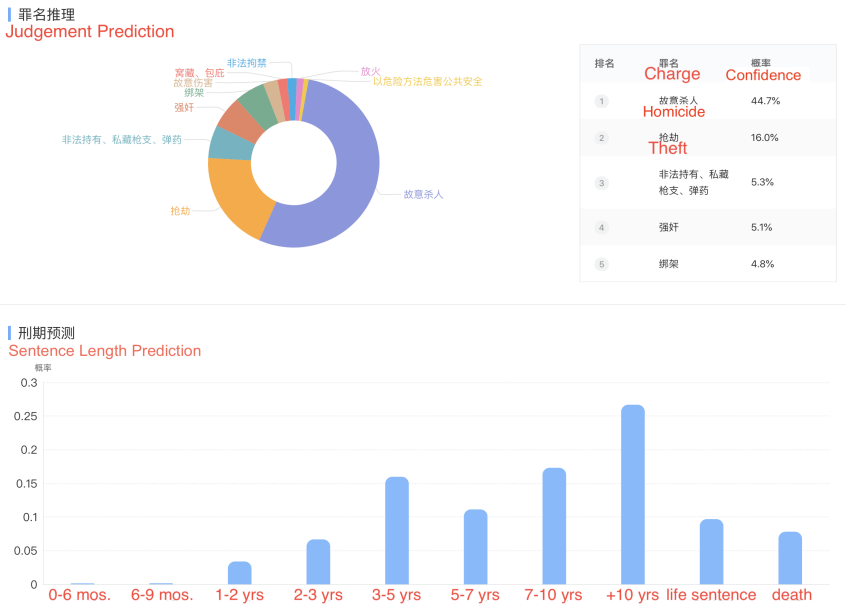

Figure B.1. Criminal Sentencing Software User Interface

B.1.3 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.

B.1.4 Model performance

Table B.1 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.

B.1.5 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).

B.2 Further details on the audit experiment

B.2.1 Graphical representation of the audit experiment

Figure B.2 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.2.4 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.

