# The Limits of AI for Authoritarian Control

Eddie Yang*

December 2024

**Abstract**

An emerging literature suggests Artificial Intelligence (AI) can greatly enhance autocrats' repressive capabilities and strengthen their 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 an "authoritarian data problem," in which citizens' strategic behavior under repression diminishes the amount of useful information in the data for training AI. This poses a fundamental limitation for AI's usefulness in 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 censorship data in China. I show that AI's accuracy in censorship decreases with increasing repression, especially during times of political crisis. 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, Purdue University. A previous version of this paper was circulated under the title "The Digital Dictator's Dilemma."

# 1. Introduction

Digital technologies, particularly Artificial Intelligence (AI), have been argued as a powerful addition to the autocrat's repressive toolkit (Diamond, 2019). Facial recognition technologies used in digital surveillance systems enable autocrats to more selectively target opponents of the state (Xu, 2021) while deep learning models for natural language processing enhance information control through automated censorship (Roberts, 2020; Gohdes, 2024) and (mis)information campaigns (Kreps, McCain and Brundage, 2022). The repressive potential of AI further benefits from the massive amounts of data collected by existing authoritarian institutions, which can be leveraged for AI model training and development (Beraja, Yang and Yuchtman, 2023; Beraja et al., 2023).

Yet we know little about the limits of AI for authoritarian control, if there are any. This paper provides both a theory and empirical evidence that authoritarian institutions limit AI's repressive capabilities, making it less omnipotent than scholars have previously argued. A key driver for these limits is the inherent tension between AI and authoritarian control: to effectively enforce control and repression, AI requires enough politically relevant information in its training data but institutions of control and repression by nature restrict both the quantity and quality of such information.[1]

AI relies on data to acquire its predictive capabilities. To effectively enforce authoritarian control, AI needs large amounts of politically relevant information in the training data. For instance, an automated censorship system will only be accurate at filtering censorable content if it has seen many examples of such content during its training. Under the shadow of repression, however, citizens' strategic behavior can tarnish AI's training data. When people self-censor (Shen and Truex, 2021) and falsify preferences (Kuran, 1997) to hide their critical views of the regime, the amount of censorable content is reduced. Other forms of strategic

[1]See also Farrell, Newman and Wallace (2022), Wallace (2022), and Trinh (2023) for similar arguments on data quality issues in the authoritarian context.

behavior (e.g., using coded language to circumvent censorship or wearing masks to evade surveillance) can also degrade the collected data. Therefore, the higher the cost of dissent because of repression, the more people will manipulate their public behavior, corrupting the training data and causing AI to be less effective and accurate at carrying out authoritarian control.

The authoritarian data problem can have a more negative effect on the performance of AI during political crises, such as protests and revolutions. This is because the sudden change in people's behavior during crisis, especially behavior that was previously suppressed for fear of repression, causes the data that AI has to predict on to be very different from its training data. For example, without additional training, AI would have never recognized that people holding a blank piece of paper on the street was a protest against China's COVID-19 restrictions. The shift in data distribution in crises is particularly bad for autocrats: they need AI to perform the best during times of political turmoil but it is exactly in such times that AI fumbles in performance.

While autocrats can collect more data to further train their repressive AI systems, the sheer volume of data is unlikely to solve the AI performance problem - as long as people maintain strategic behavior, the additional data will suffer from the same quality issues that cause AI's underperformance. However, data from democracies – generated largely without the same political constraints as in the authoritarian context – can potentially mitigate the authoritarian data problem.

To empirically test the theory, I focus on the use of AI for censorship as a case study. Specifically, I use a novel experiment to recreate commercial censorship AI systems and test their performance given different political conditions. I use millions of social media posts from the Chinese social media platform, Weibo, and Chinese tweets from Twitter as

the training data.[2] Notably, I exploit a rare opportunity to use an automated censorship service from a technology company in China to obtain the political sensitivity of each post. The political sensitivity scores allow me to model self-censorship and preference falsification by creating missingness in the training data. For example, to model a highly repressive environment where there is a high degree of preference falsification and self-censorship, the training data will have few social media posts with high political sensitivity scores. The set-up also allows me to test AI's performance during crises as well as the effect of data from international sources (Twitter) on AI's censorship accuracy.

The experiment establishes three sets of results. One, as the regime becomes more repressive and there is more preference falsification and self-censorship, the resulting information loss in the data causes a drop in the accuracy of AI in classifying which social media posts should be censored. Two, the drop in AI's accuracy is substantially larger during crises than normal times, with the majority of the misclassifications being false negatives (censorable posts misclassified as "safe"). Three, doubling the amount of data from Weibo (domestic data source) has a marginal effect on the accuracy of censorship AI, while data from Twitter (international data source) substantially increases accuracy, despite being a fraction of the Weibo data in size. The improvement from the Twitter data, however, does not fully close the accuracy gap caused by the authoritarian data problem. Through text analysis of the Weibo and Twitter data, I give suggestive evidence that the insufficiency of the Twitter data is likely due to its differences in discourse from the domestic Weibo data. This is consistent with a recent study that shows Venezuelan activists changed their discourse once they went into exile (Esberg and Siegel, 2023).

Taken together, the theory and empirical results highlight that the classic strategic behavior by citizens in authoritarian regimes (Kuran, 1997; Wintrobe, 2000; Jiang and Yang,

[2]The social media posts were about COVID-19. I focus on posts from the early period of the pandemic because this was when censorship of COVID-19 topics had not caught up in China. See the Data and Research Design section for more details.

2016; Roberts, 2018) now manifest in new forms to hamper digital dictators who wish to use AI for authoritarian control. While the literature on technology and autocracy has shown that AI can be useful for autocrats, the paper highlights the (understudied) limits of AI for authoritarian control. It does so by challenging an implicit assumption of the existing literature - that more data means more accurate predictions from AI (Feldstein, 2019). 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 world 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. 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, 2023) in highlighting the difference in content between domestic and international sources of information.

## 2. Background: AI and Autocracy

In this paper, AI refers to computer programs that are capable of performing tasks that typically require human intelligence. Examples of AI performing "intelligent" tasks include playing chess, recognizing faces in surveillance videos, and classifying whether a social media post should be censored. 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 the political nature of some content.

A key underlying technology that powers AI is deep learning - algorithms that are capable of extracting complex relationships from data. Typically, training deep learning models follows a two-stage process: 1) a pre-training stage where models are trained on diverse datasets 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, one 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 carry out censorship. In the paper, I focus on data and training in the second stage as fine-tuning has a large impact on AI's performance on specific tasks.

In the fine-tuning stage, just like OLS regression, deep learning models take as input some features $X$ with their corresponding outcomes or labels $Y$ and fit a function $Y = f(X)$. Unlike OLS regression, deep learning models 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 models are usually much more complex, involving upward of billions of parameters.

A key contributing factor to AI's recent success is the availability of large amounts of

high-quality data. Such data enables deep learning models 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 like ChatGPT are trained on trillions of words scraped from books and the internet. These models are transforming the modern way of life. Students now rely on AI chatbots for answers to questions and assignments and people increasingly use self-driving technologies to assist with their driving.

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.[3] India has used facial recognition systems to verify voter identity in elections. Perhaps even more so than democracies, authoritarian regimes have embraced AI to automate tasks like surveillance (Xu, 2023), censorship, meting out criminal sentences in place of judges (Yang, 2023), and other repressive tasks (Kendall-Taylor, Frantz and Wright, 2020). Yet despite AI's growing importance in politics, evaluations of its impact are rare in political science, with a few notable exceptions (Xu, 2021; Allie, 2023; Imai et al., 2023).

## 3. Theory: Why Authoritarian Politics Constrain AI

AI derives its capabilities from the data it is trained on. When the training data is problematic (biased, low-quality etc.), the output of AI often becomes subpar. A well-known example of bad data causing issues in AI is the case of racial bias in facial recognition models. Facial recognition systems tend to mis-recognize faces with darker skin tones at a much higher rate (Cook et al., 2019; Vangara et al., 2019; Robinson et al., 2020). The disparity in error rate is attributed to racial imbalance in the training samples, with a lack of racial minority, especially African-American, faces in the data (Buolamwini and Gebru,

[3]Mapping Pretrial Injustice, "Where are Risk Assessments Being Used?" https://pretrialrisk.com/national-landscape/where-are-prai-being-used/

2018; Leslie, 2020). Subsequently, much effort has focused on increasing the training data quality through more racially balanced samples (Kärkkäinen and Joo, 2019; Wang, Zhang and Deng, 2021).

## 3.1. Data Deficiencies in Authoritarian Regimes

Given the importance of data in dictating the performance of AI, both practitioners and scholars have claimed that authoritarian regimes may have an advantage in developing AI systems, due to their ability to collect large amounts of data on their citizens (Lee, 2018; Feldstein, 2019). In areas with little strategic behavior, this is largely true. In healthcare, for example, China is leading the global market in medical AI, thanks to its readily available data from public hospitals.[4]

In politics, however, the same argument can fall apart when people have an incentive to act strategically. Such strategic behavior - residents avoiding surveillance cameras, bureaucrats misreporting local statistics, citizens self-censoring anti-regime sentiments - reduces the quantity and quality of the relevant data for repressive tasks such as surveillance and censorship. In fact, a slew of problems in authoritarian regimes has been attributed to bad or missing data, such as inefficient governance (Wallace, 2022; Trinh, 2023) and surprise breakdowns 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.

In the context of censorship (setting for the empirical section), two mechanisms serve to degrade data. One, citizens can falsify their public preferences under the perceived threat of punishment (Kuran, 1997). While citizens may hold grudges against the autocrat, the

[4]Sinolytics, "Why China has an advantage in medical AI," Table.Media, May 14, 2024, https://table.media/en/china/sinolytics-radar/why-china-has-an-advantage-in-ai-in-medicine/

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 creates a mismatch between the distribution of citizens' private preferences and the public data that the autocrat collects. 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 (Jiang and Yang, 2016; Robinson and Tannenberg, 2019; Hale, 2022; Nicholson and Huang, 2023).

The second, related mechanism that negatively affects data 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.

