An example Journal of Peace Research paper typeset in LaTeX

Name of Author
University of Author
Word count: 9000

February 11, 2021

Abstract

Repressive states are experimenting with machine learning to identify citizens with grievances who might join dissident movements. While they are already implemented in China, the viability of existing classification techniques and citizen data remain unassesed. We train classification and regression algorithms on trust in government and democracy preferences in 6 Arab countries using Arab Barometer data. Random Forest classification does outperform logit analysis, but the gains are moderate. Results suggest that the legitimacy costs of false positives will limit applicatiosn for the near future. Accuracy and cross validation suggests that detailed personal information about movement or social media use are necessary for effective applications.

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. However, these "dissidents" had been first selected for interrogation by an algorithm, not human identification. Officials were instructed to "put measures in place according to classifications (...) different types of tags pushed out by the "integrated" platform " (China Cables).The leaked bulletin states that tens of thousands of tagged persons were investigated and sent to reeducation camps.

The CCP's integrated platform is the first identified example of machine-targeted repression. The "integrated" platform refers to a database of individual movements, administrative data and phone use in Xinjiang and an algorithm for "tagging" persons for investigation and often detention. The platform collects information from facial recognition, ID's, checkpoints, health, banking and legal records. Citizens are required to keep an app on their phone which records their activity and location. Tagged persons can be detained indefinitely (Feldstein, 2019). While Xinjiang is the most extreme modern example, Chinese firms are selling monitoring and facial recognition software in Pakistan, Singapore and Zimbabwe (Feldstein, 2019). ¡More on spread¿

Reports from china have stressed the breadth of data collection. Darren Blyer reported that "Along with DNA collection, they [the state in Xinjiang] are creating a registry of fingerprints, blood types, voice patterns, facial imagery – all of which will be correlated to ethnicity, employment, gender, age, foreign travel history, household registration, individual and family criminal history, and religious practice" (Kuo,

2017). Observers are concerned about similar data gathering on non-Uighur Chinese citizens in the social credit system. "Through the SCS, Chinese authorities can bundle with a citizen's national ID code information about matters ranging from tax payments, personal finances, and business registrations to traffic violations and more. Once credit information is linked to documents establishing the personal status of citizens, such as household registrations and ID cards (...)" (Qiang, 2012). However, the value of this data is questionable, and states may gather such data to give a false impression of omniscience. Therefore expressing undue concern about gathering of non-political data risks enhancing a repressive bluff.

Digital repression offers the possibility of increasing effectiveness and reducing the cost of repression. The core goals of repression remain the same: to raise the costs of disloyalty, to identify the opposition, and to prevent collective action against the regime. The costs of repression include the risk of backlash as unpopular punishments and surveilance enlarge the political opposition. Traditional repression also requires the regime pay, train and motivate thousands of party members or spies to surveil citizens. Furthermore, repression disincentivizes collective action, such as forming companies, universities and other associations. That atomization of society has negative consequences for growth (Acemoglu). Digital tactics can improve repression by reducing any of these costs.

Franz et al. catalogued five 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 misin-

formation to disrupt collective action. Each mechanism is distinct and deserves a detailed assessment of its viability and likely affects. We know little about the effectiveness of digital repression. Most existing work categorizes and tracks the spread of new emerging technologies. Franz et al. constructed a global dataset of digital repressive capacity, but the technology is too new for a statistically identifiable affect on macropolitical outcomes.

This article assesses the viability of using mass surveillance of citizen behavior to impute grievances in the general population. States currently have two paths to identify citizens with grievances. One is to monitor political statements and traffic to opposition-associated websites. Advances in natural language process are cheapening mass monitoring of online speech. However, a majority of citizens are not politically active or politically interested, particularly in repressive states. Therefore these strategies can identify potential first movers and prevent revolutionary action (Kuran, 1995), but not the preferences of a majority of non-elites. We lack a research method to assess natural language processing for grievance detection.

