The Limits of AI for Authoritarian Control

Eddie Yang*

August 2023

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.

Keywords: Authoritarian politics, Artificial Intelligence, censorship, repression.

*Department of Political Science, University of California San Diego. Email: z5yang@ucsd.edu. A previous version of this paper was circulated under the title "The Digital Dictator's Dilemma."

1. Introduction

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]

[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

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, 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

[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.

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 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, 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.

2. 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

[5]See Appendix B.4 for details.

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).

3. 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 institu-

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

tions to explain in detail why digital dictators struggle to realize the repressive potential that AI promises.

3.1. 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).

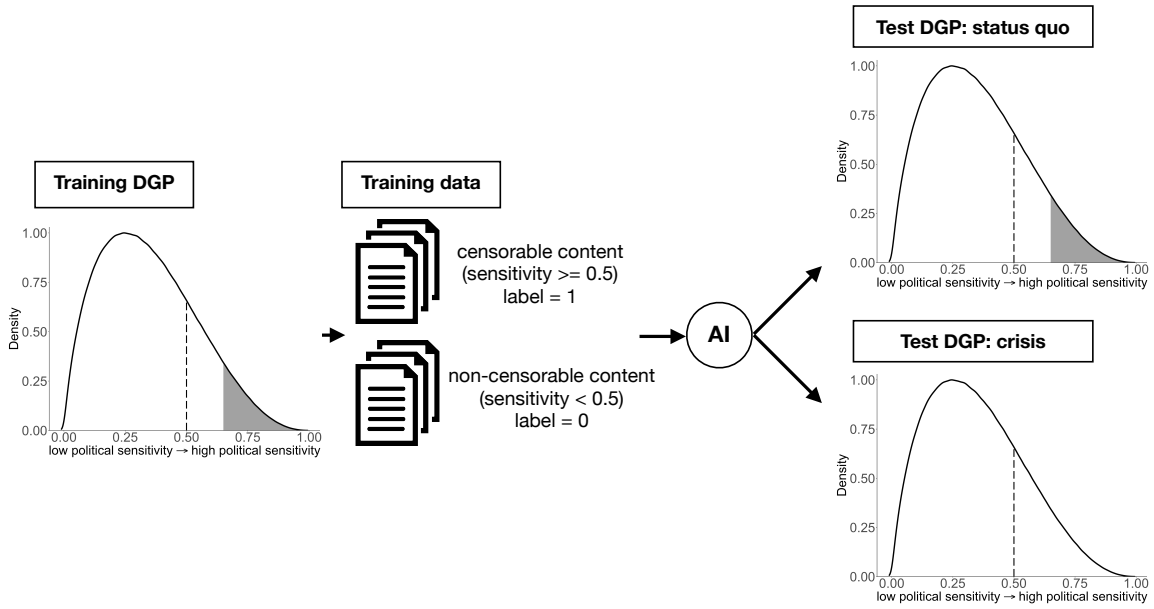
3.2. Automating Autocracy with Bad Data

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 censor-

ship AI in Figure 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).

FIGURE 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.

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

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

3.3. 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

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

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 B.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.

3.4. 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.

4. 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.

4.1. 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.

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.