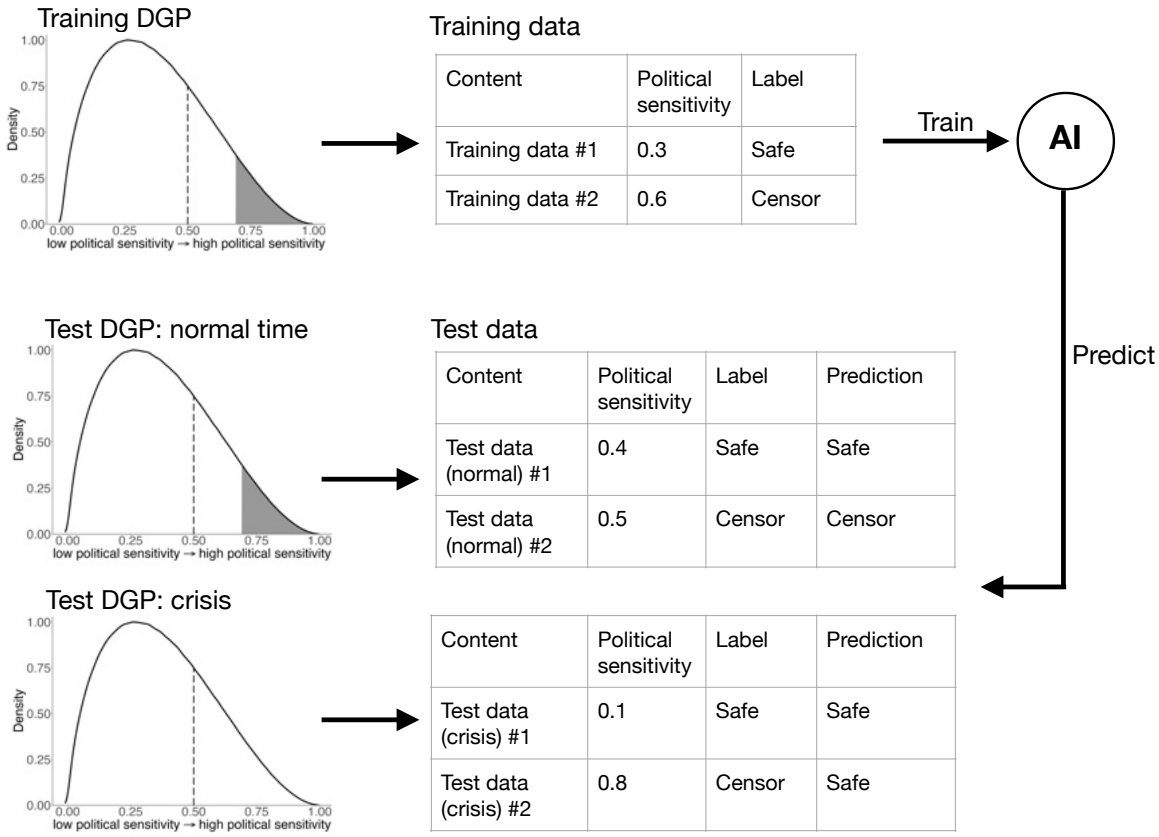
Both preference falsification and self-censorship corrupt data on political preferences and attitudes by creating a missing data problem in which politically valuable but sensitive information is missing in the observed data. The severity of the missing data problem is a function of the cost of voicing dissent - the higher the cost, the less such information will be in the data (Tannenberg, 2022).

## 3.2. Automating Autocracy with Bad Data

AI suffers two related problems from strategic behavior in the data generating process: strategic behavior 1) reduces sensitive but valuable information in the training data, and 2) increases the difficulty of the prediction task. To see why this is the case, consider the stylized example of censorship AI in Figure 1. In this example, AI's training data consists of content that should and should not be censored. Both kinds of content are generated from the unobserved distribution of political sensitivity. As an example, content with sensitivity scores above 0.5 are treated as censorable (e.g., anti-regime content) and those with scores

below 0.5 are treated as "safe". Because of preference falsification and self-censorship, the distribution of observed content is right-censored, in that content with high sensitivity scores (shaded region) will be missing from the observed data. 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 missing, it reduces the overall distance (and increases the similarity) between the observed censorable and safe content. As content becomes more homogenous, it becomes more difficult to distinguish what should be censored and what should not (Problem 2). Both of these problems become more severe as the level of repression and censorship increases and people respond with more preference falsification and self-censorship (the shaded region becomes larger).

FIGURE 1. STLYZED EXAMPLE OF CENSORSHIP AI TRAINING AND TESTING




*Notes:* The shaded regions in the data generating processes (DGP) of the training data and the normal time test data indicate right-censoring. This causes content with high sensitivity to be missing in the observed 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.

In addition, the authoritarian data problem will cause a larger drop in the performance of AI during periods of political crisis than during normal times. Normal time refers to "business as usual" in authoritarian regimes, when the level of preference falsification and self-censorship is maintained. During normal times, the data that AI has to predict on is drawn from the same distribution as the training data (Figure 1). On the other hand, crisis refers to times of political turmoil, such as protests and coups, when there is an information cascade and citizens reduce or even do away with preference falsification and self-censorship (Lohmann, 1994). When there is no preference falsification and self-censorship, the test data is drawn from the full distribution without missing data. This mismatch between the distributions of AI's training data and the test data during crises further pulls down the performance of AI, as much content that is present in crises is suppressed during normal times and thus not encountered by AI during its training.[5] This can be particularly bad for autocrats: autocrats need AI to perform the best during crises but it is exactly in times of crisis that AI fumbles in performance.

### 3.3. Irony of the Free World

What can autocrats do in light of the authoritarian constraints on AI? One 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 strategic behavior. In other words, 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

[5]The computer science literature sometimes refers to the same phenomenon as distribution shift. See e.g., Storkey et al. (2009).

reflects citizens' withheld 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, content from diaspora communities on international social media platforms such as Twitter and Facebook may contain valuable information that is suppressed domestically. A censorship AI that is trained on domestic data augmented by international data may thus be more accurate in censorship than the AI trained on domestic data alone. For autocrats, collecting data from international sources not only boosts the performance of AI but also has the advantage of keeping the level of repression and censorship unchanged domestically. Qualitative evidence suggests that such practice is already used systematically on a large scale and across authoritarian regimes.[6]

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 international sources of information are qualitatively different from domestic sources, both in terms of topical distribution as well as political sensitivity (Esberg and Siegel, 2023). Such differences will limit the effect of data augmentation on AI performance.

## 3.4. Summary

In summary, I leverage theories of citizens' strategic behavior in authoritarian regimes to explain the (under-)performance of AI for authoritarian control. Specifically, the theory implies the following two sets of hypotheses.

**Repression-performance trade-off:**

[6]See Appendix B.5 for a more detailed discussion.

1a. As repression and censorship increase and people engage in more strategic behavior, the performance of AI on authoritarian control will decrease.

1b. The drop in AI's performance is larger during times of political crisis than during normal times.

**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's performance.

# 4. Data and Research Design

I chose 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.[7]

## 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 technology is 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 an AI experiment to recreate as closely as possible the actual training of censorship AI models in practice. Specifically, the experiment uses 1)

[7]Similar set-up has been widely used in training AI for content moderation. See e.g., Google's Perspective API: https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US.

the same AI algorithm that many technology companies use, 2) training data that consists of millions of real-world user-generated content, and 3) training procedures using state-of-the-art computing hardware. Using a unique dataset of social media posts for which political sensitivity is known, the experiment compares the accuracy of censorship AI models trained on data with varying degrees of strategic behavior.

To construct the domestic training data, I first 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 data and use them to train censorship AI models specifically for COVID-19.

To get the political sensitivity of the social media posts, 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 that ranges from 0 to 1 for each post, with 1 being the most sensitive. The political sensitivity scores serve as the latent variable. I use the service's default sensitivity score of 0.5 as the threshold to generate the binary censorship labels - social media posts with scores above 0.5 have a label of 1 (censor) and posts with scores below 0.5 have a label of 0 (not censor). 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

[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. https://ai.baidu.com/support/news?action=detail&id=1819.

[9] See Appendix B.3 for more details on the service and its usage by social media companies.

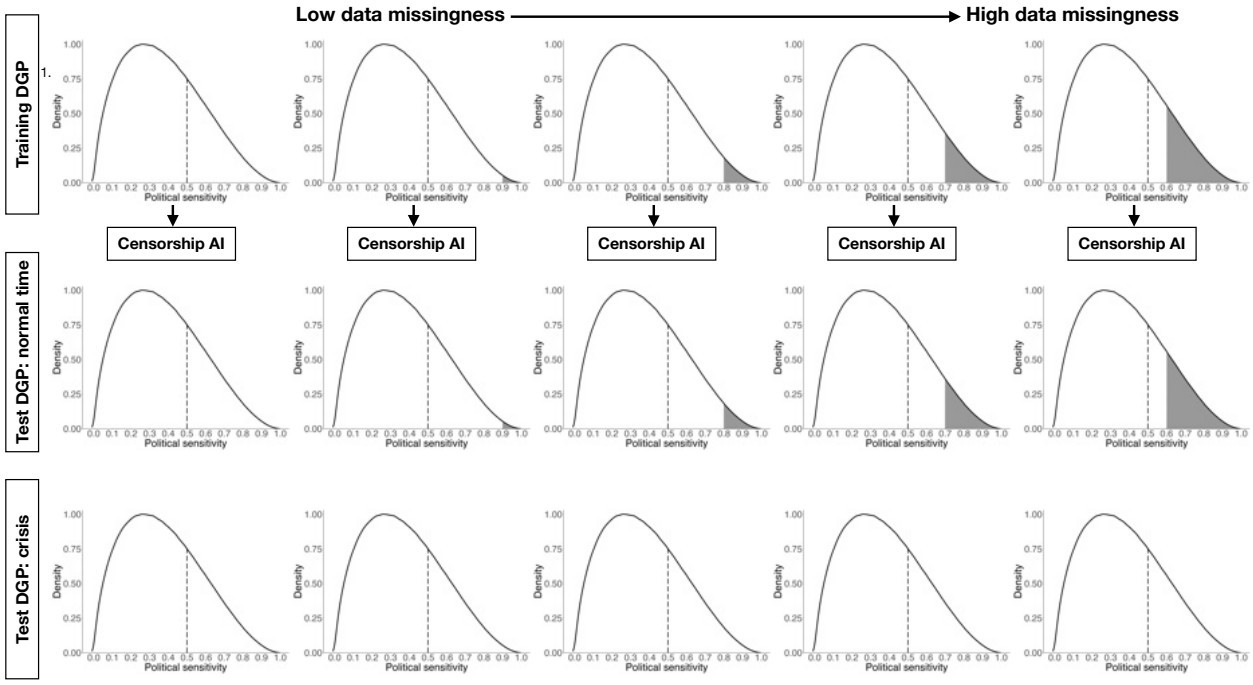
the proportion of the two classes (0 and 1) of labels. Therefore, the main sample has a size of 1 million social media posts.

To model different degrees of data missingness due to preference falsification and self-censorship, I use the main sample to construct training datasets that differ in their distribution of political sensitivity. As the top row of Figure 2 shows, I construct five versions of training dataset with varying degrees of missingness. To model the case where there is no missingness due to strategic behavior, 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. 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.

The design assumes that citizens have perfect information about what content is censorable and respond accordingly given the level of repression. In the appendix, I account for imperfect information and coordination by allowing five percent of the data from the missing part of the distribution to leak into the training datasets. The substantive conclusions remain unchanged.

For each version of the training dataset, I train a separate censorship AI model on it. Specifically, I use the 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 110 million parameters and was developed by Google. Since its introduction, BERT has been one of the most popular deep learning models for prediction and is widely used in commercial applications. In the Appendix, I provide details about how the model is used in practice for censorship based on fieldwork in technology companies. 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 25 models with the training data shuffled each time. This allows me to obtain

FIGURE 2. EXPERIMENTAL DESIGN




*Notes:* Graphical representation of the research design. There are five versions of training dataset corresponding to different degrees of strategic behavior. The "normal time" 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.

