Big Data is Watching You; Assessing the Dangers and Responses to Predictive Repression in Autocracies

Timothy Liptrot
Georgetown University
Word count: 9000

June 5, 2021

Abstract

Repressive states have begun experimenting with machine learning to identify citizens with grievances who might join dissident movements. We use survey data from 5 authoritarian states to identify data sources most dangerous for predictive repression. We find that demographics, news sources and political behavior are most predictive. We also show that use of encrypted applications is not predictive when the application is sufficiently popular.

1 Introduction

On June 25, 2017, a Xinjiang regional official of the Chinese Communist Party (CCP) circulated a bulletin about dissidents apprehended and sent to reeducation camps. While arbitrary detention in repressive states is common, these "dissidents" are the first selected for interrogation by an algorithm, rather than direct human identification. Officials were instructed to "put measures in place according to classifications (...) different types of tags pushed out by the "integrated platform" (Wilson-Chapman, Wilson-Chapman). On the basis of these tags and subsequent interrogation, tens of thousands were sent to reeducation camps in just that month.

Predictive repression is the automated imputation of citizen attitudes and ideology to prevent anti-regime activity. Predictive repression in Xinjiang relies on "big data" collected from such diverse sources as police interrogations, internet use, searches of personal computers, government records, tax data, government services, utilities, biometrics and cellphone location data. In 2019, scholars began raising the alamr about this new repressive pattern, coining the term "dataveilance" to describe intersection of big data with the surveilance state (**?**). Feldstein (2019) and Frantz, Kendall-Taylor & Wright (Frantz, Kendall-Taylor & Wright) show that preventative repression is spreading from Xinjiang across China and to other autocracies.

The current discussion of predictive repression lacks a clear picture of which behavioral variables are most dangerous in the hands of repressive actors. Both scholarly and media descriptions of AR emphasize the number and diversity of behavioral variables that states are gatheringQiang (2019). However some variables the CCP currently uses lack a plausible connection to grievances against the state

(Wang, 2019). It is possible that new machine learning techniques extract usable information from previously unrelated behaviors. We cannot assume that all used data is predictive because repressive actors can falsely overstate the effectiveness of AR to bluff complete knowledge of grievances.

A general idea of which data types are most important enables both researchers and citizens to better evaluate the state's claims. It also enables pro-democracy actors to influence dataveilance of certain states, by producing counter-surveillance software (Bock et al., 2019) or cheapening digital privacy. Assessing variable importance enables pro-democracy actors to target those efforts.

We directly measure the predictive value of a variety of citizen-level variables for imputing grievances against the state. We train Random Forest models on survey data form 6 autocracies (Algeria, Armenia, Egypt, Jordan, Kuwait and Morocco). We left as x-data all survey questions which repressive actors could plausibly access, from internet usage habits to household location to charitable contributions. We use observed instances of digital surveillance to identify all surveilable variables. We then use cross validation to assess the predictive value of the variables both separately and in general categories such as internet habits, movement, tax data, and service consumption.

The dataveillance capacity of repressive states varies widely (Frantz, Kendall-Taylor & Wright, Frantz, Kendall-Taylor & Wright). China has hired hundreds of thousands of security specialists to read internet communication and search citizens cell phones during searches and at checkpoints (Qiang, 2019) (Greitens, Lee & Yazici, 2020). In Egypt a majority of eligible citizens do not even submit tax data (Hassan

Abdel Zaher, Hassan Abdel Zaher). Our method best approximates an intermediate state between the two. While survey data is a poor proxy for state dataveillance, there are few ethical alternative research designs. Our technique gives the best possible starting measure.

We find that variables which describe either political behavior or information sources better predict ideology. Tax and employment information had very low predictive value except for income measures. We find political behavior and information consumption are highly predictive of ideology. Furthermore we show that encrypted applications, such as whatsapp, are not predictive when widely used. This suggests that digital privacy interventions will disrupt predictive repression, forcing autocrats to rely on fewer more expensive data sources.

The paper proceeds as follows. The next section explains why leaders impute the preferences of their followers, while followers falsify their preferences. The following section contextualizes prediction in the broader trend of digital repression. Next, we briefly describes the history of predictive repression in Xinjiang, the only known use case. We then describe our data sources and analysis method. The results section makes general observations about variable importance. Finally, we conclude with comment on policy priorities for preventing AR.

1.1 Why Leaders Impute Beliefs

Predictive repression automates an ancient practice of political leaders: imputing the beliefs and preferences of followers. Natural allies who share a leaders preferences betray the leader less and require less incentivizing (Kahan, 2013) (**?**). Once the leader

promotes followers to positions of power or grants them freedom, that follower can use the power to bargain for policy concessions or defect to an another coalition. Followers with ideal policies close to the leadership demand smaller policy concessions and rarely defect (**?**) slater$_{o}rdering_{2}$010.$Evenindemocracies, partyselectoratesprefercandidateseithermo$u

In autocracies, disgruntled citizens often call for a transition to democracy, not just changes in policy or leadership (for theoretical motivation, see **?** Acemoglu and Robinson, 2006; for survey analysis see Nathan (2020) Nathan, 2020). If a rebellious faction removes the autocrat through a coup or revolution, the autocrat loses any ability to punish them (**?**)(Svolik, 2012). When revolution becomes likely, coerced support evaporates and only those who strongly prefer regime continuity will stay loyal. Empowering natural allies increases state capacity by facilitating the cooperation of elite factions (**?**)(Slater, 2010). Autocratic regimes are under pressure to select loyalists or natural allies.

Autocrats respond by placing loyalists in positions of power and degrading the collective action capacity of dissenters (**?**) (DeMesquita and Smith, 2012). For example, the Jordanian monarchy selects agricultural policies to prevent Palestinian businesses dominating the food industry (**?**)(Keulertz, 2014). This goal is much easier stated than achieved.

Autocrats have a variety of mechanisms for sroting people by loyalty. Demanding public displays of devotion is a simple mechanism because citizens who highly value truth and the public good are more likely to fail (Crabtree, Kern & Siegel, 2020). Tests range from requiring photos of the leader in businesses to demanding obviously false statements about leader quality. Such displays also humiliate the supplicant

and prevent joining rival factions (Chung-Hon Shih, 2008). As cult displays become normalized, they must become more unreasonable or unpleasant to continue sorting. Autocrats also use willingness to repress as a signal of loyalty, which is Credible due to its high cost (Gregory, 2009) (Gregory, 2009). Ethnic identity, family origin (Quinlivan, 1999)(Quinlivian, 1999) and incompetence are also used as signals of loyalty (Bausch, 2018)(Bausch, 2017).