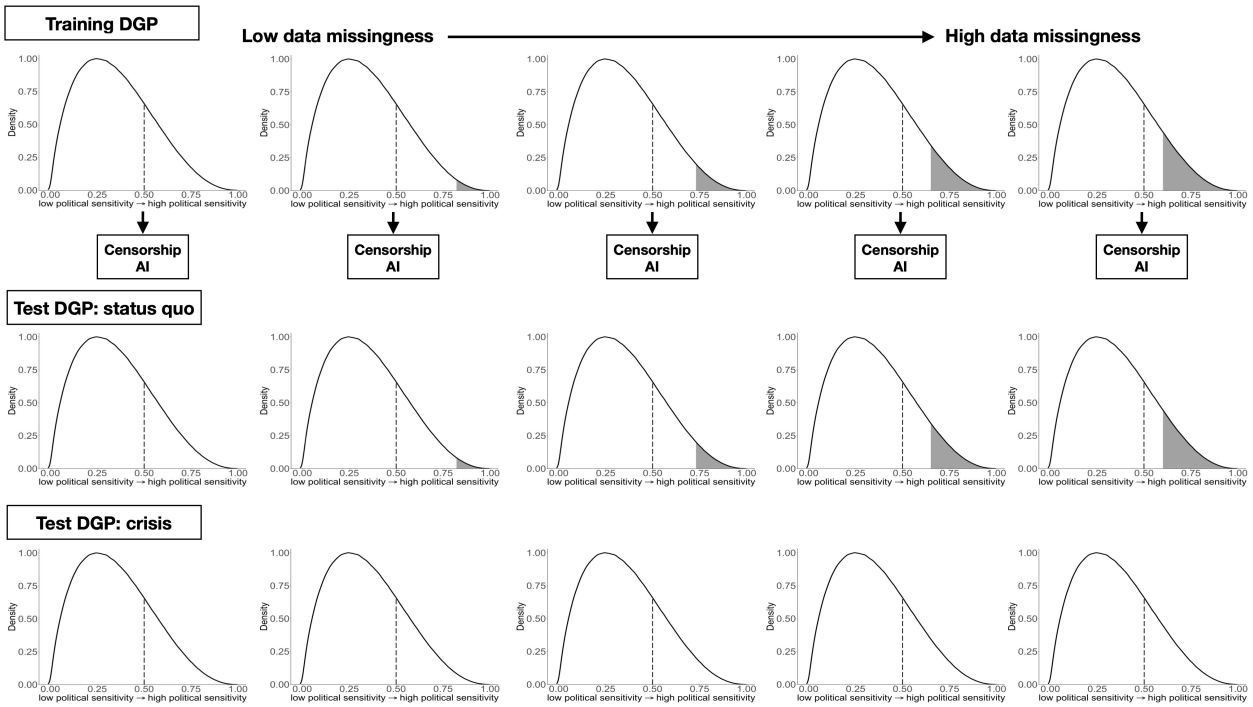
To model different degrees of data missingness due to preference falsification and self-censorship, I construct training datasets with different distributions of political sensitivity from the main sample. As Figure 2 shows, I construct five versions of training dataset with

[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.

[9] See Appendix B.1 for more details.

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

Figure 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.

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 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.

[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.

Table 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

Specifically, I use a Chinese version of BERT (Bidirectional Encoder Representations from 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 for model performance. More details about the training procedure are included in Appendix A.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 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

[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.

the Appendix.

4.2. 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.

5. Results

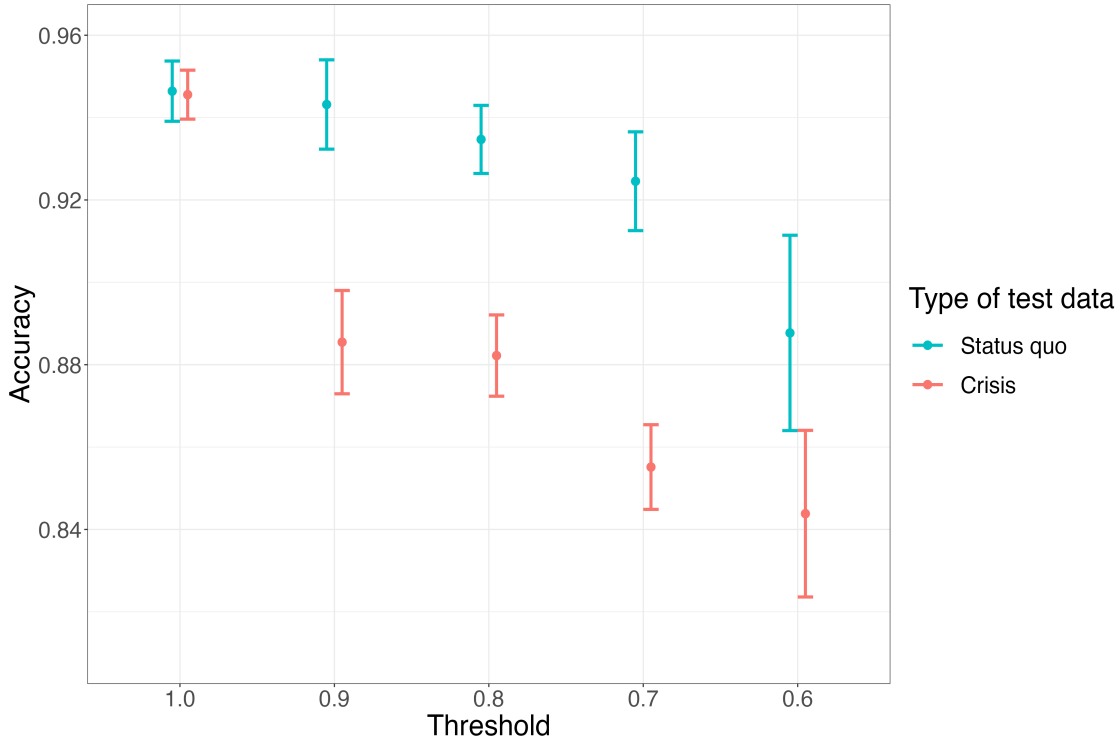
I first present evidence of the repression-performance trade-off and AI's performance gap on status quo and crisis test data. I then show evidence that adding more data to the training dataset has a marginal impact on AI's performance but augmenting with data from democratic sources results in a partial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