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 kinds of test data: normal time and crisis (second and third rows of Figure 2). Social media posts in the normal time test data are drawn with the same level of missingness as the corresponding version of the training dataset whereas the crisis test data is always drawn from the full distribution. Mirroring the theory, this is to model the situation in which citizens maintain their level of preference falsification and self-censorship during normal times but start revealing their true preferences during crises. Both sets of test data are sampled from social media posts that are not in the training data. To be able to compare performance evaluated on different test data, each test dataset is a balanced

sample of 1000 positive labels (i.e. censor) and 1000 negative labels (i.e. not censor). To measure the performance of censorship AI, I use accuracy, defined as the fraction of correct predictions over the total number of predictions, in the main text and report other measures of performance in the Appendix. Here, a correct prediction means that the model predicts a censorship label that is the same as the ground truth label.

### 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 from international sources (hypothesis 2b), I scraped all 558,322 Chinese tweets from Twitter that are on the same COVID-19 topics and were 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 219,111 tweets with political sensitivity scores above 0.5 and use it to augment the Weibo training datasets.[10] Notably, the same Twitter dataset is used to augment all training datasets, 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. I then show evidence that adding more domestic data during training has a marginal impact on AI's performance

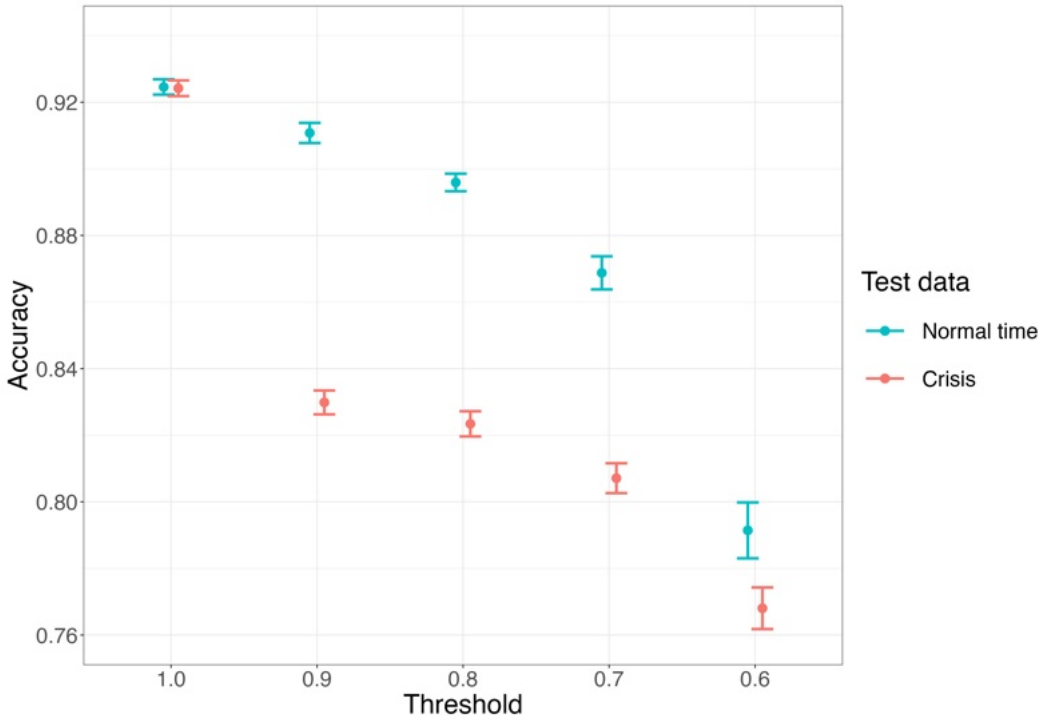
[10] I exclude tweets with labels of 0 from the Twitter dataset as missingness in the Weibo datasets only comes from social media posts with positive labels.

but data augmentation from international sources results in a larger improvement in AI's censorship accuracy.

## 5.1. More Repression, Worse AI 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 data missingness. The threshold (x-axis) indicates the political sensitivity score above which data is missing from the training dataset. The threshold of 1.0 means the training dataset has no missing data and the threshold of 0.6 has the most missing data. Model accuracy is evaluated on both the normal time and crisis test data.

FIGURE 3. MODEL PERFORMANCE ACROSS TRAINING DATASETS




*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

Evaluations on the normal time data (blue) show that as data missingness increases as a result of strategic behavior, 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). In line with the theory, the drop in model accuracy is significantly larger in crises, when people reveal their true preferences, than in normal times. In the appendix, I show that the drop in model accuracy is largely due to an increase in false negatives (censorable content predicted to be "safe"). This is a particularly bad situation for autocrats as false negatives allow transmission of politically sensitive information among citizens and thus may be more costly for autocrats than false positives (censoring more than they should).

Table 1 provides two examples of social media posts and the corresponding predictions from different models. The first example expresses disappointment at the government's handling of the pandemic and the second example mocks a research lab for promoting traditional Chinese medicine. In both cases, the censorship model trained on data with no missing data (threshold = 1.0) correctly predicts their censorship labels while the model trained with missing data (threshold = 0.6) gets both wrong.

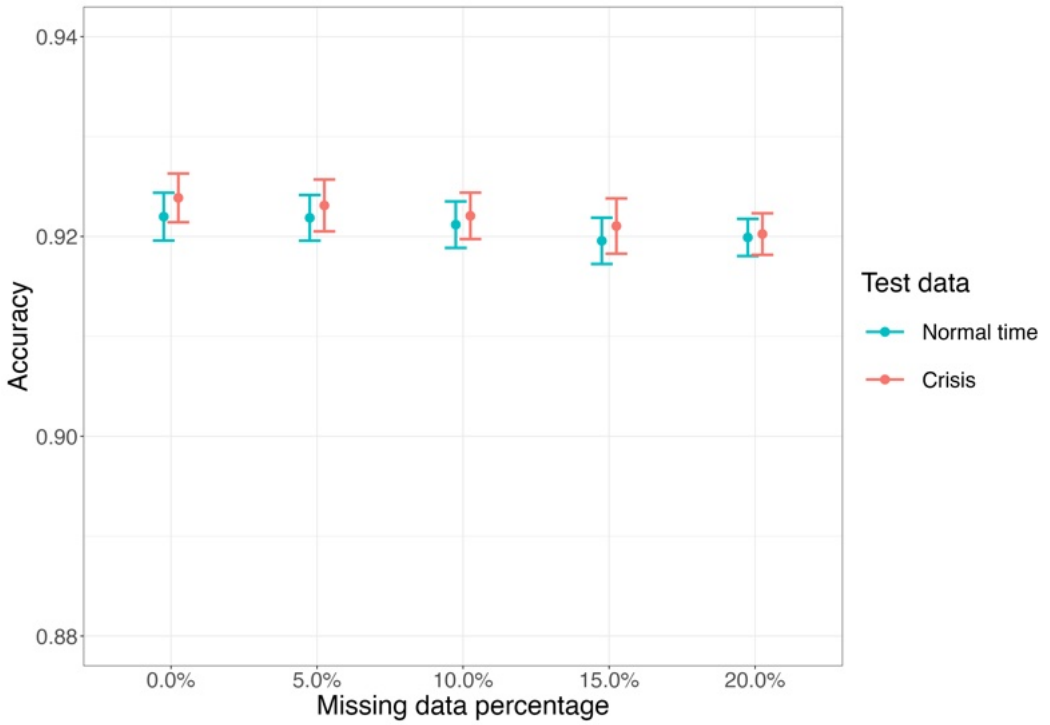
Table 1. Example Social Media Posts and Censorship Predictions

| Social media post (translated) | Ground truth | Prediction (thres.=1.0) | Prediction (thres.=0.6) |
|---|---|---|---|
| Originally, I had a lot of confidence in how the pandemic was being handled since the central government took over, but now all these developments are really disappointing. [Sad emoji] | censor | censor | not censor |
| The Wuhan Institute of Virology, a world-class P4 biosafety lab, believes that Shuanghuanglian oral liquid can inhibit the virus. Now is truly the pinnacle of traditional Chinese medicine history. [Sarcastic emoji] | censor | censor | not censor |

Critically, the decrease in model accuracy and the performance gap between normal time and crisis test data, as observed in Figure 3, depend on the strategic nature of citizen behavior. The same patterns would not be observed if people withhold content in a way that is unrelated to political sensitivity. Figure 4 shows one such scenario in which data is

missing at random. The x-axis indicates the percentage of missing data as a fraction of the original 1 million sample. Despite having missing data at the same scale as the set-up in Figure 3, Figure 4 shows no significant performance difference across models and between normal times and crises.

FIGURE 4. MODEL PERFORMANCE WITH NON-STRATEGIC MISSING DATA




*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

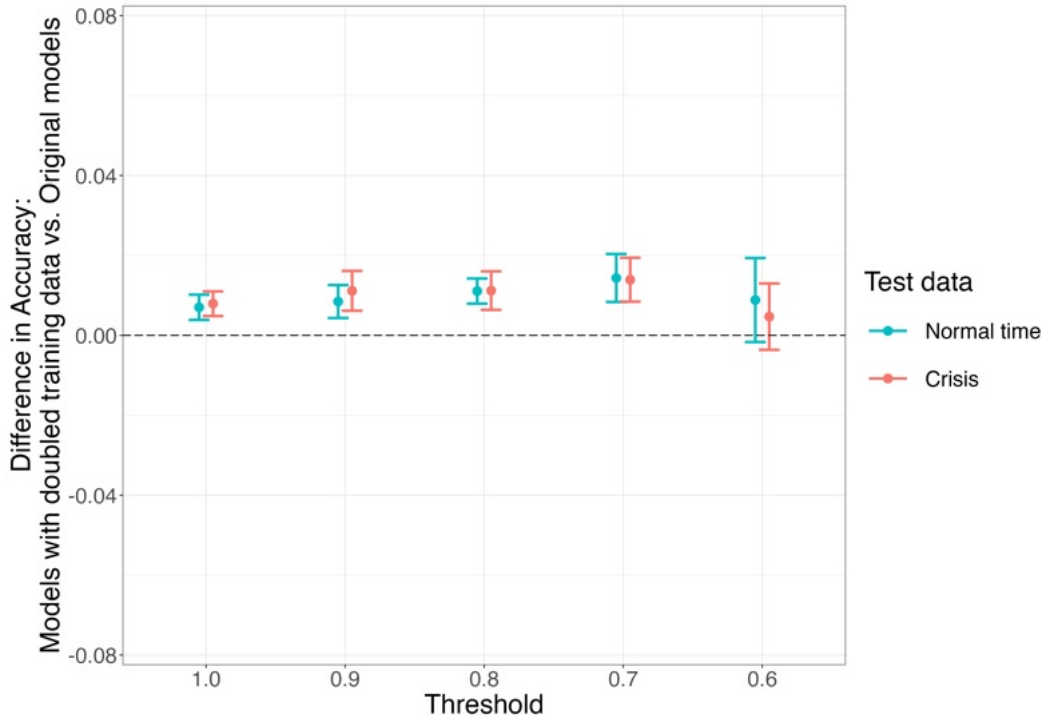
In the Appendix, I provide additional evidence that the substantive conclusions are robust to various changes to the experimental set-up, such as using a larger pre-trained model, changing the censorship decision rule (e.g., from 0.5 to 0.4), allowing some leakage of the missing data into the training data, adding various performance enhancing techniques, and using a different deep learning model architecture.