Autocratic citizens have strong incentives to pass these tests. The appearance of natural allyship opens doors in the public bureaucracy and protects from negative sanctions. Unlike democracies, public support for the opposition does not promise rewards following a majority cycle. As a result each actor wishes to know the true preferences of others, but to control their own presentation (Kuran, 1995)(Kuran, 1995).

One response is for citizens to falsify their preferences by lying (ibid). A natural experiment conducted on a purge in Shanghai showed that the purge both increased statements of regime support while decreasing actual support (Jiang and Yang, 2016), Deception can become so natural that a recent study found 20% of anonymous respondents in China falsify anonymous survey responses (**?**). Alternatively, citizens may simply adopt socially adaptive beliefs (SAB) that pass tests. Choosing to believe prevents the detection of dishonesty through social cues or physical tells (**?**). Preference falsification and SAB prevent regimes from learning their actual popularity. The collapse of the Soviet Union was proceeded in 1985 by a large shift in public opinion against communism (recorded in secret East German party surveys) (Kuran, 1995). The East German people successfully hid their beliefs, even from each other,

until a sudden cascade caught the world by surprise in 1989.

If the regime empowers many persons with hidden preferences for systemic change, it is in danger of a revolutionary cascade. As people enter the opposition, the cost of joining declines because punishment is dispersed and regime continuity becomes less likely (Kuran, 1995). The reduced loyalty incentive further expands the opposition, creating a self-enforcing cycle. This is most obviously true for preference falsifiers, but socially adaptive believers and even true loyalists will switch as the incentives for public belief shift in revolution (duffy and lafky). The regime therefore most prefers to empower true believers willing to pay costs to prevent change.

The regime has limited responses if it learns an actors secret preference. The regime must hide the size of the opposition, so dissidents are unaware of their strength in numbers (Kuran, 1995). Leaders often respond to challenges with repression (Davenport, 2007), but a belief is not a challenge unless it is common knowledge. If grievances are imputed from non-political behavior, direct repression will deter the innocuos signal, rather than the actual views. Socially adaptive believers do not choose a preference until the revolutionary cascade begins, making preference deterrence impossible and preventative repression is difficult to legitimate. The coerced confessions common in Stalin and Saddam Husseins regimes were plausibly intended to legitimate preventative attacks on suspected disloyalists.

1.2 Digital Repression

Elites have always run the follower selection game, but in the past were limited to targeting other elites. Imputing beliefs throughout an entire society presents practi-

cal challenges. Autocrats may be skilled at detecting true loyalties but each regime insider can only assess a few hundred followers at best. This necessitates a bureaucracy of loyalty assessors, such as the Soviet Commissars. Modern autocracies use secret policy ministries, sometimes several (**?**) or a single political party to recruit and indoctrinate member-informants (**?**). Unsolicited enforcement by citizens is common, but unpopular autocrats have few informants. Outside of a few exceptional regimes (Russia, 1930-55; Cambodia 1976-79; China 1966-76) non-elites are rarely primary targets.

Modern predictive repression using big data could enable loyalty imputation for the masses (Feldstein, 2019)(Qiang, 2019)(Feldstein, Qiang). Digital surveilance promises low enough costs for any citizen to become a target, as long as they produce aggregatable data. Repression can expand beyond elites or highly ideological and risk-tolerant first movers (**?**)(Kuran, 1991). Additionally, the regime need not wait to observe revolutionary acts or statements to punish disloyal citizens. An accurate behavior predictor would overcome both preference falsification and SAB.

Predictive repression is likely to increase regime durability if it:

- Is cheaper to implement than traditional belief imputation
- Can be applied more widely, accessing previously unmonitored social strata
- Is more accurate than traditional methods
- Makes belief-based discrimination more acceptable to the public.

Autocrats are already taking advantage of the digital world to cling to power (Frantz, Kendall-Taylor & Wright, Frantz, Kendall-Taylor & Wright). Many regimes

commonly monitor digital communications, often by hand, to identify the opposition. Gohdes (2020) found that the Assad regime used targeted killings more in areas with higher internet access, and indiscriminate killings in areas with lower access. Activists and journalists in Egypt, Iran and elsewhere are routinely jailed for online statements. States are already monitoring digital communications to identify forming protests and proactively respond. Singapore is pioneering the use of facial recognition systems to identify protesters (Tan, 2020) and the Yanukovych regime in Ukraine mass texted all cell phones near protests in Kiev as an implicit threat though the threat failed to dissuade protesters(Feldstein, 2019). Frantz, Kendall-Taylor & Wright (Frantz, Kendall-Taylor & Wright) catalogued mechanisms of digital repression: identifying likely regime opponents by combining mass surveillance with machine learning; monitoring regime insiders; automating censorship, e.g. the Great Firewall of China; gauging public sentiment to anticipate and prevent protests; proactively spreading misinformation to disrupt collective action. Predictive repression emerged late (2016) and has spread less quickly than deterrent punishment and censorship tactics. While the techniques are new, the goals are constant: to raise the costs of disloyalty, to identify the opposition, and to prevent collective action against the regime.

1.3 Predictive Repression Today

This section summarizes public knowledge about AR in Xinjiang, the only verified implementation. ¡summarize findings below¿

There is a long history of repressive escalation, deescalation and dissent in the

resource-rich Muslim-majority province of Xinjiang (Greitens, Lee & Yazici, 2020). The most recent episode of major public resistance occurred in 2008-2009 with attacks on police stations and violent clashes between Uighur and Han. An estimated 200 people were killed in the police response. The CCP responded by escalating tested policies, like embedding police in local communities. Although public contention declined after 2009, Uyghur participation in Islamic terrorism increased. Several high profile terror attacks occurred within China and some 300 Uighur travelled to Syria to fight with ISIS. The CCP decided to eradicate Islamist extremist ideology by reeducating a significant fraction of Xinjiangs population (Greitens, Lee & Yazici, 2020). In 2016 the CCP began detaining Uighur in reeducation camps by the millions.