[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.

5.1. More Repression, Worse Performance

Figure 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.

FIGURE 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.

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,

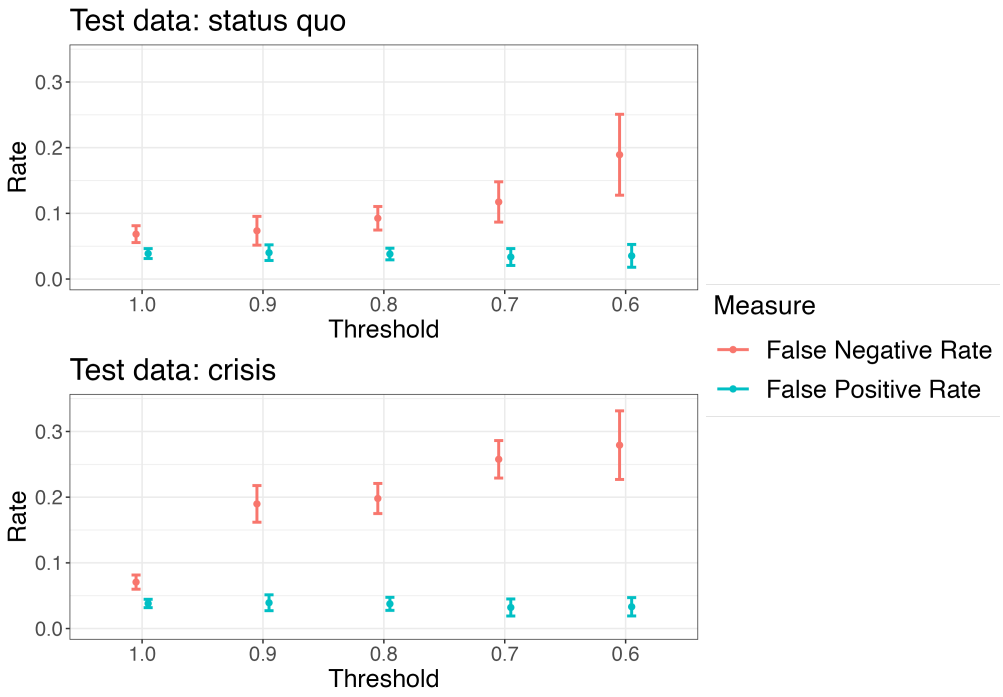
the decrease in model accuracy is significantly larger in crisis, when people reveal their true preferences, than in the status quo.

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 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 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 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.

Together, Figure 3 and Figure 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.