Table B.1. 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%

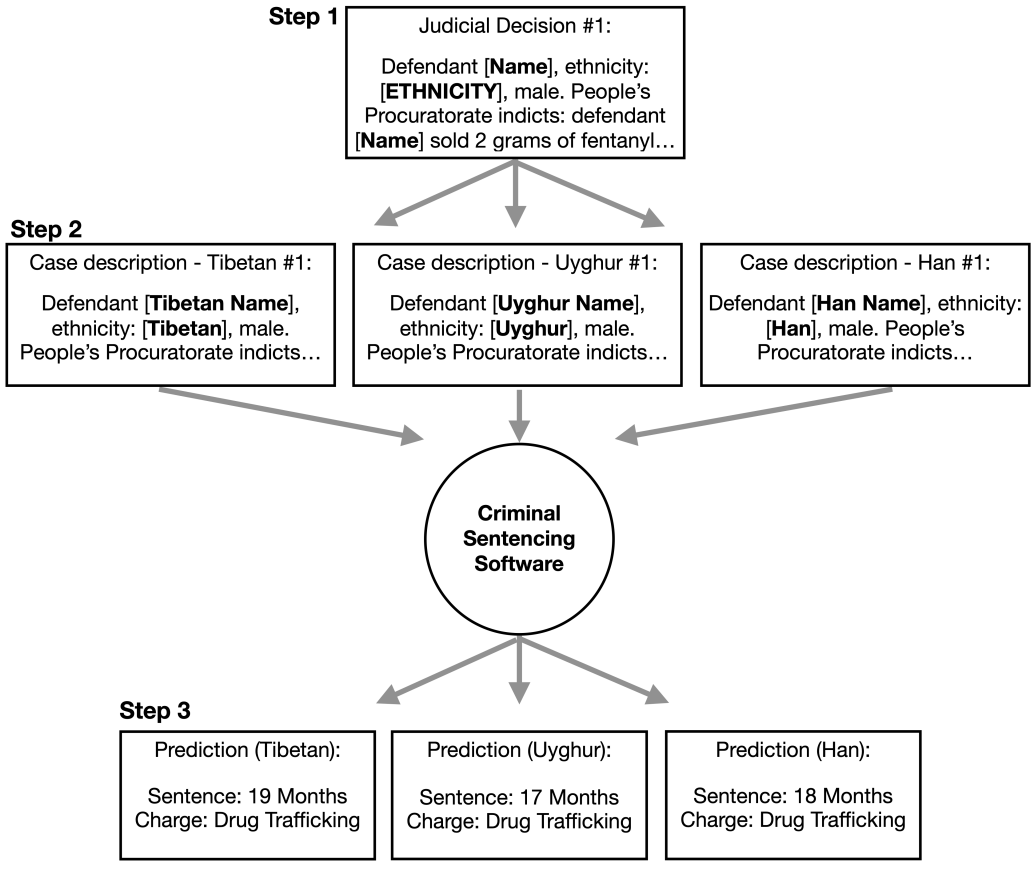

Figure B.2. Data Collection in the Audit Experiment

B.2.2 Example of case description in English

Figure B.3 is an example of a (truncated) case description translated to English.

> 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.

Figure B.3. Example of Case Description

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

B.2.3 Details on collection of court rulings

Table B.2 shows the number of court rulings across crime types.

Table B.2. 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: 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.2.4 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.

B.3 Additional results

B.3.1 Effect by ethnic cues

Table B.3 and Table B.4 show the effects of defendant name and defendant ethnicity respectively. Results in Table B.3 are based on case descriptions that exclude the ethnic identity of the defendants. Results in Table B.4 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 B.3. 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.

B.3.2 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 B.4). Here I report results from two alternative models that do not transform the outcomes. Table B.5 reports results using negative binomial regression on the un-transformed outcomes and Table B.6 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.

Table B.4. 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.

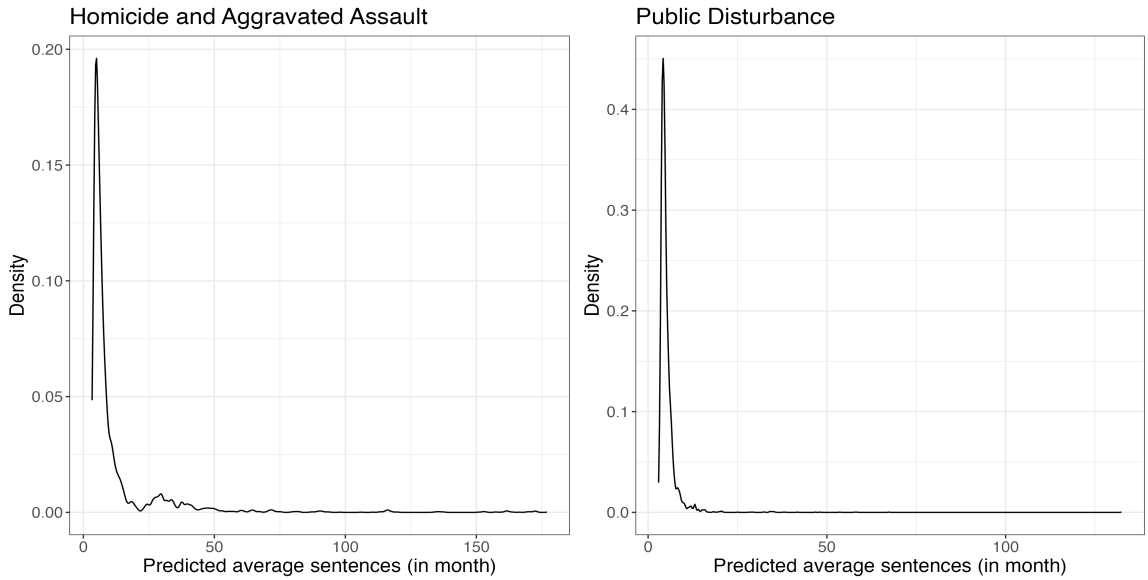

Figure B.4. Density Plot of Average Sentences for Homicides and Aggravated Assaults, Public Disturbance

Table B.5. 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 B.6. 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.

B.3.3 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:

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

where Y_i 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 B.7 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).

B.3.4 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, Jefferson and Rendleman, 2023). Figure B.5 provides a visual representation of the two causal mechanisms.

Table B.8 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 B.5). 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 B.9 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

Table B.7. 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 mis-classification is the outcome variable.