The primary purpose of the IJOP is to identify citizens with a specific political ideology (political Islam). It is not yet an all-purpose tool for attacking challengers. The context suggests the implementers were aware of its low accuracy. A party official described the plan as "You can't uproot all the weeds hidden among the crops one by one (...). Re-educating these people is like spraying chemicals on the crops. That's why it is a general re-education, not limited to a few people." (Greitens, Lee & Yazici, 2020). Perhaps because the targets are an isolated minority, the CCP tolerated a a high false positive rate.

We possess detailed knowledge of mass surveillance in Xinjiang from a mobile app used by security forces which Human Rights Watch accessed and reverse engineered Wang (2019). In addition a small number of internal party documents have been leaked (Wilson-Chapman, 2018). The app, named the Integrated Joint Operations

Platform (IJOP), collects data on a surprisingly wide set of personal activity. Some entries are plausibly connected to subversive behavior such as sharing "Wahhabism", "knowing how to make explosives", and possessing foreign messaging applications. But many entries lack plausible relevance to regime support: "unwilling to enjoy policies that benefit the people", "Collected money or materials for mosques with enthusiasm" and "household uses an abnormal amount of electricity".

Data collection is labor intensive. Tens of thousands of security contractors interrogate citizens in mandatory home visits or at checkpoints (Wang, 2019). Security personnel also routinely search residents phones for suspicious applications. The IJOP also relies on automated data streams from public and private service providers (Wang, 2019). It imports data from; CCTV cameras equip ed with facial recognition software; visitors to residential areas and schools; police checkpoints; package shipping; detailed information about package shipping; electricity consumption and gas station visits. When an "unusual" amount of electricity use is detected officers are dispatched to investigate the household and seek a plausible explanation. Biometric information including DNA samples, fingerprints, iris scans, blood types and voice samples is also collected, but lacks a plausible predictive value.

We know much less about how the IJOP analyzes this data. The app itself contains only simple conditional statements for investigation tags, such as "if the person who drives the car is not the same as the person to whom the car is registered, then investigate this person". We also know that the implicit probability threshold for detention is very low. For example, reports suggest that any person possessing the messaging application WhatsApp is detained. However, the central IJOP system

may use more sophisticated algorithms. Leaked documents show that in 2017 the IJOP was tagging tens of thosuands of individuals per month, a majority of whom were detained for reeducation (Wilson-Chapman, Wilson-Chapman).

This technology is likely to spread outside of Xinjiang if it provides real repressive value. Xinjiang is a first training ground because China is an AI development hotspot and the CCP tolerates a high false positive rate in Xinjiang. As the technology spreads and/or specificity improves, other regimes will copy these tactics.

China is already exporting the technology to implement predictive repression elsewhere (Feldstein, 2019)(Polyakova & Meserole, Polyakova & Meserole). Singapore, Malaysia, Zimbabwe and Dubai are importing facial recognition systems from Chinese state contractors. Ethiopian security services use use ZTE provided tech to digitally monitor opposition activists (**?**)(HRW, 2014). Qiang suggests that China's social credit system, ostensibly for incentivizing prosocial behavior, could easily extend the IJOP to the rest of China (Qiang, 2019). Venezuela is contracting with Chinese firm ZTE to build a "national ID card, payment system, and "fatherland database" that will track individuals' transactions alongside personal information such as birthdays and social media accounts" (Polyakova & Meserole, Polyakova & Meserole)(Berwick, Angus, Berwick, Angus). A Venezuelan justice ministry defector reported the system is directly inspired by the social credit system. While these applications have yet to begin automatically imputing grievances, they underline the threat of global spread.

Research design

The main objective of this study is to determine what behavioral data is most concerning for predictive repression. We roughly simulate grievance imputation using publicly available surveys on personal attributes and political attitudes in authoritarian states from the Arab Barometer. Our x-values are all observable behaviors and characteristics that at least one repressive state surveils. We built two models, predicting trust in government and democracy preference respectively. We train several algorithms then report the variable importance information from the most predictive models (random forests).

Data

In democracies individuals have little incentive to hide their ideological positions and often intentionally signal their regime preference. This is not the case in repressive states, where individuals have strong reasons to hide their ideology. Ideology imputation relies on public behavior, so tools for predicting ideology in democracies should be less effective in autocracy. If a state closes opposition newspapers, choice of paper no longer signals ideology. As Kuran argued "In [repressive states] the very forces that discourage truthful expression also inhibit the collection and dissemination of opinion data" (Kuran, 1995)(1995 p. 1538). Because we wish to generalize to autocracies, we only used data from countries with polity IV scores below 5 (Regan & Henderson, 2002).

We selected the Arab Barometer (AB) Wave V survey data from 2018. The

AB covers a variety of states which rely on repression and limit mass political participation. It provides detailed information about political beliefs, social behavior, economic activity, and information consumption, with especially detailed questions on internet use.

We only included countries where a salient ideological divide exists between pro-democracy and anti-democracy positions. We expect ideologies to be more predictable when discussed or personally experienced, and respondents may give random answers to non-salient questions. The Arab countries share a salient ideological divide about the degree of democracy. In Sudan, the survey was completed 12 days before the 2018 revolution began with an 8-month mass civil disobedience campaign in pursuit of democracy. A year after the survey Algeria entered a pro-democracy revolution (Volpi, 2020). Ideological positions toward democracy are also shaped by life experiences. In Egypt, **?** find that sustained revolutionary protests reduced local support for democracy by disrupting daily life.

We also did not model countries experiencing civil wars (Iraq, Liby, Yemen) or foreign occupation (Palestine). These considerations left 6 states, Algeria, Egypt, Jordan, Kuwait, Morocco and Sudan. After removing null responses, each country had at least 1400 responses suitable for use, and all but Kuwait had over 2000.

1.4 Y Variables

Our Y-variables were inspired by recent work on citizen grievances in repression (Gregory et al.; Rozenas et al; Aspinall). We interpret a grievance as a preference for anti-regime action or for regime change. A grievance is the positive dummy

variable. For robustness we included two grievance metrics as y-variables.