Figure 4. False Positive Rate vs. False Negative Rate

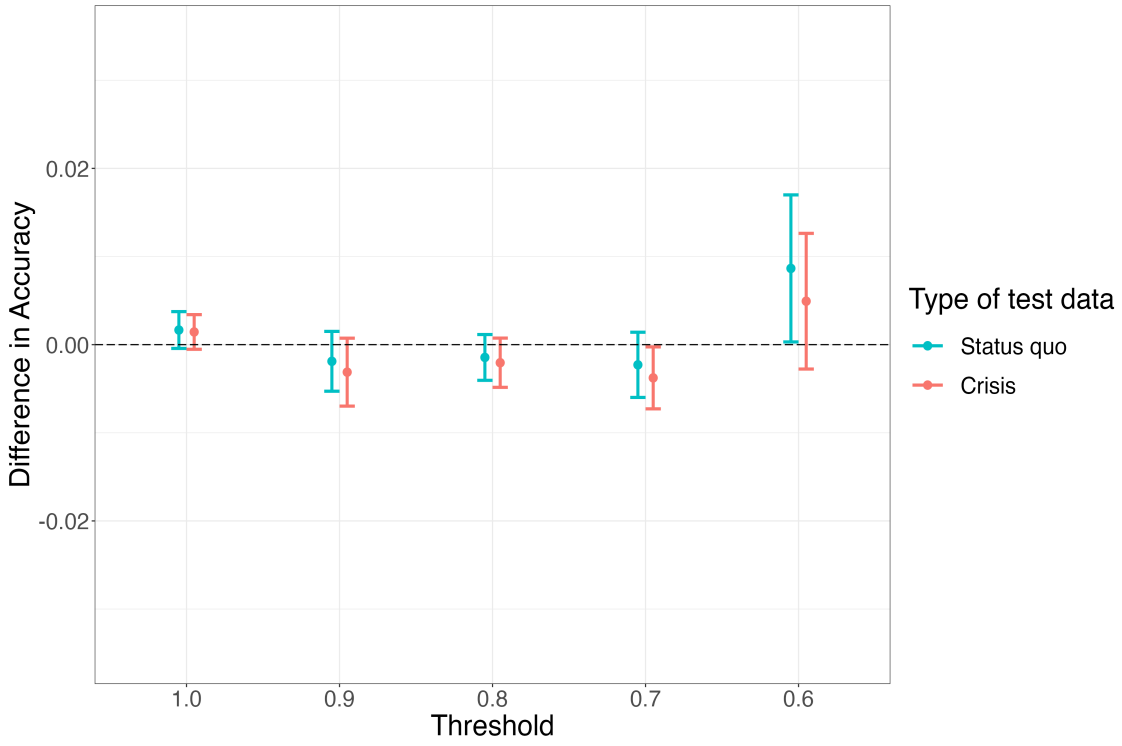


5.2. Marginal Impact of More Training Data

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 5 presents the result of doubling the size of the training dataset on model accuracy. Specifically, Figure 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 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

Figure 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.

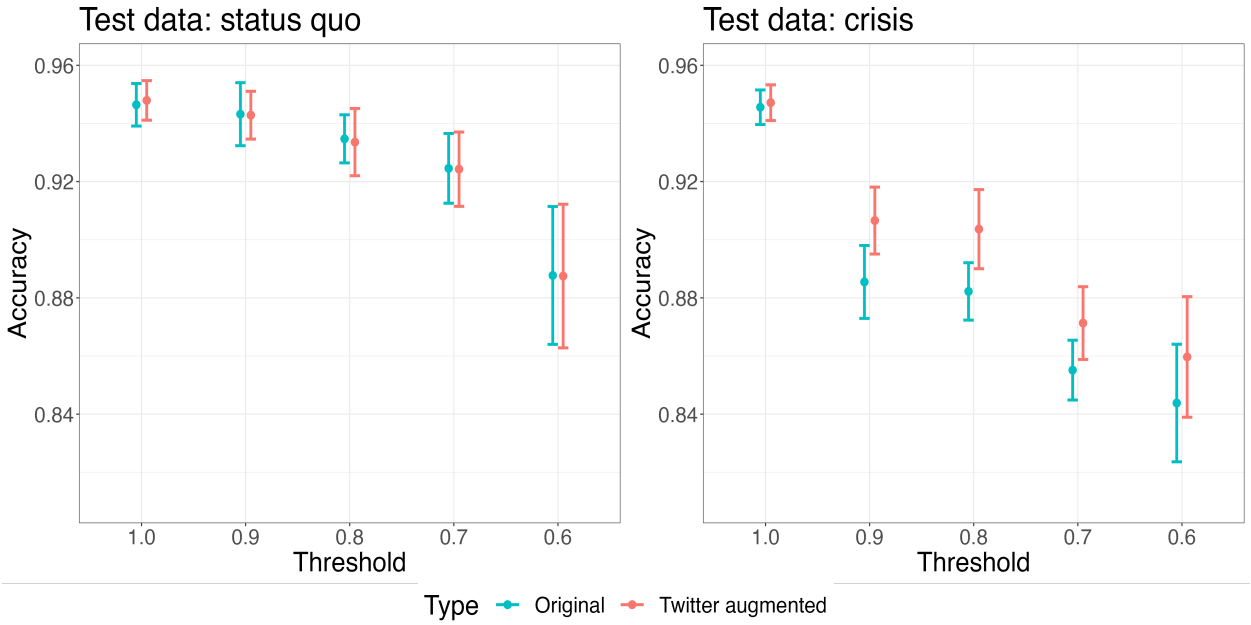
larger training datasets follows a similar trend to the original models - the false positive rate is low and stays relatively stable across different thresholds but the false negative rate increases drastically as data missingness increases.