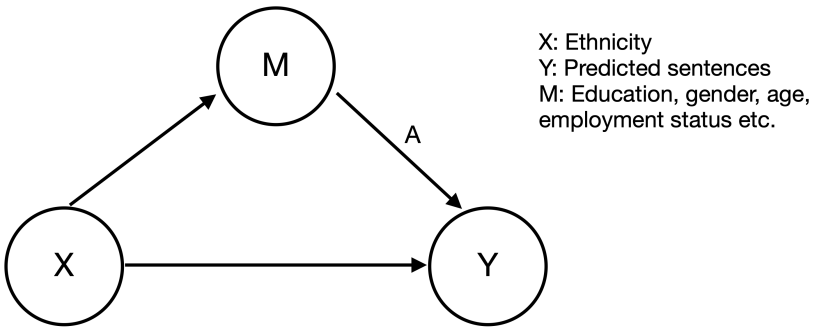

Figure B.5. Directed acyclic graph of the causal relations between ethnicity and predicted sentences

of education level on predicted sentences are small (1.2 years change in education, the largest estimate from Table B.8, 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 B.9 thus suggests that there is no strong evidence that ethnic cues result in longer predicted sentences either through education or gender.

Table B.8. 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 B.9. 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.

Bibliography

Abadji, Julien, Pedro Javier Ortiz Suárez, Laurent Romary and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event) Mannheim: Leibniz-Institut für Deutsche Sprache pp. 1 – 9.
URL: *https://nbn-resolving.org/urn:nbn:de:bsz:mh39-104688*

Abadji, Julien, Pedro Ortiz Suarez, Laurent Romary and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association pp. 4344–4355.
URL: *https://aclanthology.org/2022.lrec-1.463*

Agrawal, Ajay, Joshua S Gans and Avi Goldfarb. 2019. "Artificial intelligence: the ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives* 33(2):31–50.

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

Allie, Feyaad. 2023. "Facial Recognition Technology and Voter Turnout." *The Journal of Politics* 85(1):328–333.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine bias. ProPublica, May 23, 2016.".

Aral, Sinan and Dean Eckles. 2019. "Protecting elections from social media manipulation." *Science* 365(6456):858–861.

Baik, Jongyoon and Ling Dai. 2022. "Administrative Lawsuits Dataset." *Working Paper* .

Bamman, David, Brendan O'Connor and Noah Smith. 2012. "Censorship and deletion practices in Chinese social media." *First Monday* .

Barocas, Solon and Andrew D Selbst. 2016. "Big data's disparate impact." *Calif. L. Rev.* 104:671.

Barocas, Solon, Moritz Hardt and Arvind Narayanan. 2018. "Fairness and Machine Learning. fairmlbook.org, 2019.".

Ben-Michael, Eli, D James Greiner, Kosuke Imai and Zhichao Jiang. 2021. "Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment." *arXiv preprint arXiv:2109.11679* .

Benjamin, Ruha. 2019. *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Beraja, Martin, Andrew Kao, David Y Yang and Noam Yuchtman. 2021. AI-tocracy. Technical report National Bureau of Economic Research.

Beraja, Martin, Andrew Kao, David Y Yang and Noam Yuchtman. 2023. *Exporting the surveillance state via trade in AI*. Brookings Institution.

Beraja, Martin, David Y Yang and Noam Yuchtman. 2020. Data-intensive innovation and the State: evidence from AI firms in China. Technical report National Bureau of Economic Research.

Berinsky, Adam J. 1999. "The two faces of public opinion." *American Journal of Political Science* pp. 1209–1230.

Biderman, Stella, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit and Edward Raff. 2023. Emergent and Predictable Memorization in Large Language Models. In *Advances in Neural Information Processing Systems*, ed. A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt and S. Levine. Vol. 36 Curran Associates, Inc. pp. 28072–28090.

Bird, Sarah, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach and Kathleen Walker. 2020. "Fairlearn: A toolkit for assessing and improving fairness in AI." *Microsoft, Tech. Rep. MSR-TR-2020-32* .

Boumans, Jelle, Damian Trilling, Rens Vliegenthart and Hajo Boomgaarden. 2018. "The agency makes the (online) news world go round: The impact of news agency content on print and online news." *International Journal of Communication* 12:22.