We assess "algorithmic repression" (AR), imputing the grievances of a majority of citizens. Under AR, the regime acquires detailed information about the behavior of all citizens. This includes routine information such as taxes, employment, biopolitics and service use. States can also measure political actions such as petitioning, voting and party organizing. They may plausibly require information providers, from internet utilities to newspapers, to collect a personal identification number for SIM card or subscription. China is advancing this collection even further by combining facials and voice recognition data and information on phone and voice use (Qian,

2019).

The state must then accurately identify some citizens with grievances, preferably in a non-biased way. The state may identify grievance-bearers through its traditiofnal informant method, through online activism, through monitoring protest attendance or even through direct surveys. The state can then use widely available machine learning algorithms to impute the political preferences of all citizens. We refer to this process as algorithmic repression (AR)

In Xinjian AR has been used to target persons for indefinite detention or criminal investigation. We do not expect imprisonment to become a common application for AR, given the false positive rates. Secondly, punishing a significant portion of the population for guesses about their beliefs violates norms of fairness and reciprocity, jeopardizing state legitimacy. Instead, civil liberties restrictions pose lower legitimacy costs (Davenport,). Therefore subtler forms of discrimination or soft sanctioning with algorithmic repression should be worthwhile at achievable specificity. We discuss applications of preference imputation in discrimination in military training, public sector hiring and internet shutdowns during mass protest.

However, those strategies depend on grievances being imputable from non-opposition behavior. We assess grievances imputability by simulating algorithmic repression in 5 Arab countries. We select from the Arab Barometer survey all information that the surveillance state could plausibly gather, from newspaper subscriptions to income to voting behavior. We then train a model on the trust in government and regime type preferences of a minority of respondents. We produce both a logit and a random forest model for each country and assess its predictive power.

While random forest models do outperform the logit and base rate models, the improvements are modest. ¡something about how different the logit and RF models perform¿. The models may perform modestly because non-opposition behaviors have little relation to grievances, or that citizens in autocracies do not construct such dangerous preferences. However, in Egypt a ¡score achieved¿. This demonstrates variance in predictability between countries.

The viability of AR applications depends on the payoffs for false positives and false negatives in the particular repressive act. Therefore we construct three scenarios of repressive action to compare the models. The instances include military recruitment (80% confidence threshold), selective internet shutdowns (50% confidence threshold) and public sector recruitment (20% confidence threshold). In each scenario we compare random selection, assuming the same grievance probability for each citizen, a logit model and the random forest model. Results suggest that variation in the relative costs of false positives will limit near-term applications.

Cross validation analysis shows social media use is highly predictive. In all models internet use or social media use is significant and in the top three most useful variables. The input data recorded only the extent of social media use, not sites visited or group membership. This suggests that highly detailed surveillance of citizen-internet interactions could allow precise classification of grievances.

This article first reviews the role of belief imputation in authoritarian survival. Next it describes the research design for simulating AR. It then briefly compares the performance of different models generally and in the game scenarios. It briefly revies variable importance results. The conclusion speculates on expected early applications

of AR and advocates for improving digital privacy access for citizens in autocracies.

Imputed Beliefs

Imputing the beliefs of others is a common practice across all polities. Actors prefer coalition partners who share their beliefs and preferences (Kahan 2013; Hanson and Simmler, 2018). Cobelievers are more loyal partners than nonbelievers, even for empirically viable beliefs such as climate change, policy impacts or the affects of political institutions. Untrue believes can even represent a *stronger* signal of loyalty because a non-member would not endorse a belief contrary to evidence. We use belief and preference interchangeably because believing a policy is beneficial and preferring that policy are the same for our argument.

Even in democracies, discrimination by belief can be strong. A 2010 survey found that 30 percent of Republicans and 24 percent of democrats prefer their children not marry across parties (Pew, 2014). Iyengar and Westwood find that American partisans are 60 percentage points more likely to award a scholarship to a copartisan (partisan bias was much lower in the past however). Timur Kuran argues that in autocracies norms against dissent are enforced socially, not just by repression specialists (although the state also discriminates) (1995). People highly prize membership in coalitions (identity groups, ideological camps, political parties and factions) in all polities as it offers security, connection, and other rewards. People thus have strong incentives hold socially-approved beliefs, especially in autocracies. On surveys, "political intimidation in China is real, of course, bit its more likely effect is not to cause people to lie about what they think, but rather to shape the thoughts they have"

(Nathan, 2020).