The trust in government Y-variable is based on trust in the "Government (council of ministers)", as reported by AB. A grievance is recorded for the answers "not a lot of trust" or "no trust at all". The regime preference variable is positive if the respondent preferred the statement "Democracy is always preferable to any other kind of government" over "Under some circumstances, a non-democratic government can be preferable" or "For people like me, it doesn't matter what kind of government we have". The respondent must also report that the country is below six on a ten-point democracy scale. Because less than 10% of respondents ranked democracy as a priority issue, issue priority was not used.

1.5 X variables

We included in the X variables all behavioral information that could plausibly be collected by a high-capacity autocracy such as China. We did not vary the inclusion criteria by the capacity of the country of data collection. Wherever possible we used verified accounts of state surveillance to decide inclusion criteria.

The variables user_facebook, user_youtube, user_whatsapp, user_twitter, user_snapchat, user_instagram and user_telegram identify users of said apps. App usage is a target of the IJOP surveillance system, where it appears that possession of an end-to-end encrypted app is sufficient evidence for arbitrary detention of Uighurs. Egyptian security forces regularly demand access to citizens phones to search for anti-regime communication (Malsin and EL-Fekki, 2019). During the 2020 Belarus protest movement the dominant messaging app was not encrypted, forcing activists to switch to

Telegram. The security forces responded by accosting citizens in the street and searching their phones for telegram (Gerdžiūnas, 2020). In extreme cases citizens were tortured until they granted access to their phones. In all identified cases the regime either coerced citizens to give access or used a physical connection to hack into the phone, rather than surveiling app usage digitally en masse. The diversity and security of cellphones may inhibit mass surveillance for the moment.

Repressive states also spend significant resources monitoring social media (Frantz, Kendall-Taylor & Wright, Frantz, Kendall-Taylor & Wright). One report leaked that the CCP employed 2 million persons to monitor digital communications identified by keyword in 2014 (Xu & Albert, 2014). Our variables socmed_use and internet_use measures numbers of hours of use per week, respectively. Internet is a binary variable for access to internet. Infs_socmed is true for all respondents who reported social media as their main source of news. We expect this coarse data greatly underestimates the relative importance of social media data due to the richness of text data. Furthermore, current revolutionary movements are shifting toward secure social media applications. The Belarus revolution of 2020 was organized mainly on Telegram. Iranians moved to the secure app Telegram for both texting and commenting, and Iranians have repeatedly defeated state attempts to block the app (Radio Free Europe). We revisit the surveillance of social media in our conclusion.

We include data on political behavior such as petitioning, protesting and voting. Unfortunately only local voting behavior was available in the Arab Barometer data set. Petition signatures are generally public knowledge. Voting behavior is often secret but could conceivably be surveilled, especially when the regime controls the

voting format. We included protest attendance in light of a 2014 incident in Ukraine in which the regime mass texted all phones in the area of opposition protests with a threat. The Chinese government is also developing facial recognition software to identify protest participants (**?**).

We also included a variety of data that states routinely collect in the process of tax and service provision. These include income, education (educ), age, gender, charitable contributions as a dummy variable, marriage, and employment. We further breakdown employment status between housewife, unemployed, student, retired, self employed and between the private and public sector. The variable orgmem records membership in an organization or club, which the IJOP monitors.

We excluded from x-data all responses based on the private beliefs or preferences of respondents. That includes trust in other institutions, beliefs about women's rights, and preferences for other state institutions. It is possible that NLP will grant access to some such preferences to future AR implementations using social media statements. However this data would be null for most person-variable dyads because social media statements are much fewer than all possible beliefs. Supposing that 1% of citizens tweet about their trust in capitalism, using those tweets would require an inverse increase in training data. Training data is constrained because unbiased information about citizen beliefs is expensive. Social media data which reveals grievances directly do not require prerdictive imputation and are beyond the scope of this study. Secondly, we would expect trust in other institutions to correlate with trust in regime if citizens do not distinguish within the state, regardless if regime-change preferences.

Some questions were not asked in all countries (local voting and Palestinian or Jordanian descent). Questions were also dropped if more than 100 respondents refused per country, which included income for Algeria and Morocco. Algerian respondents were not asked about trust in government. Kuwaiti respondents were not asked about their preferred political system.

The survey data may suffer from preference falsification or self-censorship. Robinson & Tannenberg (2019) find self-censorship of 25% in list experiments in China. However, we would not expect self-censorship to systematically bias the predictive value of different variables, particularly across the five countries. Furthermore, any actual implementation of AR would suffer from similar or higher levels of data falsification. The Xinjiang leaks show that citizens evade surveillance by faking documents, discarding their phones, and registering under the identities of dead relatives (Wilson-Chapman, Wilson-Chapman).

Data Analysis

First we identified the most effective classifier algorithm for the dataset using cross validation. Cross validation allows us to estimate the best model before running the test data through it by "holding out" a subset of the training observations and then using them as a stand-in to test. The Cross-Validation method used here is k-Fold validation, which splits the observations into k groups, with k-1 groups used to train the model and the last group used as the test set. This is then repeated k times, with the cross validation estimate computed based on the average of the k test groups. For the purposes of this analysis, k was set to 5.

The classifiers that were tested were:

- Naive Bayes- a probabilistic classifier based on the Bayes theorem. It calculates the probability that there is backsliding, given the values of the independent variables, assuming that the independent variables are also independent from one another

- K Nearest Neighbors- a classifier that identifies the number of 'K' points from the training data that are closest to the observation of interest, and then determines its category based on the majority category of the nearest points.

- Decision Trees- A classifier which sorts observations by splitting them based upon specific decision criteria. It then predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs

- Random Forests- A classifier which builds several different decision trees by pulling several training sets from the training data along with a random number of predictors. It then creates its decision criteria by averaging across the predictions from each tree

- Support Vector Machine- A classifier which sorts objects into categories by creating "decision boundaries" distinguishing between classes.

In order to evaluate the performance of each algorithm, the full data set was split into two subsets: a training set and a testing set. The algorithm was first fit on the training set, and then based on the information it gained from the training data it

predicted the categories of the test set. The model was then judged on its ability to correctly the predict the dependent variables of the test set. The segregation of training data from testing prevents over fitting so model performance is accurately scored.

The Random Forest algorithm outperformed all other classifiers in every data set. The remainder of the paper focuses on them.