5.3. 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 6 provides evidence that augmenting the original Weibo training dataset with data from Twitter improves model accuracy during crisis. Specifically, Figure 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

FIGURE 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.

status quo test data is sampled from the same distribution as the original training data. In this 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.

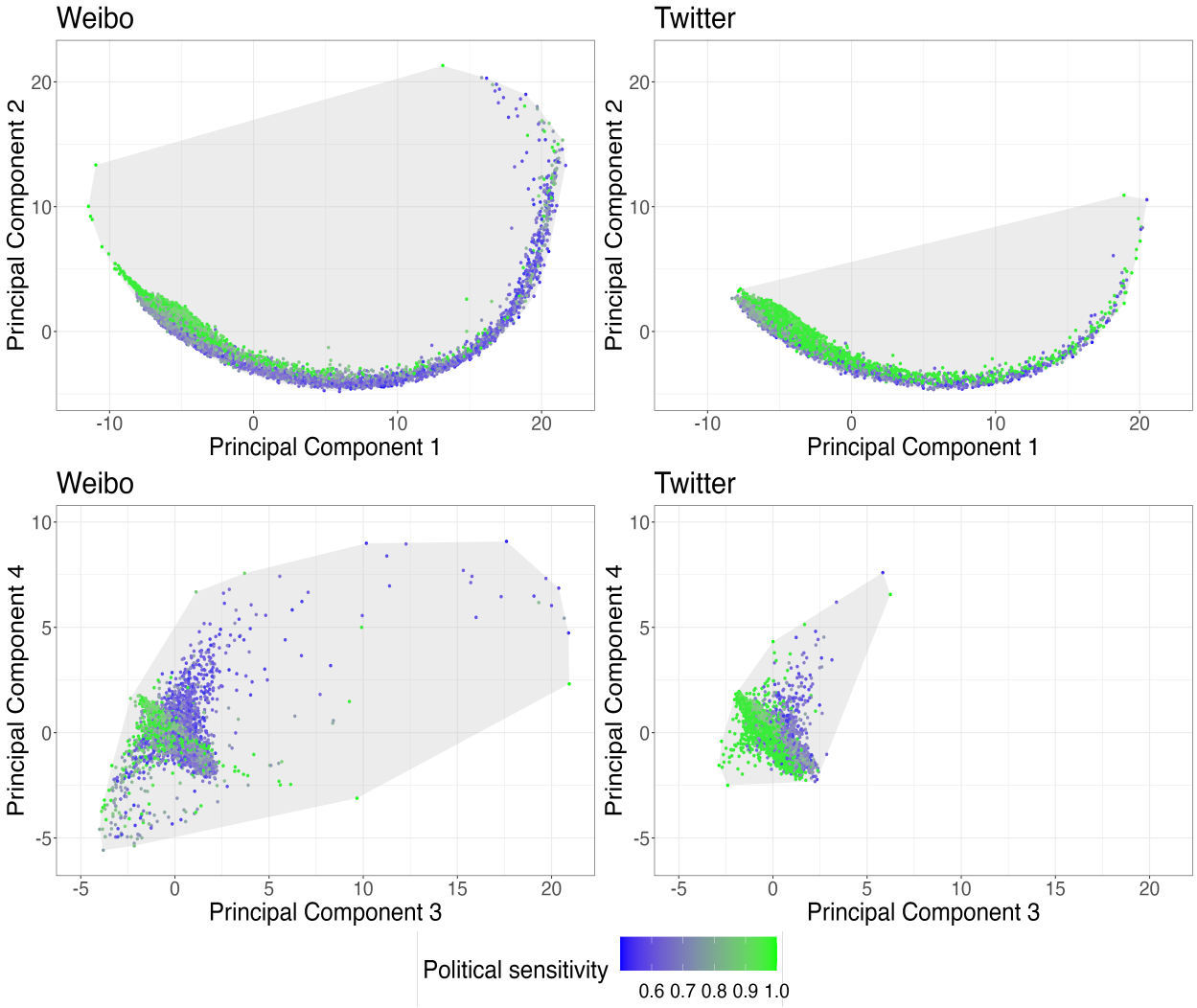
On the other hand, Twitter data augmentation improves the accuracy of censorship AI models that suffer from missing data during crisis. The right panel of Figure 6 shows that, when there is missing data (thresholds $0.9 - 0.6$), the accuracy of the models trained on the augmented datasets is higher than 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 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 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 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 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.