People respond by adopting and dissembling. In socially adaptive belief theory (SAB), actors respond by simply adopting socially beneficial beliefs. In SAB the actor sacrifices epsitemically accurate beliefs. In preference falsification, they disguise their preferences. Falsifying preferences retains an accurate underlying map of the world, but is dangerous because humans possess mind reading powers (through social cues, hesitation, speech patterns) (Williams, 2020). SAB is most effective when accurate beliefs are not valuable to the individual, such as in voting, revolution and punditry, but consequences for disloyalty are high. Coalition politics thus creates an arms race between discriminating coalition formers and potential members. Each actor wishes to know the true beliefs and preferences of others, while others prevent disagreement from surfacing (Kuran, 1995).

An authoritarian regime can secure itself by placing loyalists in all positions of power. They also prefer to increase collective action capacity of loyalists and decrease that of dissenters. However, this goal is much easier stated than achieved. Firstly, humans are skilled at this game. During human evolution coalition politics was critical to accessing safety and mates (De Waal, 2007). In the United States political "ideology" predicts believes about climate change better than scientific knowledge, especially among 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). But the East German people successfully hid their beliefs until a sudden cascade caught the world by surprise in 1989. If the regime empowers many persons with socially adaptive beliefs or false preferences, it is in danger of a cascade. As a few

persons of strong conviction enter opposition, the social rewards for loyalty decline (Kuran, 1995). The reduced loyalty incentive expands the opposition, creating a self-enforcing cycle that expands the opposition. This is true for preference falsifiers, socially adaptive believers and even true loyalists willing to preference falsify to an overwhelming opposition. Thus the regime desires the most stable underlying beliefs, if imputable.

Creating a belief incentive system through an entire society presents deep practical challenges. Autocratic politics should select for leaders skilled at imputing the loyalty of potential followers, as this helps the would-be autocrat rise up. But even a skilled leader can only assess a few hundred close followers. Down the hierarchy she must create a principal agent chain of loyalty assessors. The political commissars in Soviet military units fulfilled this function. Modern autocracies use secret policy ministries, sometimes several (sultanistic regimes guy) or a single political party to recruit and indoctrinate member-informants (Geddes). Secret policy are expensive and many autocracies struggle to form strong parties despite their effectiveness. A primary role of both institutions is to extend the imputation of loyal beliefs down through society and implement discrimination.

The CCP's experiments in Xinjian raise a serious question if mass surveilance and machine learning can improve the imputation of loyal beliefs. Gathering information on activities from taxes to movement to news consumption is expensive, but may be cheaper than the lavish secret police of the sultanistic regimes and the CCP. Algorithmic repression should 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 (Jung at al. 2017) find that regression-derived checklists outperform intuition in bail setting)
- Makes belief-based discrimination more acceptable to the public.

In both competitive and non-competitive polities, citizens value government fairness highly (Nathan, 2010). The last Asia Barometer round found that "in nine of the fourteen countries, citizens give even more weight to government fairness than they do to economic performance" (Nathan, 2020; 162). Punishment or discrimination based on hunches about unstated beliefs violate widely-held justice norms. In elite politics coerced confessions or corruption allegations evade those norms. The legitimacy costs of AR-based discrimination in the general population are untested but expected to be large.

Research design

The research design simulates a future AR platform as closely as possible. We assumed that a state implementing AR would begin by routinely monitoring citizens political and economic information and theirn information sources. They would aggregate information at the individual level. They would then invest in finding the true beliefs of a small group of citizens as training data. Finally they would use the most effective available algorithm to impute the preferences of the broader polity.