Brady, Anne-Marie. 2009. *Marketing dictatorship: Propaganda and thought work in contemporary China*. Rowman & Littlefield Publishers.

Broussard, Meredith. 2023. *More than a glitch: Confronting race, gender, and ability bias in tech*. MIT Press.

Brox, Trine and Ildikó Bellér-Hann. 2014. *On the fringes of the harmonious society: Tibetans*

and Uyghurs in socialist China. Nias Press.

Budach, Lukas, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann and Hazar Harmouch. 2022. "The Effects of Data Quality on Machine Learning Performance.".

Buolamwini, Joy and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR pp. 77–91.

Cagé, Julia, Nicolas Hervé and Marie-Luce Viaud. 2020. "The production of information in an online world." *The Review of economic studies* 87(5):2126–2164. Publisher: Oxford University Press.

Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer and Chiyuan Zhang. 2022. "Quantifying memorization across neural language models." *arXiv preprint arXiv:2202.07646* .

Carter, Erin Baggott and Brett L Carter. 2021. "When autocrats threaten citizens with violence: Evidence from China." *British Journal of Political Science* pp. 1–26.

Carter, Erin Baggott and Brett L Carter. 2023. *Propaganda in autocracies: institutions, information, and the politics of belief*. Cambridge University Press.

Chen, Jidong and Yiqing Xu. 2017. "Why do authoritarian regimes allow citizens to voice opinions publicly?" *The Journal of Politics* 79(3):792–803.

Chen, Lichang, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang et al. 2023. "Alpagasus: Training a better alpaca with fewer data." *arXiv preprint arXiv:2307.08701* .

Choi, Donghyun Danny, J Andrew Harris and Fiona Shen-Bayh. 2022. "Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya." *American Political Science Review* pp. 1–14.

Christin, Angèle. 2018. "Counting clicks: Quantification and variation in web journalism in the United States and France." *American Journal of Sociology* 123(5):1382–1415.

Clark, Kevin, Minh-Thang Luong, Quoc V Le and Christopher D Manning. 2020. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* .

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

Cook, Cynthia M, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton and Arun R Vemury. 2019. "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1(1):32–41.

Cox, Gary W. 2009. Authoritarian elections and leadership succession, 1975-2004. In *APSA 2009 Toronto meeting paper*.

Crabtree, Charles, Holger L Kern and David A Siegel. 2020. "Cults of personality, preference falsification, and the dictator's dilemma." *Journal of Theoretical Politics* 32(3):409–434.

Creemers, Rogier. 2017. "Cyber China: Upgrading propaganda, public opinion work and social management for the twenty-first century." *Journal of contemporary China* 26(103):85–100.

Cui, Yiming, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang and Guoping Hu. 2020. "Revisiting pre-trained models for Chinese natural language processing." *arXiv preprint arXiv:2004.13922* .

Dai, Andrew M and Quoc V Le. 2015. "Semi-supervised sequence learning." *Advances in neural information processing systems* 28.

Davenport, Lauren, Hakeem Jefferson and Hunter Rendleman. 2023. "Deconstructing Race." *Working Paper* .

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* .

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

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024. "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models." *Advances in Neural Information Processing Systems* 36.

Egorov, Georgy and Konstantin Sonin. 2020. The political economics of non-democracy. Technical report National Bureau of Economic Research.

Egorov, Georgy, Sergei Guriev and Konstantin Sonin. 2009. "Why resource-poor dictators allow freer media: A theory and evidence from panel data." *American political science Review* 103(4):645–668.

Esarey, Ashley. 2015. "Winning hearts and minds? Cadres as microbloggers in China." *Journal*

of Current Chinese Affairs 44(2):69–103.

Esberg, Jane. 2022. "Employment Restriction as Repression: Evidence from Argentina's Film Industry." *Working Paper* .

Esberg, Jane and Alexandra A Siegel. 2021. "How Exile Shapes Online Opposition: Evidence from Venezuela." *American Political Science Review* pp. 1–18.

Farrell, Henry, Abraham Newman and Jeremy Wallace. 2022. "Spirals of Delusion: How AI Distorts Decision-Making and Makes Dictators More Dangerous." *Foreign Aff.* 101:168.

Feldstein, Steven. 2019*a*. *The global expansion of AI surveillance*. Vol. 17 Carnegie Endowment for International Peace Washington, DC.