## 5.2. Marginal Impact of More Domestic Training Data

Given the previous results, one of the ways autocrats may choose to respond to the data problem 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 double the amount of training data and those trained with the original training datasets. Across settings, the improvement in model accuracy from doubling the size of the training data is marginal - the largest accuracy improvement is smaller than two percentage points.

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.

FIGURE 5. EFFECT OF DOUBLING DOMESTIC TRAINING DATA



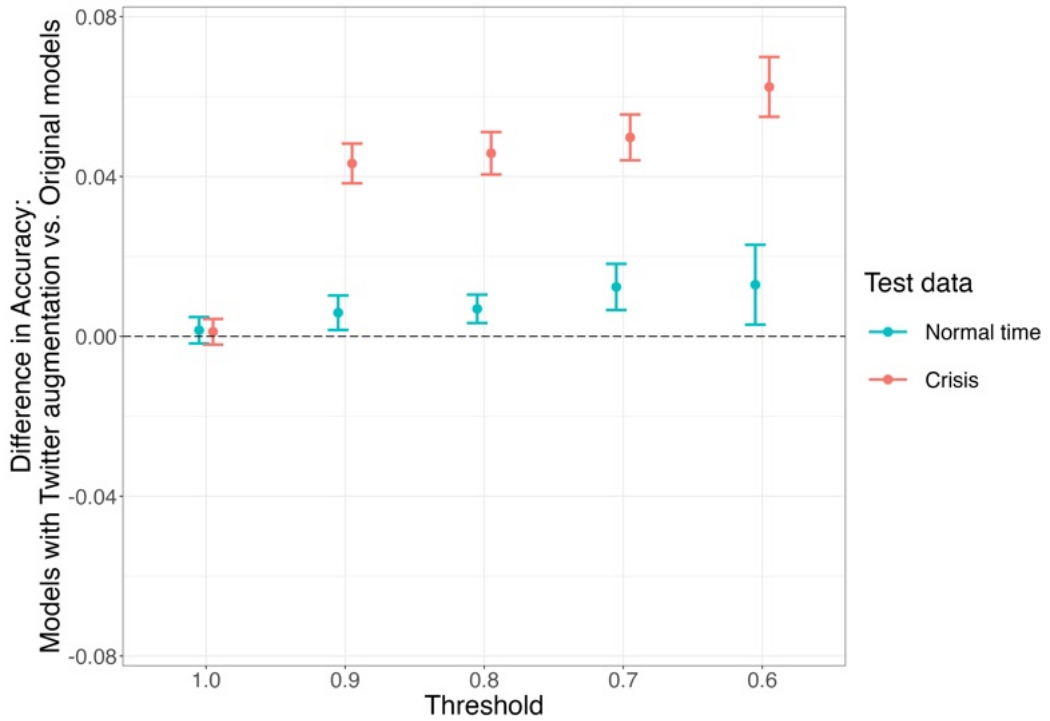

*Notes:* Y-axis shows the average difference in accuracy between models trained on the original Weibo data and models trained on the larger (doubled in size) Weibo data. Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

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.

## 5.3. Accuracy Improvement from International Data

While additional data collected domestically provides little improvement to model accuracy, data from international sources, generated largely without the same political constraints, should boost model performance. Figure 6 provides evidence that augmenting the original Weibo training data with data from Twitter improves model accuracy, especially for performance during crises.

Figure 6. Effect of Twitter Data Augmentation




*Notes:* Y-axis shows the average difference in accuracy between models trained on the original Weibo data and models trained on the Twitter-augmented data. Each threshold value represents a version of the original Weibo training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

Figure 6 compares the accuracy of models trained on datasets augmented by the Twitter data with models trained on the original Weibo training datasets. Similar to domestic data augmentation, the Twitter data augmentation provides a marginal improvement on the normal time test data. This is because the normal time 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 "safe" 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 normal time data becomes more difficult as the threshold decreases, the Twitter data thus provides litter accuracy improvement.

On the other hand, Twitter data augmentation improves the accuracy of censorship AI models during crises. Figure 6 shows that, when there is missing data (thresholds $0.6 - 0.9$), the accuracy of the models trained on the augmented datasets is substantially higher than the models trained on the original Weibo data, with the largest improvement being more than six percentage points. This is despite the fact that the Twitter data is only about one-fifth of the Weibo data in size. Figure 6 thus shows that Twitter data can partially compensate for the missing data from Weibo and reduce the mismatch in distribution between the training data and crisis test data.

It is important to note, however, that the accuracy improvement from Twitter data is limited, in that the models' accuracy is still significantly 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 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 the topic prevalence across the two data sources. Topic prevalence is estimated with a structural topic model (Roberts et al.,

2014) using the combined Weibo and Twitter data[11], where the number of topics is set to 15. Topic labels are written manually based on the top keywords and the most representative posts for each topic.[12]

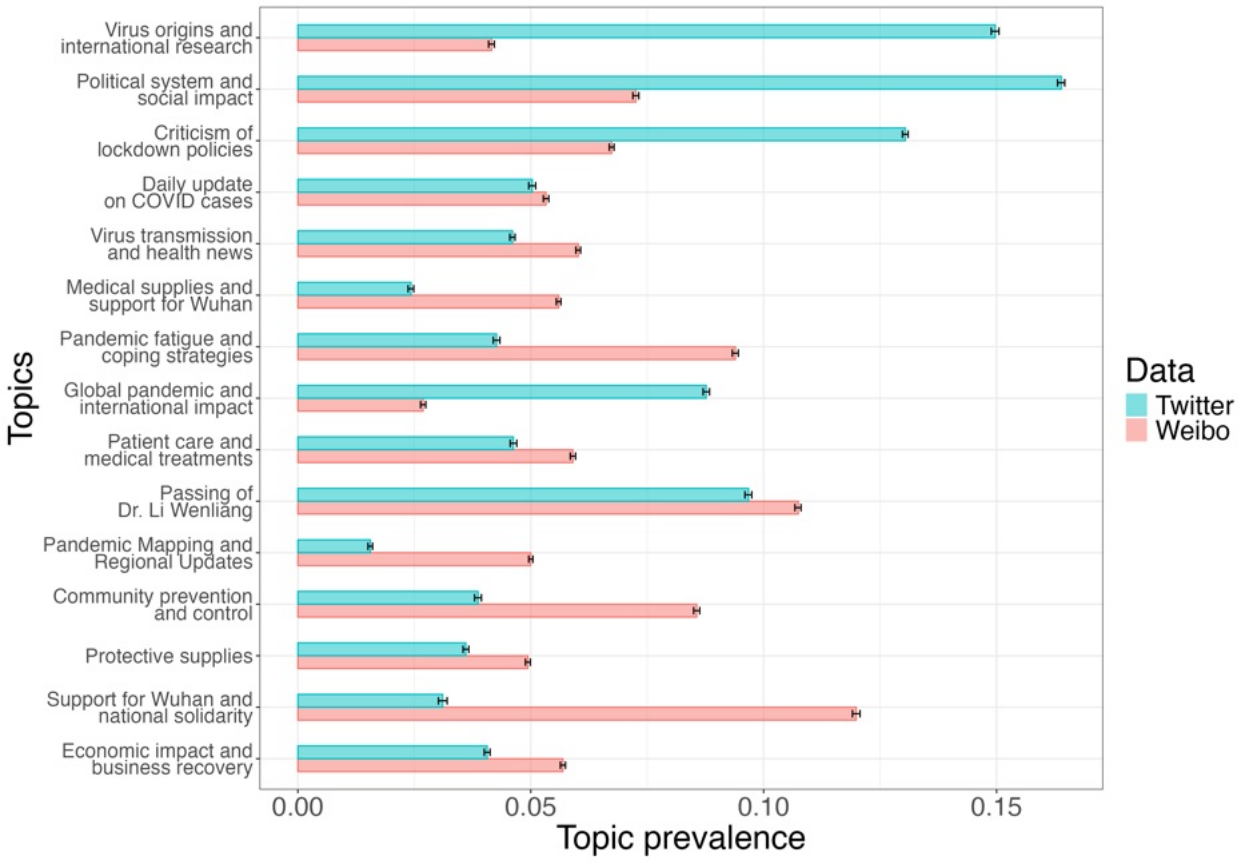
Figure 7 shows that politically sensitive topics (e.g., discussions of virus origins, political system, and criticism of lockdown policies) are substantially more prevalent in the Twitter data than in the Weibo data. For the topic on virus origins and international research, for example, it is more than three times as prevalent in the Twitter data than in the Weibo data. Additionally, international issues (topic on global pandemic and international impact) are also more prevalent in the Twitter data. In contrast, content on Weibo consists of more discussions of domestic issues such as daily updates on COVID-19 cases and local community prevention and control. Furthermore, content on Weibo includes more support (instead of criticism) for local and national COVID measures, with the topic on support for Wuhan and national solidarity being 8.8 percentage points more prevalent in the Weibo data than in the Twitter data. In Appendix E.2, I provide evidence that the difference in content propagates to censorship AI models, where the models' internal representation of the Twitter and Weibo data shows spatial differences between the two data sources.

Together, Figure 5 and Figure 6 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. Figure 6 also shows the limit of data augmentation from international sources in boosting censorship AI's performance and Figure 7 suggests that the difference in content between domestic and international sources likely contributes to the limit.

[11]For this analysis, both censorable and safe posts from Twitter and Weibo are included.

[12]See Appendix E.1 for the list of topic keywords.

Figure 7. Topic Prevalence across Weibo and Twitter




# 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 data and make predictions based on the data no doubt enables contemporary autocrats to sieve through information at a scale their counterparts 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 autocrats 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 empirical findings of the 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), studies have shown that repression can generate both chilling and backlash effects (Huang, 2018; Pan and Siegel, 2020).[13] 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 may 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 autocrats' 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

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

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, 2023; 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, 2023). 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 are worth exploring in 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. 2023. "AI-tocracy." *The Quarterly Journal of Economics* 138(3):1349–1402.

Beraja, Martin, David Y Yang and Noam Yuchtman. 2023. "Data-intensive innovation and the State: evidence from AI firms in China." *The Review of Economic Studies* 90(4):1701–1723.

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.

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.

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. 2023. "How exile shapes online opposition: Evidence from Venezuela." *American Political Science Review* 117(4):1361–1378.

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. "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. 2024. *Repression in the Digital Age: Surveillance, Censorship, and the Dynamics of State Violence*. Oxford University Press.
**URL:** *https://doi.org/10.1093/oso/9780197743577.001.0001*

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, D James Greiner, Ryan Halen and Sooahn Shin. 2023. "Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186(2):167–189.

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.

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.

Kreps, Sarah, R Miles McCain and Miles Brundage. 2022. "All the news that's fit to fabricate: AI-generated text as a tool of media misinformation." *Journal of experimental political science* 9(1):104–117.

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.

Leslie, David. 2020. "Understanding bias in facial recognition technologies." *arXiv preprint arXiv:2010.07023* .