To simulate imputability we used a "split, train, test" process. The input variables are divided between X-data, which is used for imputation, and Y-data which

contains the true outcome information (in this case grievances). The observations (respondents) in each data set are randomly split between a larger training and smaller testing dataset. The predictive algorithms use both the X and Y data in the training set to construct their models. The model then uses the X data for the test set to impute Y values. We then analyze the accuracy of the model in predicting the Y-values.

Data

Arab Barometer (AB) survey data was used for all internal samples. The AB covers a variety of states which rely on repression and limit mass political participation, making them plausible future AR users. It provides detailed information about political beliefs (for classifying grievances), political and economic behavior and information consumption. The survey questions vary only slightly between countries. Each country had at least 2000 responses suitable for use (except Kuwait, which had 1400 usable responses).

Our Y-variables were inspired by recent work on citizen grievances in repression (Gregiry et al.; Rozenas et al; Aspinall). We interpret a grievance as a preference for anti-regime action or for regime change. Following formal theory, a grievance is the positive dummy variable (Rozenas). 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", otherwise no grievance. 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. The majority of grievance bearers did not report democracy as a high priority issue.

The regime-preference variable has greater specificity, as it excludes low-trust citizens who still prefer the current regime. However, not all revolutionary movements have democratic goals and the regime must defend from autocratic challengers as well as democratic movements.

We included in the X variables all behavioral information that could plausibly be collected by the state. Behaviors were included if a high-capacity autocracy such as China could gather the information, even if the state of data collection lacks the capacity or interest. The resulting net is quite wide. We included

- Demographics: age, gender, religion, marital status, neighborhood income, residence in the capital province
- Tax information: income, retirement status, employment status (public, private, self or none), student status
- Social behavior: education level, charitable contribution
- Access to information: newspaper reading (Borzyskowski and Kuhn), social media use (hours per day), radio use, and television use, internet access

- Political behavior: petitioning, protesting, campaign rally attendance and voting in local elections

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 algorithmic 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.

Certain questions were only usable in some countries. 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. 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 social desirability bias or preference falsification. However, even respondents in the most repressive country (Egypt) expressed majority agreement that democracy is the best political system, which indicates little belief

censorship. 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.

Data Analysis

¡Youssef section here¿

Results

¡section about the ROC curves of the models¿

Usage Cases

The added value of AR depends strongly on the payoffs to the regime of false positives and false negatives. A repressive act has a positive payoff when applied to a true positive, through deterring or disabling opposition. Repressing a false positive has a negative payoff due to decreasing deterrence, increasing opposition, decreasing state legitimacy. Of course often both have negative payoffs and rarely both are positive (Rozenas). Imputation is only relevant when the payoffs differ in sign. When the payoffs differ in sign, the expected utility of a repressive act depends on the subjective probability (SP) of a grievance to the regime. When the payoffs for a false negative are much larger than a false positive, a high SP threshold is rational, and vice versa.

A true positive also has a legitimacy cost, which is usually higher than the value,

hence the globally low level of repression. Because repression is a rational strategy for political survival, we expect it is highly contingent on the threats and resources facing the regime. During the great purge, Stalin advocated for reporting disloyal persons at a 5% true positive / false positive ratio: "Every communist is a possible hidden enemy. And because it is not easy to recognize the enemy, the goal is achieved even if only 5 percent of those killed are true enemies" (Gregory et al. 2007). Obviously this is a rare extreme of behavior (and is unimplementable because such a threshold requires repressing the entire society with accurate calibration). The relatively low current levels of state violence imply that modern autocracies face higher costs for false positives. Edel and Josua show that the Egyptian and Uzbek governments targeted opposition protesters for massacres and afterward argued that no "loyal" or "peaceful" citizens were killed. Th regime therefore is unwilling to legitimate accidental violence against loyalists, even while openly killing hundreds including women and children. Even under optimistic assumptions, the false positive rates of AR would be unacceptable for violent repression for regimes even minimally concerned about legitimacy.

Restrictions on civil liberties incur lower legitimacy costs than political terror. Because legitimacy costs constrain AR, we expect initial applications in discriminatory withdrawal of civil liberties or discrimination. Targeting distant ethnic minorities similarly lower legitimacy costs, which explains the Integrated Platform in Xinjiang (Rozenas).