Feldstein, Steven. 2019*b*. "The road to digital unfreedom: How artificial intelligence is reshaping repression." *Journal of Democracy* 30(1):40–52.

Field, Anjalie, Su Lin Blodgett, Zeerak Waseem and Yulia Tsvetkov. 2021. "A survey of race, racism, and anti-racism in NLP." *arXiv preprint arXiv:2106.11410* .

Fourcade, Marion and Kieran Healy. 2024. *The Ordinal Society*. Harvard University Press.

Fu, King-wa, Chung-hong Chan and Michael Chau. 2013. "Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy." *IEEE internet computing* 17(3):42–50.

Fu, King-wa and Yuner Zhu. 2020. "Did the world overlook the media's early warning of COVID-19?" *Journal of Risk Research* 23(7-8):1047–1051.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gohdes, Anita R. 2020. "Repression technology: Internet accessibility and state violence." *American Journal of Political Science* 64(3):488–503.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. "Fake news on Twitter during the 2016 US presidential election." *Science* 363(6425):374–378.

Grossman, Guy, Oren Gazal-Ayal, Samuel D Pimentel and Jeremy M Weinstein. 2016. "Descriptive representation and judicial outcomes in multiethnic societies." *American Journal of Political Science* 60(1):44–69.

Grother, Patrick, Mei Ngan and Kayee Hanaoka. 2019. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD.

Guriev, Sergei and Daniel Treisman. 2019. "Informational autocrats." *Journal of economic perspectives* 33(4):100–127.

Guriev, Sergei and Daniel Treisman. 2020. "A theory of informational autocracy." *Journal of public economics* 186:104158.

Guriev, Sergei and Daniel Treisman. 2022. *Spin dictators: The changing face of tyranny in the 21st century*. Princeton University Press.

Hale, Henry E. 2022. "Authoritarian rallying as reputational cascade? Evidence from Putin's popularity surge after Crimea." *American Political Science Review* 116(2):580–594.

Han, Rongbin. 2018. *Contesting cyberspace in China: Online expression and authoritarian resilience*. Columbia University Press.

Hobbs, William R and Margaret E Roberts. 2018. "How sudden censorship can increase access to information." *American Political Science Review* 112(3):621–636.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier et al. 2021. "Knowledge graphs." *ACM Computing Surveys (CSUR)* 54(4):1–37.

Hou, Yue, Chuyu Liu and Charles Crabtree. 2020. "Anti-muslim bias in the Chinese labor market." *Journal of Comparative Economics* 48(2):235–250.

Hou, Yue and Rory Truex. 2022. "Ethnic discrimination in criminal sentencing in china." *The Journal of Politics* 84(4):000–000.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang and Weizhu Chen. 2021. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* .

Hu, Yong, Heyan Huang, Anfan Chen and Xian-Ling Mao. 2020. "Weibo-COV: A large-scale COVID-19 social media dataset from Weibo." *arXiv preprint arXiv:2005.09174* .

Huang, Haifeng. 2015. "Propaganda as signaling." *Comparative Politics* 47(4):419–444.

Huang, Haifeng. 2018. "The pathology of hard propaganda." *The Journal of Politics* 80(3):1034–1038.

Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen and Sooahn Shin. 2020. "Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment." *arXiv preprint arXiv:2012.02845* .

Jiang, Junyan and Dali L Yang. 2016. "Lying or believing? Measuring preference falsification from a political purge in China." *Comparative Political Studies* 49(5):600–634.

Johnson, Rebecca Ann and Simone Zhang. 2022. What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in US Social Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. pp. 1671–1682.

Jowett, Garth S and Victoria O'Donnell. 2018. *Propaganda & persuasion*. Sage publications.

Kalinin, Kirill. 2016. "The social desirability bias in autocrat's electoral ratings: evidence from the 2012 Russian presidential elections." *Journal of elections, public opinion and parties* 26(2):191–211.

Kärkkäinen, Kimmo and Jungseock Joo. 2019. "Fairface: Face attribute dataset for balanced race, gender, and age." *arXiv preprint arXiv:1908.04913* .

Kay, Matthew, Cynthia Matuszek and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*. pp. 3819–3828.