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. 2023. "Making the list: reevaluating political trust and social desirability in China." *American Political Science Review* 117(3):1158–1165.

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.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American journal of political science* 58(4):1064–1082.

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.

Yang, Eddie and Margaret E Roberts. 2023. "The Authoritarian Data Problem." *Journal of Democracy* 34(4):141–150.

# Appendix *for*

# The Limits of AI for Authoritarian Control

Eddie Yang
(Purdue)

## Table of Contents

# A. Further Details on Censorship AI

## A.1. Model Details

Except for the results on larger model and alternative model architecture in Section D, the pre-trained BERT model used in the paper is the Chinese BERT (bert-base-chinese).[A1] The model has 110 million parameters and has been shown to perform well on a variety of Chinese prediction tasks.

Based on information gathered in fieldwork, the model and its variants have 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 ChatGPT, 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 data 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 F1 score on the validation set did not improve for two epochs. To speed up training,

[A1] https://huggingface.co/google-bert/bert-base-chinese

[A2] https://huggingface.co/docs/transformers/index

I used mixed precision training (fp16) for all models in the experiment. The full list of hyperparameter values is provided in Table A3

TABLE A1. HYPERPARAMETERS

| Hyperparameter | Value |
|---|---|
| maximum token length | 128 |
| fp16 | True |
| batch size | 512 |
| learning rate | 0.0001 |
| learning rate scheduler | cosine |
| early stopping patience | 3 |
| warmup steps | 500 |
| maximum training steps | 10,000 |
| optimizer | AdamW (fused) |

### A.3. Hardware

Models in the paper were fine-tuned using Nvidia A100 GPU with 80GB memory. The total GPU hours are around 1600.

## B. Further Details on the Weibo and Twitter Data

### B.1. 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.2. Details on the Twitter Data

Twitter data was collected using the Twitter research API with the following restrictions: 1) the tweets were posted between December 1, 2019 and February 27, 2020; 2) the tweets contain at least one of the 40 keywords in Fu and Zhu (2020); and 3) the language of the tweets 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 was anonymized to remove tags and other user information.

## B.3. Political Sensitivity Service

The political sensitivity service is provided by Baidu, a major Chinese technology company, and is publicly available.[A3] 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

| Keyword(s) | 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 |

[A3] https://ai.baidu.com/tech/textcensoring

## B.4. Details on the Training Datasets

From the combined Weibo data, a 2.3 million sample was labeled by the political sensitivity service. The size of the sample was primarily determined by resource constraints. From the 2.3 million labeled data, a 2 million sample was constructed where censorable posts were upsampled to reduce label imbalance (there are many more "safe" content than censorable content). The 2 million sample is then randomly split into two datasets of 1 million social media posts, one of which serves as the baseline Weibo data for fine-tuning models. Table A3 shows the summary statistics of different versions of the baseline training datasets.

The entirety of the Twitter data was labeled by the political sensitivity service. Twitter posts for which the sensitivity scores are above 0.5 are then used as the Twitter augmentation dataset.

Table A3. 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) | 183,511 | 119,139 | 92,872 | 64,142 | 36,738 |
| No. of negative labels (not censor) | 816,489 | 816,489 | 816,489 | 816,489 | 816,489 |

## B.5. 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.[A4] The Peacock Chinese Twitter Corpus (PCTC) is another dataset of 4.9 millionn Chinese tweets.[A5] 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 Tiktok and is in the process of using such data to develop automated censorship systems.[A6] 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.[A7]

# C. Additional Performance 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

[A4] 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/

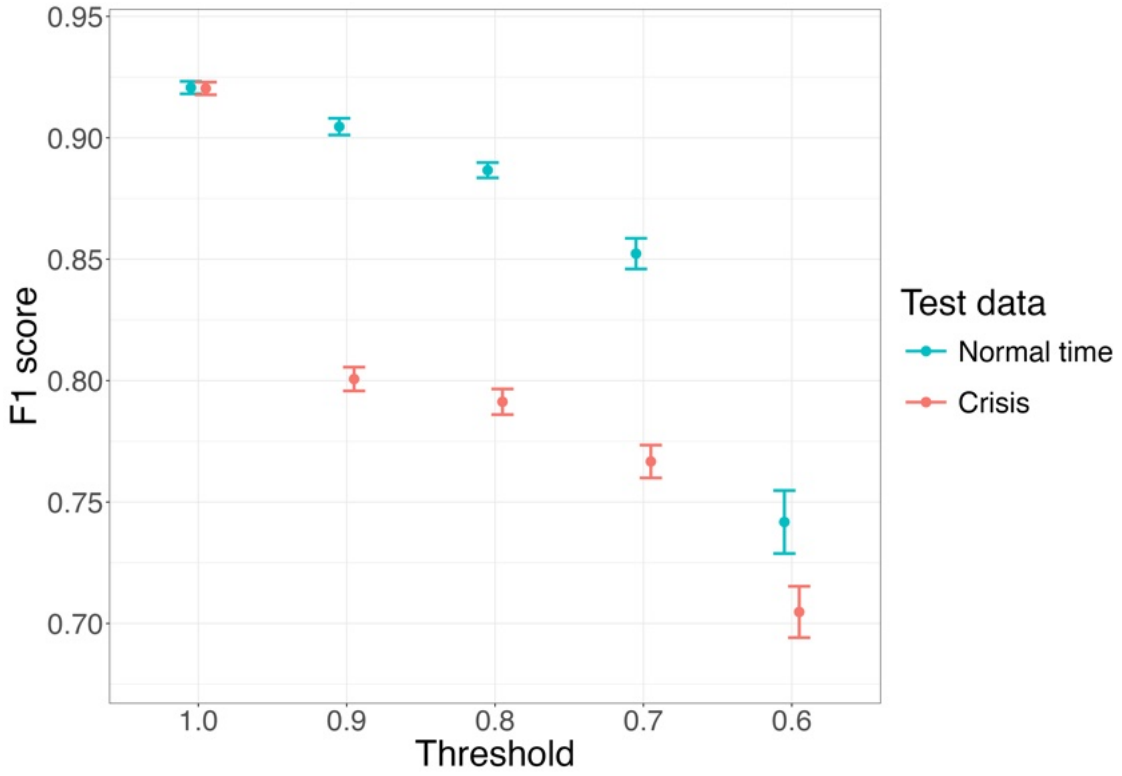
[A5] https://figshare.com/articles/dataset/Peacock_Chinese_Twitter_Corpus_PCTC_/13489239/1

[A6] See e.g., https://istories.media/stories/2023/02/08/vnutri-mashini-tsenzuri/

[A7] https://static.istories.media/uploaded/documents/0b809ea16feb42c7b8c91b022e45bd6b.pdf

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).

FIGURE A1. MODEL PERFORMANCE ACROSS TRAINING DATASETS




*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

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 normal time

test data.

## C.2. Error Rate Results

While accuracy/F1 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 may be more costly for autocrats than false positives (censoring more than they should).

Figure A2. Error Rate by Error Type - Original Models

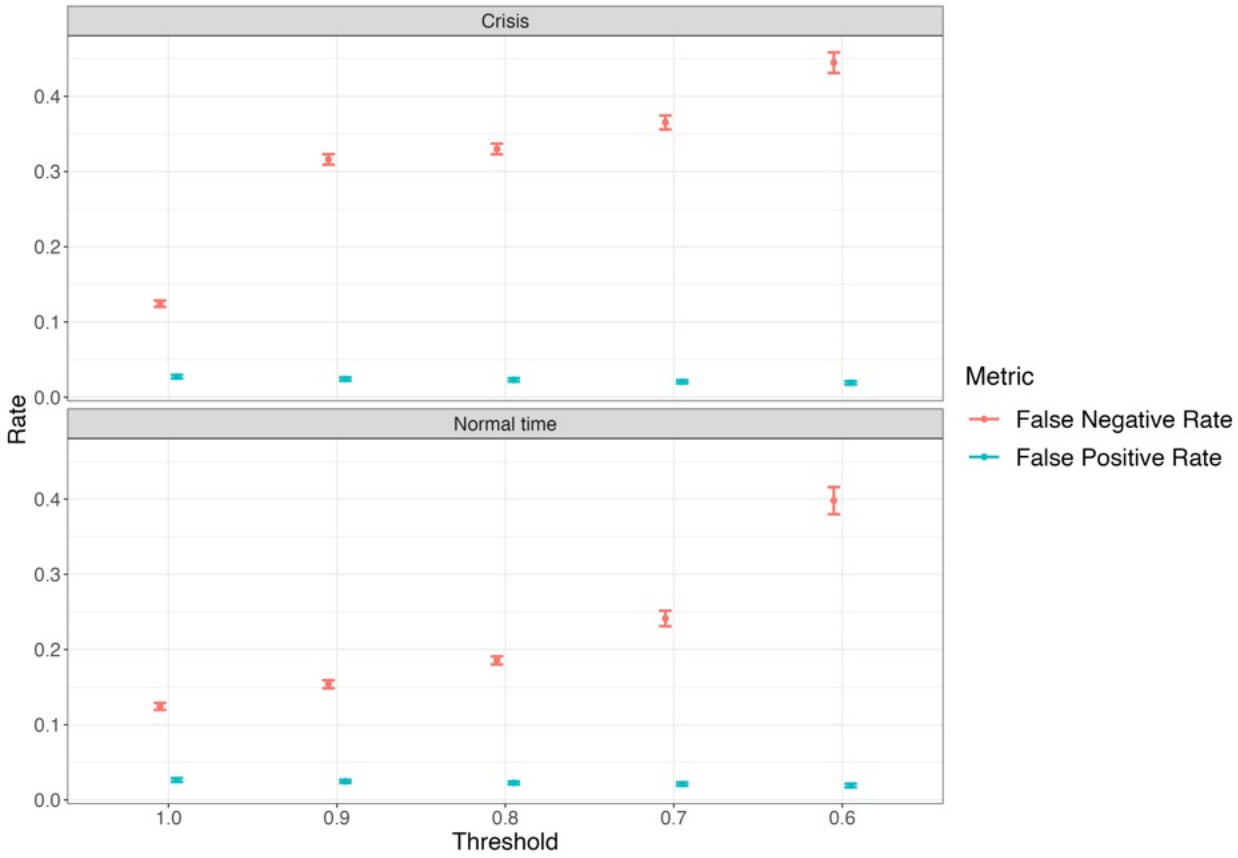
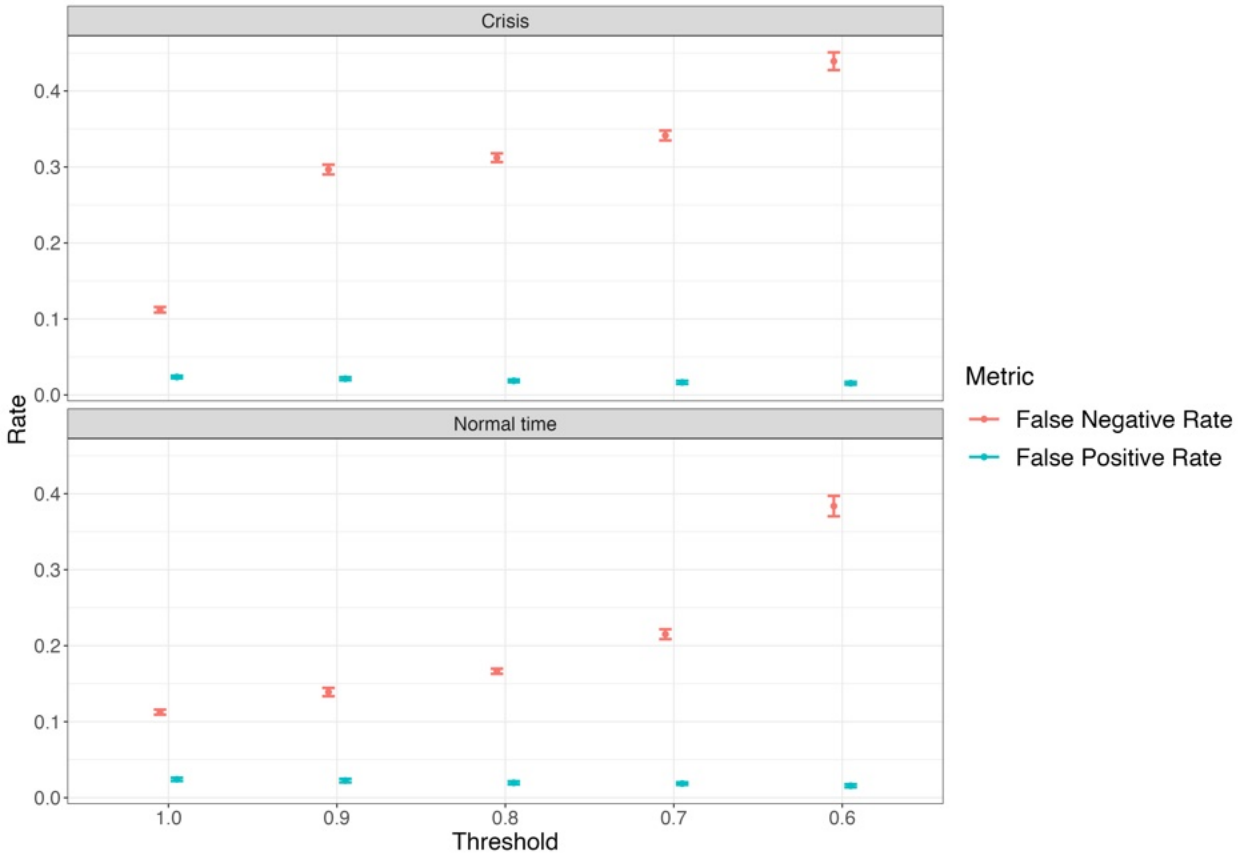