Decision trees consist of a series of splits on the original dataset, known as a tree. Each split, which can be thought of as a branch, is made on a "decision node." This decision node denotes a criteria that the data is being split on, for example countries with a GDP per capita greater than 25,000. In that countries with a GDP greater than 25,000 would go on one branch, while countries with GDP less than 25,000 would go on another branch. These groups would then be split on another criteria. The farthest branches of these trees are referred to as "terminal nodes" or leaves. In this way a decision tree is actually like an upside down tree, with the leaves on the bottom. The model utilizes a "top down" and "greedy" approach in deciding what features to split on. Put simply, this means that starting from the top of the tree, the algorithm makes each split based on what best minimizes classification errors at that specific step (ie grouping as many 1s together and 0s together as possible while minimizing the number of members of the other class in the group).

Figure X gives an example of a single decision tree fit on the democracy preference variable using only education level and Facebook use. The education level has been simplified to a scaler from 0 to 1, with no education represented by 0, elementary by 1/6, secondary by 2/6 and so on. At the first node the model selects completing high

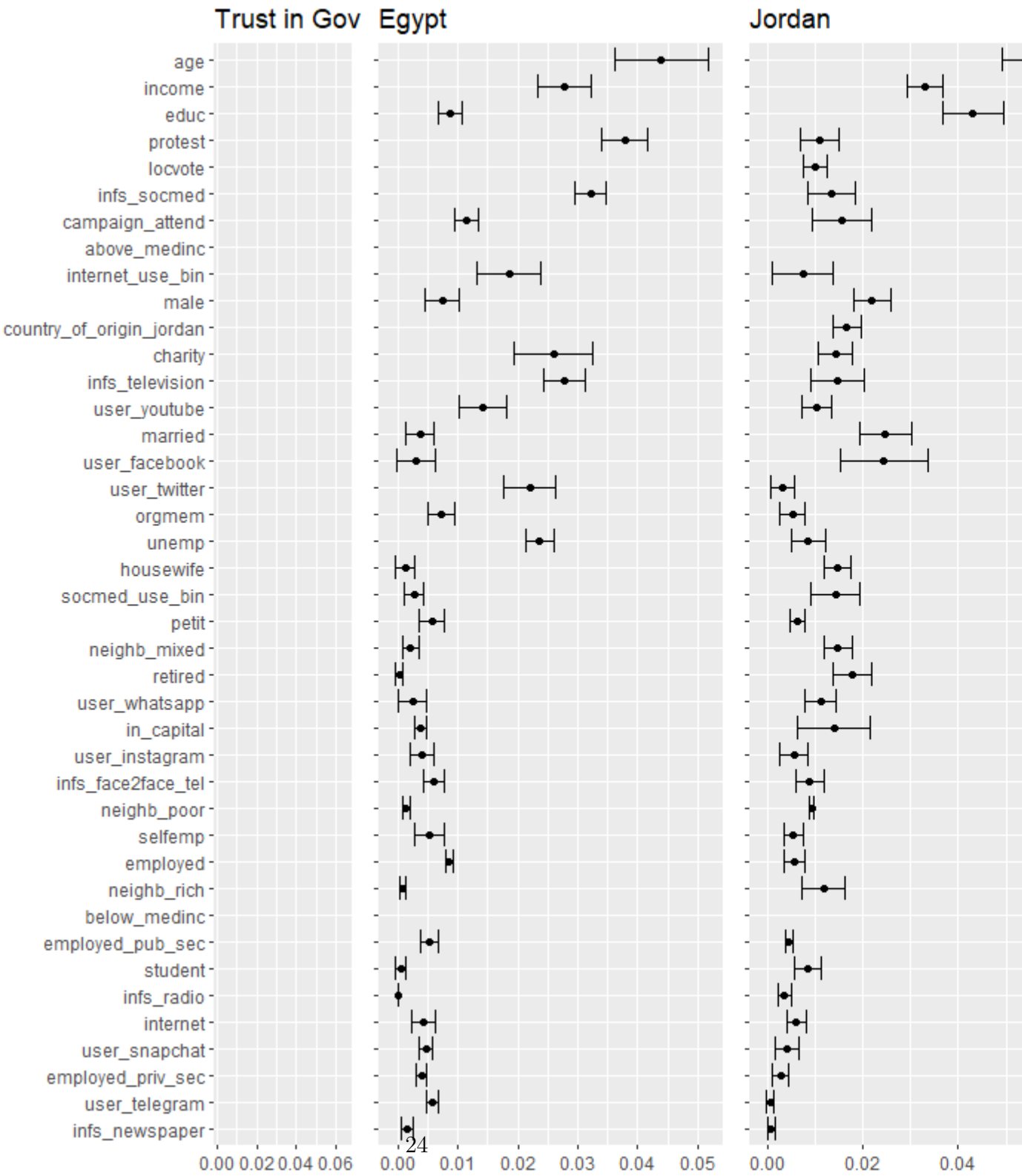
school to maximize the difference in the y variable between the resulting groups. The right daughter node is further partitioned by graduate education or none, then again by Facebook use. The daughter nodes continue partitioning until the gain of another partition is below an arbitrary stopping threshold. The observations are then sorted into terminal nodes with different average outcome values.