Kendall-Taylor, Andrea, Erica Frantz and Joseph Wright. 2020. "The digital dictators: How technology strengthens autocracy." *Foreign Aff.* 99:103.

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American political science Review* 107(2):326–343.

King, Gary, Jennifer Pan and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199):1251722.

King, Gary, Jennifer Pan and Margaret E Roberts. 2017. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American political science review* 111(3):484–501.

Kiros, Ryan, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba and Sanja Fidler. 2015. "Skip-thought vectors." *Advances in neural information processing systems* 28.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The quarterly journal of economics* 133(1):237–293.

Kokas, Aynne. 2023. *Trafficking data: How China is winning the battle for digital sovereignty*.

Oxford University Press.

Kotek, Hadas, Rikker Dockum and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*. pp. 12–24.

Kuran, Timur. 1991. "Now out of never: The element of surprise in the East European revolution of 1989." *World politics* 44(1):7–48.

Kuran, Timur. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al. 2018. "The science of fake news." *Science* 359(6380):1094–1096.

Lee, Kai-Fu. 2018. *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin.

Liang, Fan, Yuchen Chen and Fangwei Zhao. 2021. "The platformization of propaganda: How Xuexi Qiangguo expands persuasion and assesses citizens in China." *International Journal of Communication* 15:20.

Liebman, Benjamin L. 2015. "Leniency in Chinese criminal law: everyday justice in Henan." *Berkeley J. Int'l L.* 33:153.

Liebman, Benjamin L, Margaret E Roberts, Rachel E Stern and Alice 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.

Lohmann, Susanne. 1994. "The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91." *World politics* 47(1):42–101.

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of political science* 58(2):402–414.

Lu, Yingdan and Jennifer Pan. 2021. "Capturing clicks: How the Chinese government uses clickbait to compete for visibility." *Political Communication* 38(1-2):23–54.

Luo, Ruixuan, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren and Xu Sun. 2019. "PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation." *CoRR* abs/1906.11455. **URL:** *https://arxiv.org/abs/1906.11455*

Mandal, Paul K and Rakeshkumar Mahto. 2022. "An FNet based Auto Encoder for Long Sequence News Story Generation." *arXiv preprint arXiv:2211.08295* .

Mattingly, Daniel C and Elaine Yao. 2022. "How soft propaganda persuades." *Comparative Political Studies* 55(9):1569–1594.

McNamee, Lachlan and Anna Zhang. 2019. "Demographic engineering and international conflict: Evidence from China and the Former USSR." *International Organization* 73(2):291–327.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. "A survey on bias and fairness in machine learning." *ACM Computing Surveys (CSUR)* 54(6):1–35.

Metaxa, Danaë, Michelle A Gan, Su Goh, Jeff Hancock and James A Landay. 2021. "An image of society: Gender and racial representation and impact in image search results for occupations." *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–23.

Miller, Michael K. 2015. "Elections, information, and policy responsiveness in autocratic regimes." *Comparative Political Studies* 48(6):691–727.

Nguyen, Thuat, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi and Thien Huu Nguyen. 2023. "CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages.".

Nicholls, Tom. 2019. "Detecting textual reuse in news stories, at scale." *International Journal of Communication* 13(2019). Publisher: University of Southern California, Annenberg School for Communication ....

Nicholson, Stephen P and Haifeng Huang. 2022. "Making the List: Reevaluating Political Trust and Social Desirability in China." *American Political Science Review* pp. 1–8.

Noble, Safiya Umoja. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Omiye, Jesutofunmi A, Jenna C Lester, Simon Spichak, Veronica Rotemberg and Roxana Daneshjou. 2023. "Large language models propagate race-based medicine." *NPJ Digital Medicine* 6(1):195.

O'neil, Cathy. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ortiz Su'arez, Pedro Javier, Benoit Sagot and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019 Mannheim: Leibniz-Institut f"ur Deutsche Sprache pp. 9 – 16.
URL: *http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215*

Ortiz Su'arez, Pedro Javier, Laurent Romary and Benoit Sagot. 2020. A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics pp. 1703–1714.
URL: *https://www.aclweb.org/anthology/2020.acl-main.156*

Pan, Jennifer and Alexandra A Siegel. 2020. "How Saudi crackdowns fail to silence online dissent." *American Political Science Review* 114(1):109–125.