Figure A2 breaks down the models' errors by type. It reports the false positive rate, defined as $\frac{\text{No. of false negatives}}{\text{Total no. of positives}}$, and false negative rate, defined as $\frac{\text{No. of false positives}}{\text{Total no. of negatives}}$, for different censorship AI models. As 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 substantially, with the largest false negative rate more than three times that of the smallest. This is true for both the normal time test data and the crisis test data, with a larger increase in false negative rate during crises. Therefore, Figure A2 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 A3 reports the error rates for models trained on double the amount of domestic data and the patterns are similar.

FIGURE A3. ERROR RATE BY ERROR TYPE - DOUBLED TRAINING DATA



Intuitively, when censorable information is missing in the training data, signals about what should be censored (e.g., keywords that should trigger censorship) become more sparse in the data. As a result, more censorable posts will be able to pass the censorship models undetected because the models could not identify censorable markers from these posts.

# D. Robustness to Weaker Assumptions and Performance Enhancing Techniques

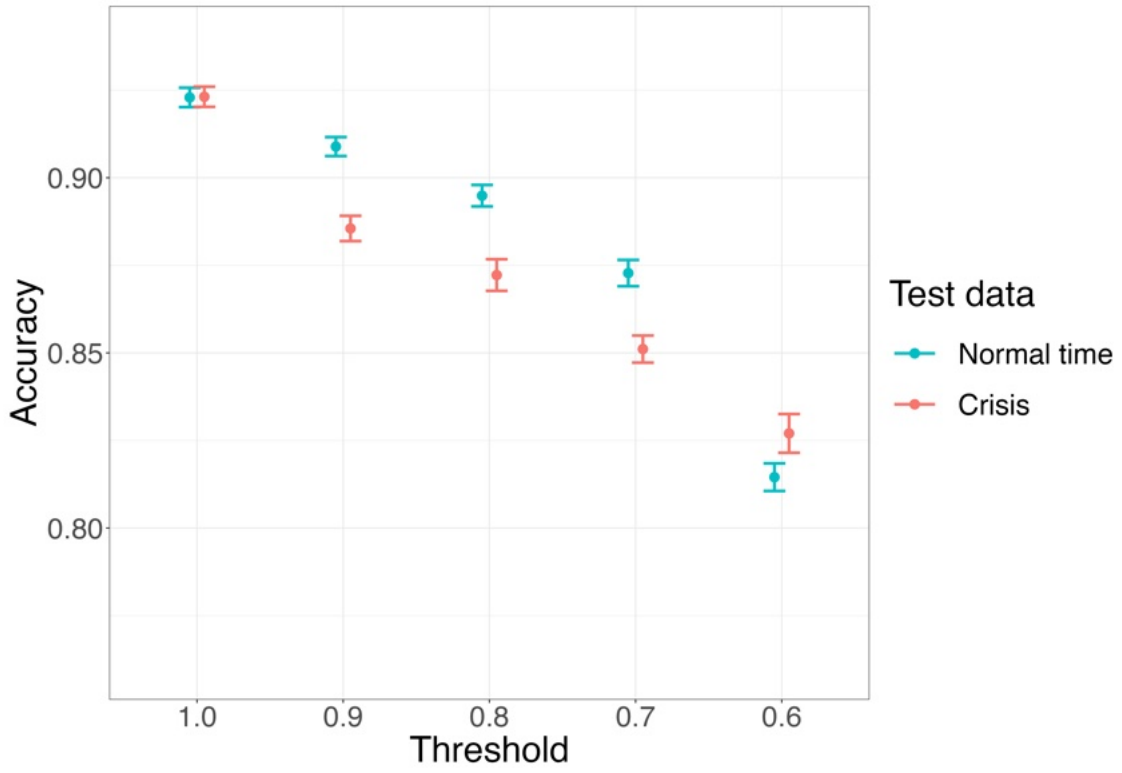
This section reports additional results from weaker assumptions and performance enhancing techniques commonly used in deep learning. The substantive conclusions from the main results hold for the additional tests reported in this section.

## D.1. Data Leakage

Figure A4 shows the accuracy results when 5% of the data above the threshold is allowed in the training data. This relaxes the assumption that citizens have perfect information about what content is censorable and that they are using a pure strategy of self-censorship and preference falsification according to the threshold. Noticeably, allowing data leakages improves the accuracy of the models trained on data with missing data (threshold $< 1$). The improvement is larger for the crisis test data.[A8] However, the accuracy gaps between models of different thresholds still persist. Except for threshold $= 0.6$, the drop in accuracy is also larger for crises than normal times. Essentially, the model's accuracy in crises is determined by a horse race between accuracy loss due to distribution shift (between the training and test data) and accuracy increase due to the prediction problem becoming easier (as the most politically sensitive hence easier to classify posts are now in the test data).

[A8]There is no change for the model trained on the full distribution of data (threhold $= 1$) as there is no missing data to begin with.

FIGURE A4. MODEL PERFORMANCE ACROSS TRAINING DATASETS

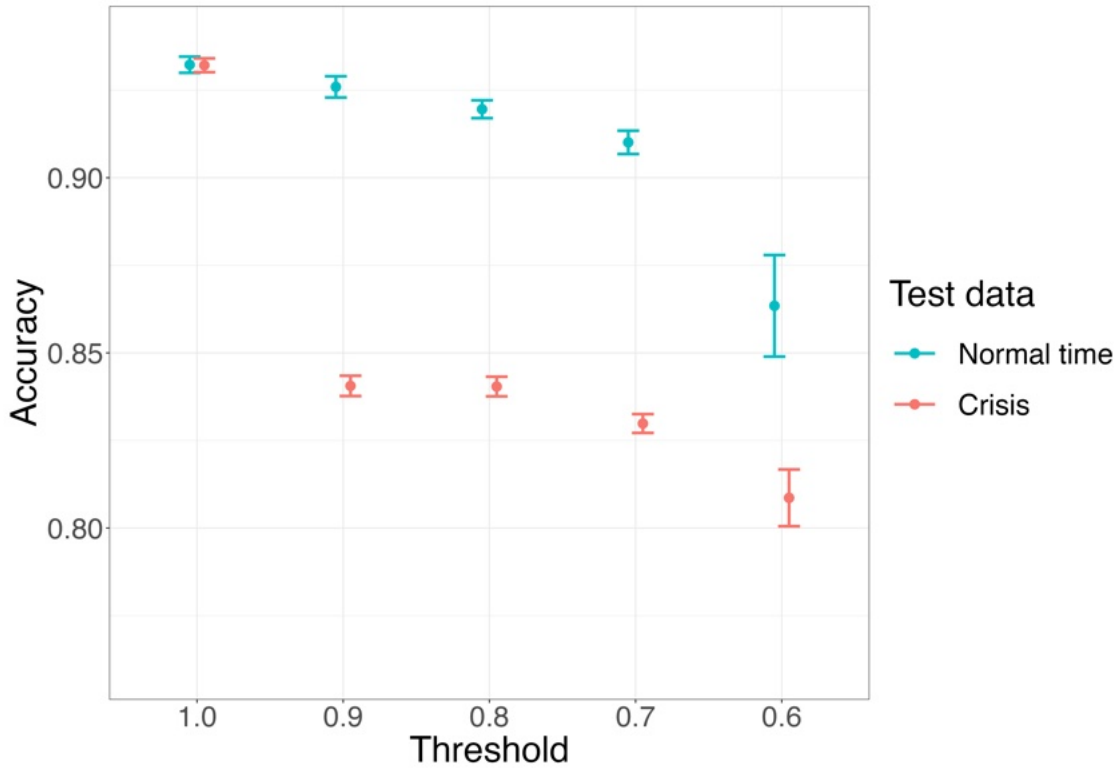



*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

## D.2. Sample Weighting

Figure A5 shows the results of using sample weights in the training. As Table A3 shows, there is an imbalance in the number of censorable and "safe" posts in the training data. To correct for this imbalance, I use $\frac{1}{prop_c} + 1$ as sample weights in the training of censorship AI models, where $prop_c$, $c \in \{\text{censorable, safe}\}$ indicates the proportions of censorable and "safe" content in the training data. 1 is added to the inverse weighting for training stability. Sample weighting improves the performance of the models trained on data with missing data but does not eliminate the performance gaps between 1) models trained on data with different degrees of missingness and 2) the normal time and crisis test data.