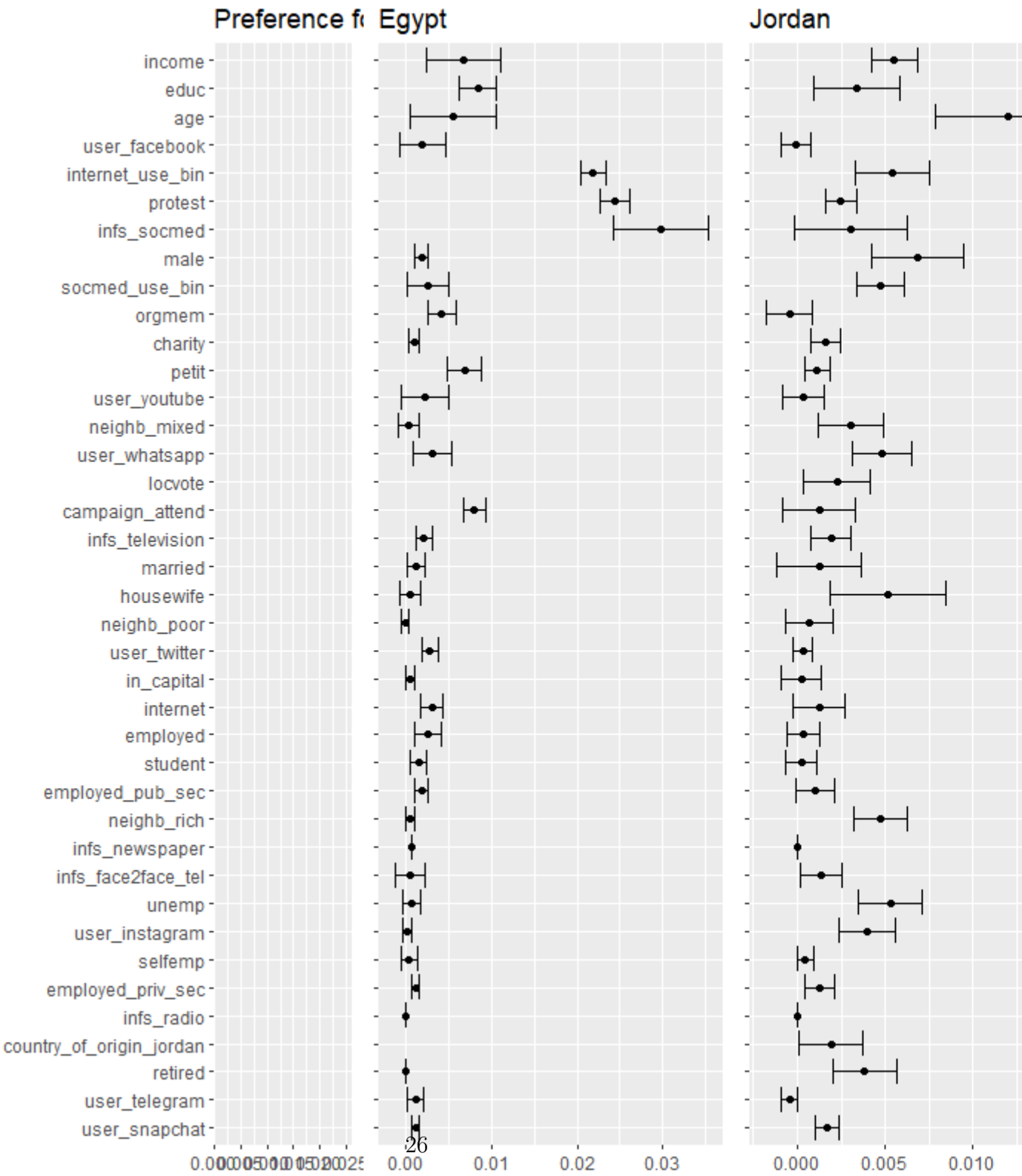
Individual decision trees have the advantage of being both relatively easy to interpret and a closer approximation of human decision making. However they lack the predictive accuracy of other models, are sensitive to small changes in the data, and tend to overfit, especially as the tree gets deeper and decisions are made on smaller and smaller subsamples of the data.

The Random Forest model improves on decision trees by building many trees, each created from a sub sample of the data, then merging them together and taking the most common predictions across each terminal node for each tree. Each tree is based on a random subset of the independent variables. The added variation caused by training on different samples of the data and utilizing various subset of the variables allows the model to create a more accurate and stable prediction upon taking the average of all the trees. The use of random subsets which are later combined also helps prevent overfitting . Additionally, the algorithm was tuned on several parameters in order to maximize its predictive power. These parameters are the maximum depth of each tree, the number of trees to build, and maximum number of features to consider for each tree. The ideal parameters differed across countries and y variables. For a more comprehensive work on random forests and decision trees, see **?**.

1.6 Cross Validation

Random forest models are not intuitively interpretable; "A forest of trees is impenetrable as far as simple interpretations of its mechanism go" (Shepsle, 2009). To understand the relative importance of different values we use a permutation test to measure the decrease in accuracy after permuting (randomizing) each variable. For each variable we randomize the values and rerun the model with the information-less random variable, comparing the increase in the prediction error with the new permuted values. In theory permuting a critical variable causes a large decrease in accuracy, while irrelevant values should have no change.


Trust in Gov
Egypt
Jordan
age
income
educ
protest
locvote
infs_socmed
campaign_attend
above_medinc
internet_use_bin
male
country_of_origin_jordan
charity
infs_television
user_youtube
married
user_facebook
user_twitter
orgmem
unemp
housewife
socmed_use_bin
petit
neighb_mixed
retired
user_whatsapp
in_capital
user_instagram
infs_face2face_tel
neighb_poor
selfemp
employed
neighb_rich
below_medinc
employed_pub_sec
student
infs_radio
internet
user_snapchat
employed_priv_sec
user_telegram
infs_newspaper
0.00 0.02 0.04 0.06
0.00 0.01 0.02 0.03 0.04 0.05
0.00 0.02 0.04


Preference f
Egypt
Jordan
income
educ
age
user_facebook
internet_use_bin
protest
infs_socmed
male
socmed_use_bin
orgmem
charity
petit
user_youtube
neighb_mixed
user_whatsapp
locvote
campaign_attend
infs_television
married
housewife
neighb_poor
user_twitter
in_capital
internet
employed
student
employed_pub_sec
neighb_rich
infs_newspaper
infs_face2face_tel
unemp
user_instagram
selfemp
employed_priv_sec
infs_radio
country_of_origin_jordan
retired
user_telegram
user_snapchat
0.00
0.01
0.02
0.03
0.000
0.005
0.010

1.7 Trust in Government Results

We depict the permutation importance measures from the random forests for each y variable in figures 1 and 2. The variables are listed in order of their unweighted average importance.

The most striking result is the variance between countries. Protest attendance is the second most revealing variable in Kuwait, but is within two standard deviations of the null in Sudan. Twitter is use is predictive in Egypt, but near null in Jordan, Morocco and Kuwait. Charity and unemployment are highly relevant only in Egypt.

Income, age and education have high importance in all countries. This is not surprising given how those variables strongly shape life experiences, economic interests, and political knowledge. Gender and charitable contributions also achieve consistent significance across the cases. Surprisingly, marriage and employment sector all perform relatively poorly. Although they are usually distinguishable from noise, the employment variables rarely pass .01 units of importance.