FIGURE 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.

Together, Figure 6 and Figure 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 6 and Figure 7 show the limits to which such augmentation techniques can boost censorship AI's performance.

6. 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,

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

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. 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.

References

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.

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

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.

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.

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

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* .

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.

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* .

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.

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.

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.

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.

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

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.

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.

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

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.

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* .

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.

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.

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

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.

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

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.

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

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*.

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

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.

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.

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

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* .

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.

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

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.

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* .

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.

Appendix *for*

The Digital Dictator's Dilemma

Eddie Yang
(UCSD)

Table of Contents

A. Further Details on Censorship AI

A.1. Model Details

Except for the result on alternative model architecture in Section C.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).[A1] 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.

A.2. Training Details

The BERT models are fine-tuned using the **transformers** library provided by Hugging Face.[A2] 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 (no. of true positives)/(no. of true positives + no. of false positives) and recall is given by (no. of true positives)/(no. of true positives + no. of false negatives).

Early stopping was used to prevent over-fitting. Specifically, training was stopped if the

[A1] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[A2] https://huggingface.co/docs/transformers/index

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.3. Hyperparameters

Table A1 reports the hyperparameter values used in training.

TABLE A1. HYPERPARAMETERS

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

A.4. Hardware

Models in the paper were fine-tuned using 8 NVIDIA A100 GPUs with 40GB of memory each.

B. Further Details on the Weibo and the Twitter Data

B.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 A2 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 A2. 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

B.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.

B.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.

B.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.[A3] The Peacock Chinese Twitter Corpus (PCTC) is another dataset of 4.9 millionn Chinese tweets.[A4] 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

[A3] 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/

[A4] 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.[A5] 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.[A6]

[A5] 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/

[A6] https://static.istories.media/uploaded/documents/0b809ea16feb42c7b8c91b022e45bd6b.pdf