Figure A5. Model Performance across Training Datasets

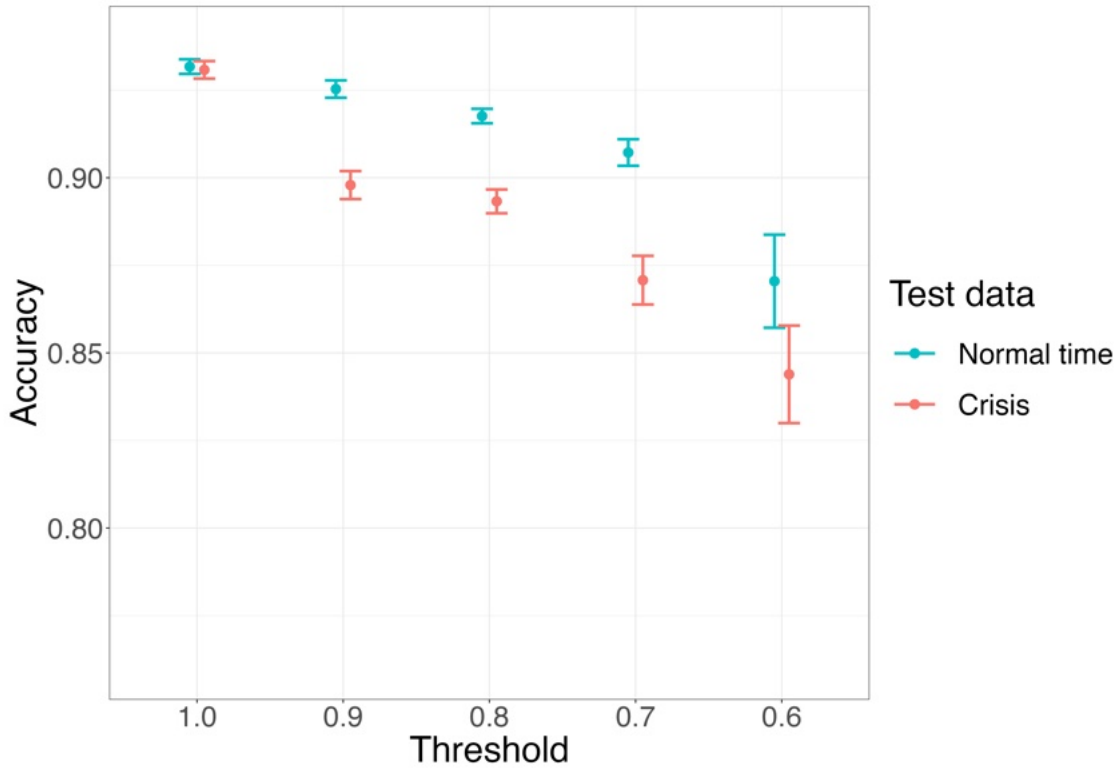



*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

## D.3. Data Leakage + Sample Weighting

Figure A6 combines data leakage and sample weighting. Here we see the benefits from both measures - the accuracy of the model goes up across the board, although the general patterns from the main results still hold.

Figure A6. Model Performance across Training Datasets



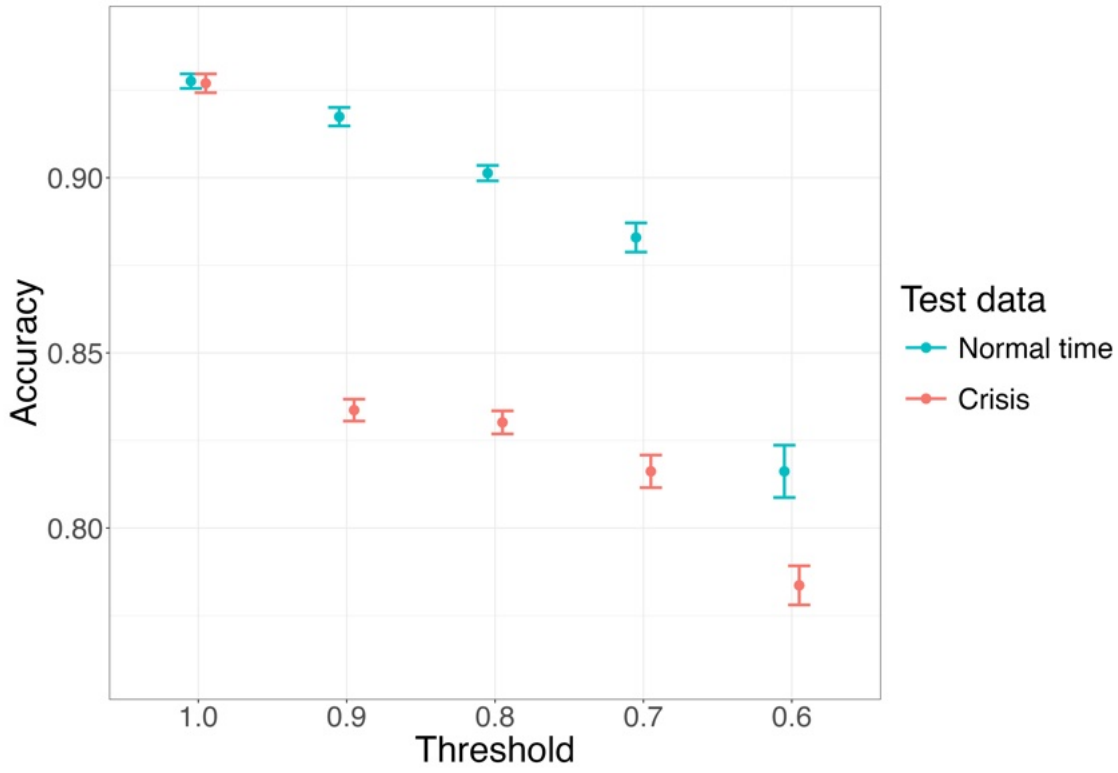

*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

## D.4. Lower Decision Rule

Another measure that can potentially change the accuracy of the censorship model is changing the decision rule - the value of political sensitivity above which posts are labeled as censorable. The main results in the paper uses 0.5 as the decision rule. However, if the sensitivity of many censorable posts is predicted just below 0.5, a lower decision rule can potentially improve the performance of the model. On the other hand, lowering the decision rule can also introduce more false positives, thus negatively affecting model performance.

Figure A7 reports the accuracy results from using 0.4 as the decision rule. The results are little changed as comapared to the main results and the substantial conclusions remain unchanged.

Figure A7. Model Performance across Training Datasets

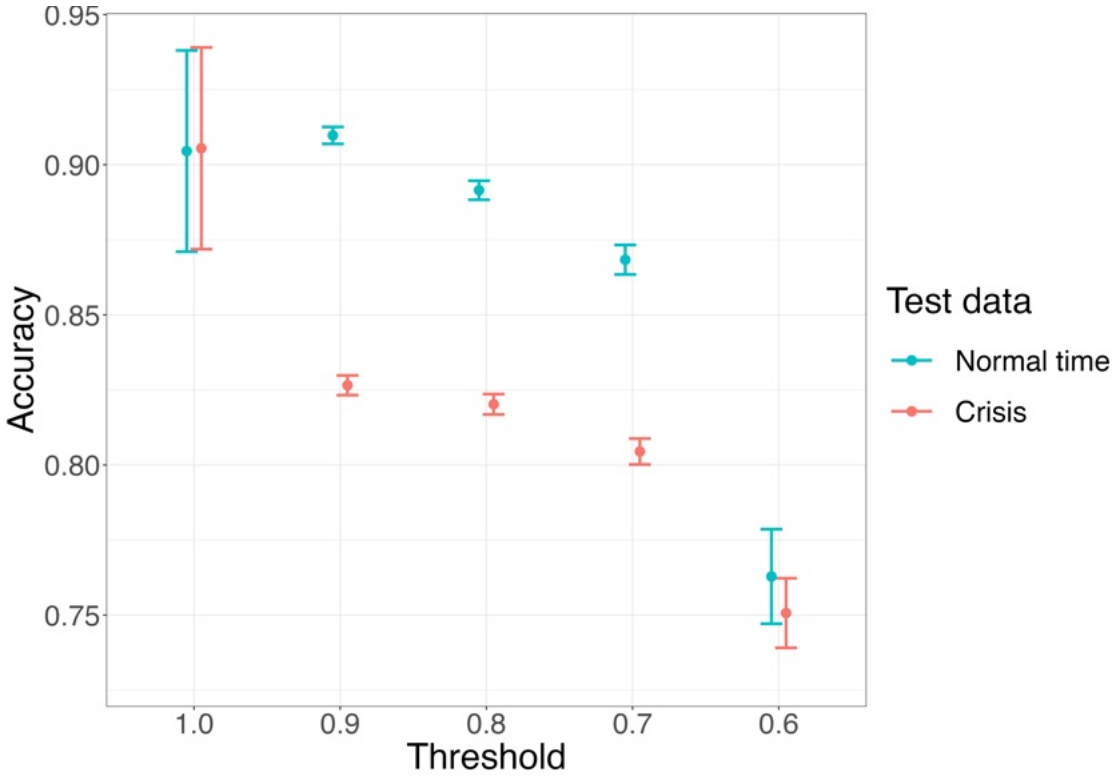



*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

## D.5. Larger Model

Figure A8 shows the accuracy results for a set of larger models. Instead of the 110 million parameter BERT model used for the main results, Figure A8 uses the larger 340 million parameter BERT model. Despite being three times larger, the larger models do not show a significant accuracy improvement or any deviation from the patterns of the main results.

Figure A8. Model Performance across Training Datasets

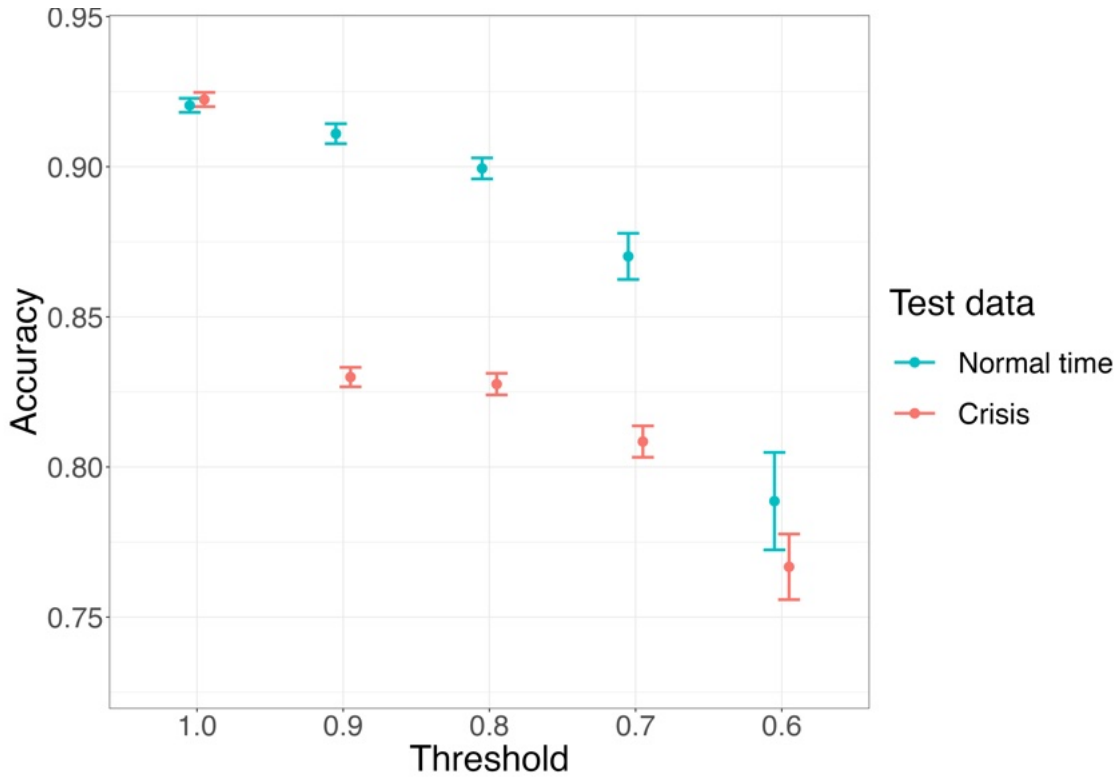



*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

## D.6. Alternative Model

Figure A9 shows the accuracy results for a set of models with a different model architecture. Instead of the BERT model, Figure A9 uses the Chinese version of the ELECTRA model (Clark et al., 2020) as the deep learning model for training. The model has 102 million parameters and is trained using an adversarial framework. Figure A9 shows that that the alternative model architecture yields similar results and the substantial conclusions are unchanged.