As expected, the political behavior variables perform well. Protest attendance was the third most important variable on average, and voting the fourth. Organization membership performed poorly, potentially due to vague wording. Petitioning, a rarity in our countries of focus, also preformed poorly. The highest petitioning rate occurred in Morocco at 20%. Campaign attendance had varied performance, highest in Algeria. Voting and campaigning in the authoritarian countries is usually more clientelistic than ideological, which could explain the inconsistency (Blaydes, 2006).

The informational and internet variables perform moderately. Use of social media as information source is the 6th most important on average, and number of hours of internet use is 9th. In Kuwait social media news sources is second, and in Egypt third, but in the other three states it performs poorly. The various apps all have moderate or low performance.

In general, these results suggest that trust or attitude imputation is sensitive to diverse set of variables. This validates Qiang's concern about the variety of data sources being gathered by repressive actors (2019).

1.8 Regime-type preference results

The non-political and non-informational variables perform more poorly at predicting ideology. While income, age and education remain the strongest predictors on average, their lead is greatly reduced. The average value for age dropped from .07 to .015, and income is reduced by half. Education has a more moderate reduction, perhaps because it strongly interacts with information consumption. The employment and location variable have moved entirely to the bottom half of the distribution, and the highest financial variable other than income is charitable contributions.

The Variables relating to political behavior and information consumption perform comparatively well. Facebook use is now the fourth most important variable, and is most important in Morocco. Hours per week of internet use is now the fifth most important on average, followed by protesting and social media information sourcing. The starkest illustration is Egypt, where hours of internet use, protest attendance, and social media use are each 4 times the value of any other variable. None of the

non-political and non-media variables pass .01 in Egypt.

Except for Facebook, the time and information consumption on social media is more relevant than the particular app. Facebook use was the most predictive variable in Morocco, but had moderate performance in other states. Twitter has low relevance even in Egypt where 439 respondents use it. Telegram and Snapchat have low importance because they are rare in all samples.

Surprisingly, Whats-app is not highly predictive, despite end-to-end encrypted apps being grounds for detention in Belarus and China. We do not reject the null in Algeria and Morocco and performance in Algeria and Sudan is modest. This demonstrates that once secure apps become popular across the population, it cannot identify opposition ideology even with nonlinear models. This observation is critical to countering state surveillance and we return to it in the conclusion.

These results suggest that non-political data has weak relevance to ideology in autocracies. Figure three summarises this result. One explanation is that most citizens of non-democracies do not form strong ideological positions on the basis of life experiences and material interests, perhaps because they lack safe avenues for political expression. While life experiences condition attitudes toward the state, they do not affect ideology. Ideology does have a two-way relationship with news consumption and political expression. More ideologically minded individuals seek out news and news changes ideological positions.

Type of Behavior	Trust in Gov	Regime-Type Preference
Political behavior	High	High
Non-Political behavior	High	Low

This suggests the danger of regimes imputing opposition ideology from non-political data is lower. Unfortunately data limitations prevent assessing the variety of behaviors states now seek to monitor from purchases to finances to shipping and geolocation. It remains possible that accumulating enough low-importance data could overcome the low value per observation, but marginally less likely.

1.9 Conclusion

For the moment, predictive repression is rare and peripheral to regime survival (Greitens). If the tactic spreads and becomes effective at preventing ressistance, pro-democracy actors should consider countermeasures. This section discusses how pro-democracy actors can effectively respond to predictive repression in light of our results.

Treaties and international norms are unlikely to stop predictive repression if leaders value it. The Convention Against Torture attacked a similar problem. While many autocracies ratified the convention including China, Egypt, Pakistan and Ethiopia, they did not change their behavior (Lupu, 2013).

Pro-democracy actors should instead directly influence the costs and benefits of repressive strategies. As reviewed in section 1.3, the current pattern of predictive repression suggests that data gathering is expensive and the results have high false positive rates. Because repressive actors have few and expensive data sources, protecting the most valuable data can make predictive repression non-viable.

Our results suggest that increasing the encryption of personal communication is a high leverage intervention. Phone use and news sources are important variables even

in our coarse sample. A variety of end-to-end encrypted messaging and social media apps are used by citizens of autocracies which make state interception impossible or prohibitively expensive. Telegram and Whatsapp are the most prominent examples.

While Belarus and China have targeted encrpted app users, our results point to a simple solution. Whatsapp had only the 14th highest average var imp for ideology, below YouTube and Facebook. Whatsapp was also used by more than 40% of respondents in each country except Algeria. This demonstrates that if only opposition members use secure communication, regimes can exploit app selection, but if secure communications are popular enough they no longer signal any political position.

Popularizing and normalizing secure applications has the double benefit of protecting the particular communication and camouflaging a privacy-hungry opposition. Fortunately, secure communication is becoming the norm in most autocracies. Whatsapp, Telegram and Facebook messenger all offer encrypted communication (as an option in Facebook's case). ¡describe problems with facebook¿

Our results suggest that a reader's choice in news is valuable information. Whether respondents received news mainly from social media was the 7th most predictive variable for ideology. Pro-democracy actors should worry about autocrats monitoring information consumption, not just expression. States may already be monitoring site visits, follows and or censored TPS packets per internet user. Even if states lack training data, they could simply label opposition websites using the same apparatus as web censorship.

VPN subsidies would both protect encrypted apps and prevent information con-

sumption monitoring. Iran has attempted to block Telegram, forcing users to either use virtual private networks (VPNs) or unsecured "forks" of Telegram. The attempt failed as Telegram retains a huge share of internet traffic (60 by some estimates). Nonetheless, such attacks remain a major threat to the privacy of ideology. A promising response is to temporarily subsidise VPN traffic in any country that launches a censorship attack on secure communications. Telegram CEO Pavel Durov has already piloted such a subsidy (cite fix)(Who Votes in Authoritarian Elections & of Voter Turnout, noa). If citizens respond to banning by adopting VPNs, the regime also loses future data access and embarrass itself.