Pan, Jennifer, Zijie Shao and Yiqing Xu. 2022. "How government-controlled media shifts policy attitudes through framing." *Political Science Research and Methods* 10(2):317–332.

Peisakhin, Leonid and Arturas Rozenas. 2018. "Electoral effects of biased media: Russian television in Ukraine." *American journal of political science* 62(3):535–550.

Qin, Bei, David Strömberg and Yanhui Wu. 2018. "Media bias in China." *American Economic Review* 108(9):2442–2476.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. 2019. "Language models are unsupervised multitask learners." *OpenAI blog* 1(8):9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu. 2019. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *arXiv e-prints* .

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu. 2023. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.".

Repnikova, Maria and Kecheng Fang. 2019. "Digital media experiments in China:"Revolutionizing" persuasion under Xi Jinping." *The China Quarterly* 239:679–701.

Roberts, Margaret E. 2018. Censored. In *Censored*. Princeton University Press.

Roberts, Margaret E. 2020. "Resilience to online censorship." *Annual Review of Political Science* 23:401–419.

Robinson, Darrel and Marcus Tannenberg. 2019. "Self-censorship of regime support in authoritarian states: Evidence from list experiments in China." *Research & Politics* 6(3):2053168019856449.

Robinson, Joseph P, Gennady Livitz, Yann Henon, Can Qin, Yun Fu and Samson Timoner. 2020. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–1.

Rozenas, Arturas. 2010. Forced consent: information and power in non-democratic elections. In *APSA 2010 Annual Meeting Paper*.

Rozenas, Arturas and Denis Stukal. 2019. "How autocrats manipulate economic news: Evidence from Russia's state-controlled television." *The Journal of Politics* 81(3):982–996.

Sanovich, Sergey, Denis Stukal and Joshua A Tucker. 2018. "Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia." *Comparative politics* 50(3):435–482.

Scheible, Raphael, Fabian Thomczyk, Patric Tippmann, Victor Jaravine and Martin Boeker. 2020. "GottBERT: a pure German Language Model.".

Selb, Peter and Simon Munzert. 2018. "Examining a most likely case for strong campaign effects: Hitler's speeches and the rise of the Nazi party, 1927–1933." *American Political Science Review* 112(4):1050–1066.

Sen, Maya and Omar 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.

Serrano, Alejandro Vaca, Guillem Garcia Subies, Helena Montoro Zamorano, Nuria Aldama Garcia, Doaa Samy, David Betancur Sanchez, Antonio Moreno Sandoval, Marta Guerrero Nieto and Alvaro Barbero Jimenez. 2022. "Rigoberta: A state-of-the-art language model for spanish." *arXiv preprint arXiv:2205.10233* .

Shalumov, Vitaly and Harel Haskey. 2023. "Hero: Roberta and longformer hebrew language models." *arXiv preprint arXiv:2304.11077* .

Shambaugh, David. 2017. China's propaganda system: Institutions, processes and efficacy. In *Critical readings on the communist party of China (4 Vols. Set)*. Brill pp. 713–751.

Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan and Huazhu Fu. 2022. "Transformers in medical imaging: A survey." *arXiv preprint arXiv:2201.09873* .

Shayegani, Erfan, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong and Nael Abu-Ghazaleh. 2023. "Survey of vulnerabilities in large language models revealed by adversarial attacks." *arXiv preprint arXiv:2310.10844* .

Shen-Bayh, Fiona Feiang. 2022. *Undue Process*. Cambridge University Press.

Shen, Xiaoxiao and Rory Truex. 2021. "In search of self-censorship." *British Journal of Political Science* 51(4):1672–1684.

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan and Nanyun Peng. 2019. "The woman worked as a babysitter: On biases in language generation." *arXiv preprint arXiv:1909.01326* .

Shih, Victor Chung-Hon. 2008. ""Nauseating" displays of loyalty: Monitoring the factional bargain through ideological campaigns in China." *The Journal of Politics* 70(4):1177–1192.

Shliazhko, Oleh, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova and Tatiana Shavrina. 2022. "mgpt: Few-shot learners go multilingual." *arXiv preprint arXiv:2204.07580* .

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton et al. 2017. "Mastering the game of go without human knowledge." *nature* 550(7676):354–359.