Figure A9. Model Performance across Training Datasets




*Notes:* Each threshold value represents a version of the training dataset. Uncertainty estimates are obtained based on the predictions of 25 models for each threshold.

# E. Content Difference Between Weibo and Twitter

## E.1. Keywords for Each Topic

**Topic 1:**

Highest Probability Words: Virus, USA, China, COVID-19, Wuhan, Research, Expert

FREX Words: Influenza, Vaccine, Seafood, Huanan, Inhibition, Paper, Coptis

Score Words: USA, Virus, Influenza, Vaccine, Iran, WHO, SARS

**Topic 2:**

Highest Probability Words: Epidemic, China, Country, Government, Society, People, Control

FREX Words: Stock Market, Democracy, Highly Pathogenic, Subtype, Culling, Socialism,

Denmark

Score Words: Humanity, China, Chinese People, World, Epidemic, Economy, Disaster

**Topic 3:**

Highest Probability Words: Wuhan, Lockdown, Quarantine, Really, Reason, Sadness, Many

FREX Words: Everywhere, Brain, That Kind, Let Go, Owner, Fool, In the City

Score Words: Lockdown, Wuhan, Animal, Really, Wild, Sadness, Quarantine

**Topic 4:**

Highest Probability Words: Confirmed, Cases, Pneumonia, New, Cumulative, Deaths, Novel

FREX Words: Confirmed, Cases, New, Cumulative, Report, Contacts, -Time

Score Words: Cases, Confirmed, New, Cumulative, Discharged, Deaths, Recovered

**Topic 5:**

Highest Probability Words: Pneumonia, Novel, Virus, Corona, Infection, Epidemic, News

FREX Words: Aerosol, Paperclip, Manuscript, Sp, Fecal-Oral, Department Store

Score Words: Corona, Novel, Virus, Pneumonia, Infection, Health, Transmission

**Topic 6:**

Highest Probability Words: Wuhan, Supplies, Hospital, Hubei, Donation, Personnel, Support

FREX Words: Huoshenshan, Red Cross, Dali, Requisition, Leishenshan, Charity, Guidance Team

Score Words: Supplies, Hospital, Donation, Medical Team, Support, Huoshenshan, Dali

**Topic 7:**

Highest Probability Words: Epidemic, End, Hope, Reason, Stay at Home, Video, War Against the Virus

FREX Words: Check-in, Promise, Valentine's Day, Exercise, Stay Home, Hotpot, Quick Click

Score Words: Check-in, End, Go Out, Lottery, Quick Click, Promise, Happiness

**Topic 8:**

Highest Probability Words: Japan, China, South Korea, Hong Kong, COVID-19, Government, Epidemic

FREX Words: Japan, South Korea, Italy, Princess, Diamond, Cruise Ship, Tokyo

Score Words: Japan, South Korea, Diamond, Italy, Princess, Cruise Ship, Hong Kong

**Topic 9:**

Highest Probability Words: Patient, Hospital, Wuhan, Treatment, Sick, Zhong Nanshan, Quarantine

FREX Words: Fever, Nucleic Acid, Fangcang, Traditional Chinese Medicine, Clinic, Recovery, Plasma

Score Words: Patient, Hospital, Treatment, Symptoms, Fever, Sick, Zhong Nanshan

**Topic 10:**

Highest Probability Words: Doctor, Li Wenliang, Reason, Frontline, Candle, Nurse, Passed Away

FREX Words: Li Wenliang, Candle, Passed Away, Elderly, Unfortunately, Daughter, Rumor

Score Words: Doctor, Li Wenliang, Candle, Passed Away, Nurse, Elderly, Rumor

**Topic 11:**

Highest Probability Words: Epidemic, Map, Show, Henan, Reason, Zhejiang, Time

FREX Words: Map, School Opening, School, Student, Hardcore, Ahh, University

Score Words: Map, School Opening, Henan, Show, School, Jiangxi, Prison

**Topic 12:**

Highest Probability Words: Epidemic, Prevention and Control, Work, Personnel, Community, Residential Area, Epidemic Prevention

FREX Words: Highway, Registration, Traffic Police, Passenger Transport, Sub-bureau, Highway, Scenic Spot

Score Words: Prevention and Control, Police, Residential Area, Command Center, Public Security, Notice, Highway

**Topic 13:**

Highest Probability Words: Mask, Protection, Disinfection, Go Out, Wash Hands, Medical, Contact

FREX Words: Alcohol, Pharmacy, Ventilation, Taiwan, Elevator, Wuhan, Cleaning

Score Words: Mask, Disinfection, Wash Hands, Medical, Taiwan, Go Out, Alcohol

**Topic 14:**

Highest Probability Words: Stay Strong, Wuhan, Epidemic, China, Fight, Reason, Tribute

FREX Words: Fist, Believe, Relay, Heroes, Song, Hello, Creation

Score Words: Stay Strong, Tribute, Wuhan, Frontline, Medical Staff, China, Frontline

**Topic 15:**

Highest Probability Words: Epidemic, Resumption of Work, Enterprise, Company, Production, Prevention and Control, Impact

FREX Words: Resumption of Work, Enterprise, Employee, Resumption of Production, Salary, Bank, Return to Post

Score Words: Enterprise, Resumption of Work, Resumption of Production, Production, Company, Market, Employee

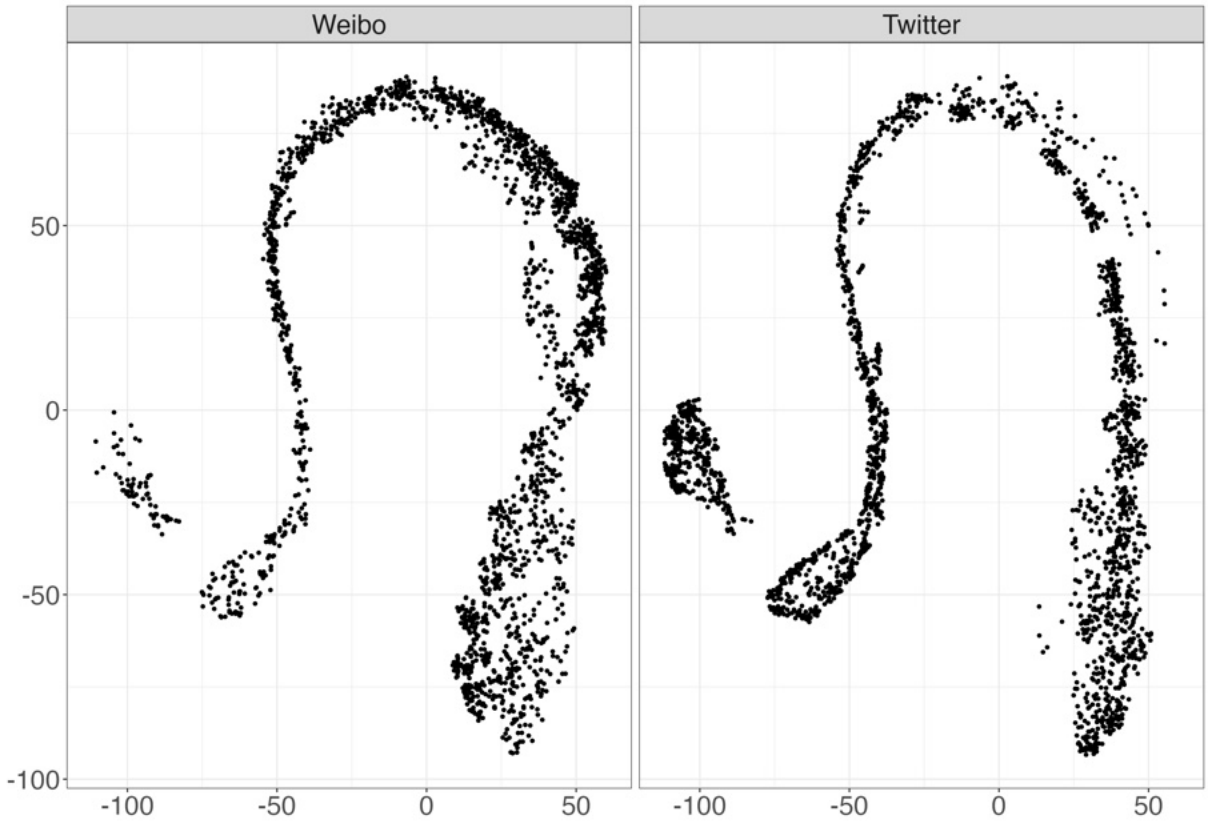
## E.2. Model's Internal Representation of Weibo and Twitter Data

Figure A10 provides evidence that the difference in content between Weibo and Twitter propagates to censorship AI models. It shows how a censorship AI model trained on the combined Weibo-Twitter 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 2000 social media posts from the Weibo and Twitter data respectively. I then use the model to obtain the embeddings (internal representation) of the combined 4000 social media posts. As the embeddings are high-dimensional, I use t-distributed stochastic neighbor embedding (t-SNE) to reduce the dimensionality of the embeddings and plot the distributions in a two-dimensional space in Figure A10. Essentially,

t-SNE is a statistical technique that models high-dimensional data in a two-dimensional space such that similar data points are closer to each other and dissimilar data points are further apart with high probability.

As Figure A10 shows, the overall shapes of the two data sources are quite similar: both plots display a curved, hook-like pattern. This is to be expected as both sources are collected based on the same COVID topics using a common list of keywords. However, the distribution of the data points is quite different between the two sources. Weibo's plot has a denser concentration of points on the upper right curve whereas Twitter's plot has a sparser concentration around the same area but a denser concentration on the separate cluster on the left loop region.

FIGURE A10. MODEL'S INTERNAL REPRESENTATION OF WEIBO AND TWITTER DATA



# References

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* .

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.

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* .