C. Additional Results

C.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

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

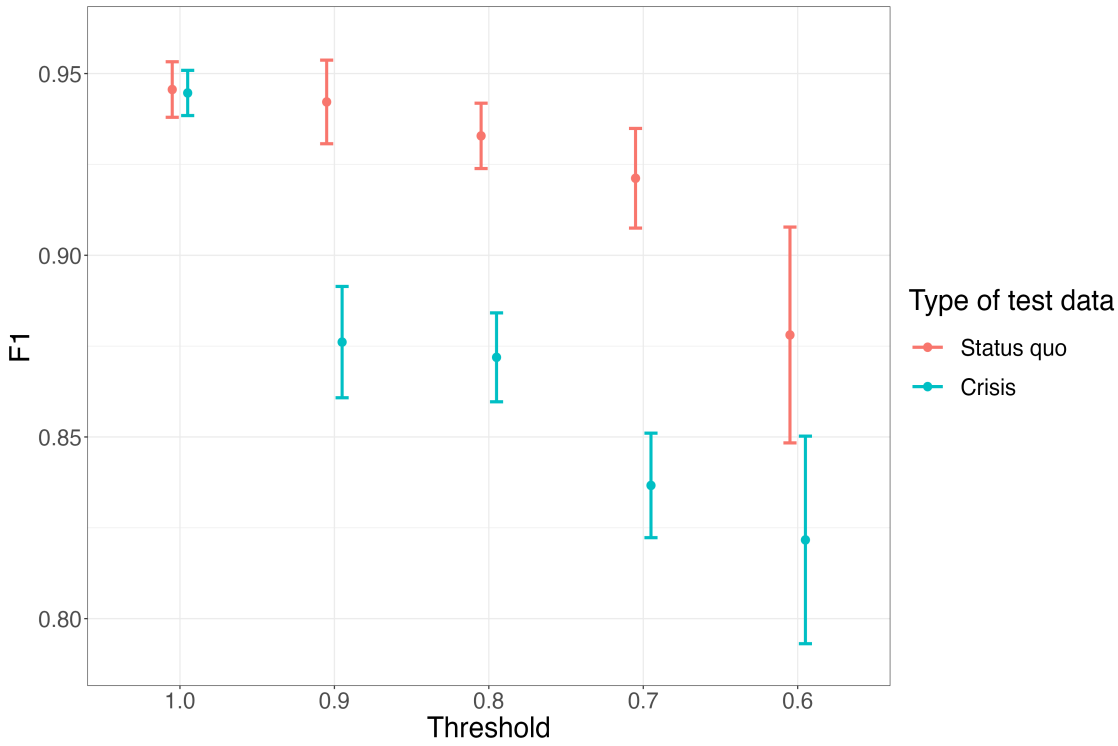
where precision is given by (no. of true positives)/(no. of true positives + no. of false positives) and recall is given by (no. of true positives)/(no. of true positives + no. of false negatives).

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 A1 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 4, as the increasing false negative rate will pull down the F1 score.

FIGURE A1. MODEL PERFORMANCE ACROSS TRAINING DATASETS

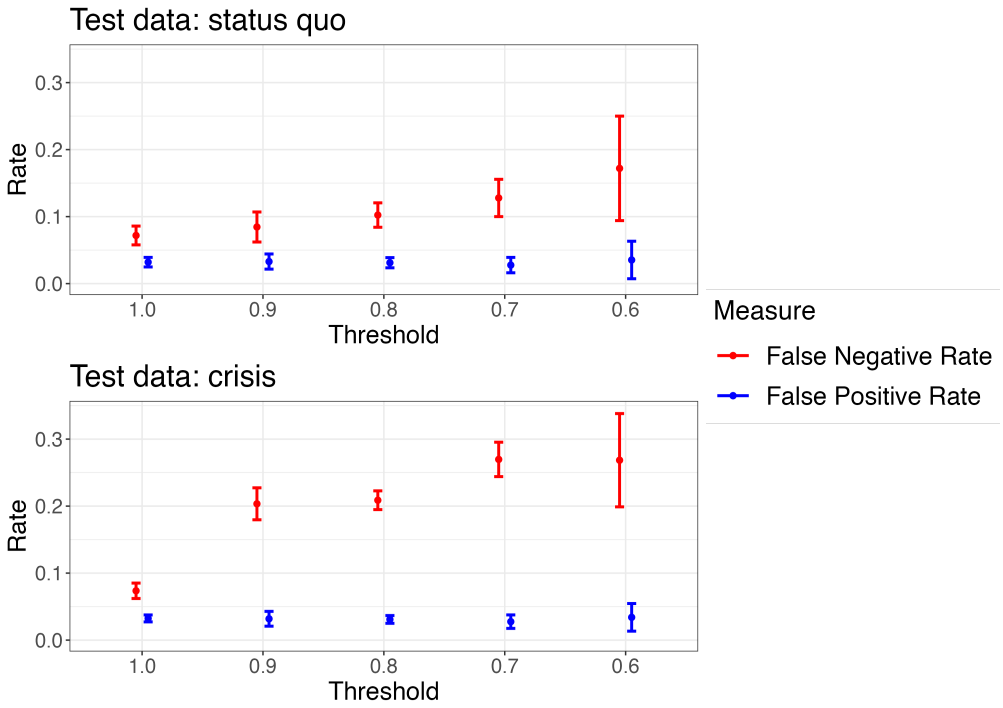


Note: Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 20 models for each threshold.