Alternatively, server side software could hide visitor origins. The success of the Geneva genetic algorithm in evading in-network monitoring (Bock et al., 2019) may point to a cost effective solution. Information consumers would not need to install software on their own devices. In the low-tech direction, receiving radio transmissions remains unmonitorable. Since the second World War, dictatorships have failed to prevent citizens from receiving outside radio broadcasts. If citizens fear their phone use is monitored, outside radio broadcasting will remain safe and viable. US-funded station Radio Free Asia continues to broadcast today.

Geolocation data remains a serious concern. While the residence variables performed poorly, protest attendance was highly predictive. The Ukrainian incident already shows that regimes will use geolocation data (Walker, 2014), and Singapore has purchased Chinese facial recognition software to identify protesters(Feldstein, 2019).

Finally, predictive repression has spread slowly relative to other digital tactics

such as automatically detecting dissenting statements on social media, disinformation campaigns and censorship. This suggests that it is not yet altering the costs and benefits of repression. One possibility is that the long theory of change and steep technical challenges compromise predictive repression. More research is needed to understand why predictive repression is so unpopular and how marginal improvements in accuracy alter regime stability.

References

Telegram: Contact @durov (https://t.me/durov/77).

Bausch, Andrew W (2018) Coup-Proofing and Military Inefficiencies: An Experiment. *International Interactions* 44(1): 1–32 Publisher: Routledge _eprint: https://doi.org/10.1080/03050629.2017.1289938 (https://doi.org/10.1080/03050629.2017.1289938).

Berwick, Angus A new Venezuelan ID, created with China's ZTE, tracks citizen behavior (https://www.reuters.com/investigates/special-report/venezuela-zte/).

Blaydes, Lisa (2006). Who Votes in Authoritarian Elections and Why? Determinants of Voter Turnout. In: University of California, Los Angeles.

Bock, Kevin; George Hughey, Xiao Qiang & Dave Levin (2019). Geneva: Evolving censorship evasion strategies. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security , 2199–2214.

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

Crabtree, Charles; Holger L Kern & David A Siegel (2020) Cults of personality, preference falsification, and the dictator's dilemma. *Journal of Theoretical Politics* 32(3): 409–434 Publisher: SAGE Publications Ltd (https://doi.org/10.1177/0951629820927790).

Davenport, Christian (2007) State Repression and Political Order: 25.

Feldstein, Steven (2019) How Artificial Intelligence is Reshaping Repression. *Journal of Democracy* 30(1): 40–52 (https://muse.jhu.edu/article/713721).

Frantz, Erica; Andrea Kendall-Taylor & Joseph Wright Digital Repression in Autocracies: 54.

Gerdžiūnas, Benas (2020) Revolution will be Telegrammed: Social media channels drive Belarus protests Section: Digital (https://www.euractiv.com/section/digital/news/revolution-will-be-telegrammed-social-media-channels-drive-belarus-protests/).

Gohdes, Anita R (2020) Repression Technology: Internet Accessibility and State Violence. *American Journal of Political Science* 64(3): 488–503 (https://onlinelibrary.wiley.com/doi/10.1111/ajps.12509).

Gregory, Paul R (2009) *Terror by Quota: State Security from Lenin to Stalin:(an Archival Study)*. Yale University Press.

Greitens, Sheena Chestnut; Myunghee Lee & Emir Yazici (2020) Counterterrorism and Preventive Repression: China's Changing Strategy in Xinjiang. *International Security* 44(3): 9–47 Publisher: MIT Press.

Hassan Abdel Zaher Bitter options for Egypt as tax evasion persists | Hassan Abdel Zaher (https://thearabweekly.com/bitter-options-egypt-tax-evasion-persists).

Iyengar, Shanto & Sean J Westwood (2015) Fear and Loathing across Party Lines: New Evidence on Group Polarization: FEAR AND LOATHING ACROSS PARTY LINES. *American Journal of Political Science* 59(3): 690–707 (http://doi.wiley.com/10.1111/ajps.12152).

Kahan, Dan M (2013) Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making* 8(4): 18.

Kuran, Timur (1995) The inevitability of future revolutionary surprises. *American Journal of Sociology* 100(6): 1528–1551 Publisher: University of Chicago Press.

Nathan, Andrew J (2020) The Puzzle of Authoritarian Legitimacy. *Journal of Democracy* 31(1): 158–168 Publisher: Johns Hopkins University Press (https://muse.jhu.edu/article/745962).

Polyakova, Alina & Chris Meserole Exporting digital authoritarianism: The Russian and Chinese models: 22.

Qiang, Xiao (2019) President XI's Surveillance State. *Journal of Democracy* 30(1): 53–67 (https://muse.jhu.edu/article/713722).

Quinlivan, James T (1999) Coup-proofing: Its practice and consequences in the Middle East. *International Security* 24(2): 131–165 Publisher: MIT Press.

Regan, Patrick & Errol Henderson (2002) Democracy, threats and political repression in developing countries: Are democracies internally less violent? *Third World Quarterly - THIRD WORLD Q 23*.

Robinson, Darrel & Marcus Tannenberg (2019) Self-censorship of regime support in authoritarian states: Evidence from list experiments in China. *Research & Politics* 6(3): 2053168019856449 Publisher: SAGE Publications Ltd (https://doi.org/10.1177/2053168019856449).

Tan, Netina (2020) Digital learning and extending electoral authoritarianism in Singapore. *Democratization* 27(6): 1073–1091 (https://www.tandfonline.com/doi/full/10.1080/13510347.2020.1770731).

Volpi, Frédéric (2020) Algeria: When Elections Hurt Democracy. *Journal of Democracy* 31(2): 152–165 (https://muse.jhu.edu/article/753201).

Walker, Shaun, Oksana Grytsenko (2014) Text messages warn Ukraine protesters they are 'participants in mass riot' Section: World news (http://www.theguardian.com/world/2014/jan/21/ukraine-unrest-text-messages-protesters-mass-riot).

Wang, Maya (2019) *China's algorithms of repression: Reverse engineering a Xinjiang police mass surveillance app*. New York: Human Rights Watch OCLC: on1102056042.

Wilson-Chapman, Amy China Cables | IJOP Daily Bulletin 14 English (http://www.documentcloud.org/documents/6558506-China-Cables-IJOP-Daily-Bulletin-14-English.html).

Xu, Beina & Eleanor Albert (2014) Media censorship in China. *Council on Foreign Relations 25*: 243.

Biographical statement