Stern, Rachel E, Benjamin L Liebman, Margaret E Roberts and Alice Z Wang. 2020. "Automating fairness? Artificial intelligence in the Chinese courts." *Colum. J. Transnat'l L.* 59:515.

Stockmann, Daniela. 2013. *Media commercialization and authoritarian rule in China*. Cambridge University Press.

Storkey, Amos et al. 2009. "When training and test sets are different: characterizing learning transfer." *Dataset shift in machine learning* 30:3–28.

Sun, Yu, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian and Hua Wu. 2019. "Ernie: Enhanced representation through knowledge integration." *arXiv preprint arXiv:1904.09223* .

Svolik, Milan. 2018. "When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents." *Available at SSRN 3243470* .

Svolik, Milan W. 2019. "Polarization versus democracy." *Journal of Democracy* 30(3):20–32.

Tanash, Rima S, Zhouhan Chen, Dan S Wallach and Melissa Marschall. 2017. The Decline of Social Media Censorship and the Rise of Self-Censorship after the 2016 Failed Turkish Coup. In *FOCI@ USENIX Security Symposium*.

Tannenberg, Marcus. 2022. "The autocratic bias: self-censorship of regime support." *Democratization* 29(4):591–610.

Trinh, Minh D. 2023. "Statistical Misreporting Debilitates Authoritarian Governance." *Working Paper* .

Truex, Rory. 2019. "Focal points, dissident calendars, and preemptive repression." *Journal of Conflict Resolution* 63(4):1032–1052.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* .

Vangara, Kushal, Michael C King, Vitor Albiero, Kevin Bowyer et al. 2019. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0.

Voigtländer, Nico and Hans-Joachim Voth. 2015. "Nazi indoctrination and anti-Semitic beliefs in Germany." *Proceedings of the National Academy of Sciences* 112(26):7931–7936.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359(6380):1146–1151.

Waight, Hannah, Yin Yuan, Margaret E. Roberts and Stewart Brandon. n.d. "Speaking with the State's Voice: The Decade-Long Growth of Government-Authored News Media in China under Xi Jinping.".

Wallace, Jeremy L. 2022. *Seeking Truth and Hiding Facts: Information, Ideology, and Authoritarianism in China*. Oxford University Press.

Wan, Yixin, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang and Nanyun Peng. 2023. ""Kelly is a warm person, Joseph is a role model": Gender biases in llm-generated reference letters." *arXiv preprint arXiv:2310.09219* .

Wang, Haiyan and Colin Sparks. 2019. "Chinese newspaper groups in the digital era: The resurgence of the party press." *Journal of Communication* 69(1):94–119.

Wang, Mei, Yaobin Zhang and Weihong Deng. 2021. "Meta balanced network for fair face recognition." *IEEE transactions on pattern analysis and machine intelligence* 44(11):8433–8448.

Wedeen, Lisa. 2015. *Ambiguities of domination: Politics, rhetoric, and symbols in contemporary Syria*. University of Chicago Press.

Wintrobe, Ronald. 2000. *The political economy of dictatorship*. Cambridge University Press.

Wu, Xiaohan, Margaret E Roberts, Rachel E Stern, Benjamin L Liebman, Amarnath Gupta and Luke Sanford. 2022. "Augmenting Serialized Bureaucratic Data: The Case of Chinese Courts." *Available at SSRN 4124433* .

Xiao, Chaojun, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang et al. 2018. "Cail2018: A large-scale legal dataset

for judgment prediction." *arXiv preprint arXiv:1807.02478* .

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

Xu, Xu. 2023. "The Unintrusive Nature of Digital Surveillance and Its Social Consequences." *Working Paper* .

Yablokov, Ilya and Precious N Chatterje-Doody. 2021. *Russia Today and conspiracy theories: People, power and politics on RT*. Routledge.

Yang, Eddie. 2023. "Automated Repression: Ethnic Discrimination in AI-assisted Criminal Sentencing in China." *Working Paper* .

Yang, Eddie and Margaret 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, Mingcan and Xiaoming Li. 2009. "Empirical Analysis of Homicide: the perspective of 493 murder-cases (: 493)." *Criminal Science ( )* 18:1.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez and Kai-Wei Chang. 2017. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." *arXiv preprint arXiv:1707.09457* .