C.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 4). Figure A2 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 A2. FALSE POSITIVE RATE VS. FALSE NEGATIVE RATE



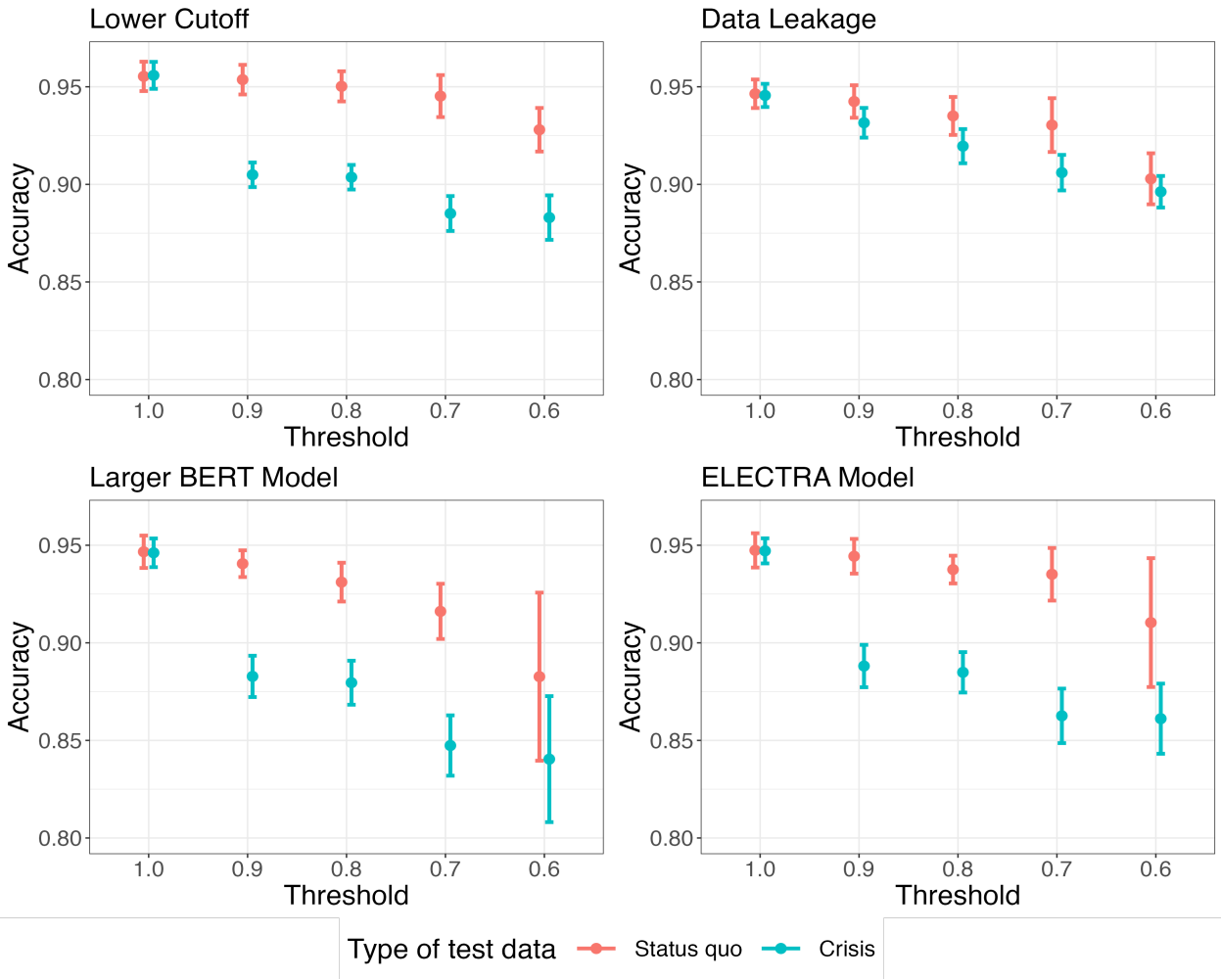
C.3. Lower Cutoff, Data Leakage, Larger Model, and Alternative Model

Figure A3 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[A7] (Clark et al., 2020) instead of BERT.

[A7] https://huggingface.co/hfl/chinese-electra-base-discriminator

FIGURE A3. 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.