To assess the added value of the random forest models, we designed three games in which a regime decides who to target for discrimination or rights withdrawal.

The analytical content of each game is a payoff matrix for true negatives adn true positives which specify a SP threshold. We include specific details for each game to clarify the real life implications of model specificity. In the first game, the regime discriminates in recruitment for military training (25% SP threshold). In the second game the regime shuts down internet access selectively during an uprising (50% SP threshold). In the third game the regime discriminates in public sector hiring (75% SP threshold).

In each scenario the government has previously built a national mass surveillance system gathering the above information with accuracy equivalent to the Arab Barometer survey. Suppose that distributors must report identity codes to the state on SIM cards, internet connections and newspaper subscriptions. The state includes tax data on employment, income and charities. The strength of this assumption depends on state capacity. For reference, over half of eligible Egyptians do not file their taxes (Zaher, 2018).

The government has accurate information about the preferences of ¡percentage¿ of the population and wishes to impute ¡split percentage¿ people's trust in government and regime-type preference. The regime may use a logit model, a random forest model, may take all, may take none, or may conduct a survey and assign all citizens the same probability.

Scenario 1: Military Training

The regime wishes to recruit for its officer training school from the general population. Ideally, the regime could trust all candidates and take the brightest students in the

country. But the regime knows some citizens hold grievances, and is concerned about pro-democracy officers betraying it. For each admittee with a grievance, the regime receives -3 points. For each admittee without a grievance, the regime receive +1 points from a more meritocratic army. If the regime identifies no persons with a 75% SP of pro-regime beliefs, it prefers to admit no one. It instead recruits from a separate, nepotistic pool, as in Saudi Arabia's Mujahideen (Quinlivian, 2006). The nepotistic outcome is suboptimal because the military is lower quality and the regime implicitly reveals its unpopularity to the people.

Table I. Model Performance when False Positives Cost 3 Times a True Positive

Country	Y	True	False	True_Frac	Max	BaseRate	Logit	RF
Algeria	DemPref	131.00	389.00	25.19	131.00	0.00	0	0.00
Egypt	DemPref	123.00	435.00	22.04	123.00	0.00	0	4.00
Jordan	DemPref	204.00	387.00	34.52	204.00	0.00	0	0.00
Morocco	DemPref	350.00	213.00	62.17	350.00	0.00	0	47.00
Egypt	TrustGov	155.00	403.00	27.78	155.00	0.00	0	2.00
Jordan	TrustGov	348.00	243.00	58.88	348.00	0.00	0	1.00
Kuwait	TrustGov	171.00	155.00	52.45	171.00	0.00	0	0.00
Morocco	TrustGov	350.00	213.00	62.17	350.00	0.00	0	50.00

Scenario 2: A Selective Internet Shutdown

A protest for regime change is cascading. The regime realizes that protester numbers are increasing exponentially and intervenes to stop them. It decides to shutdown internet access to reduce the collective action capacity of citizens with grievances. However, the regime wants to maintain the collective action capacity of citizens who privately prefer regime continuity. The dictators still wishes to tweet at her

supporters to organize counter-protests. In practice, there is also a signalling problem where internet shutdowns reveal the strength of the opposition and close businesses, incentivizing protest action. For simplicity we neglect that effect in this game. The regime receives +1 point for each citizen with a grievance that loses internet access. It receives -1 point for each citizen without a grievance targeted. Note that if the regime assigns all citizens a SP of 49% it leaves the internet on for everyone, and closes the internet for all at 51%.

Table II. Model Performance under Equal Value for False and True Positives

Country	Y	True	False	True_Frac	Max	BaseRate	Logit	RF
Algeria	DemPref	131.00	389.00	25.19	131.00	0.00	0	0.00
Egypt	DemPref	123.00	435.00	22.04	123.00	0.00	0	11.00
Jordan	DemPref	204.00	387.00	34.52	204.00	0.00	0	-9.00
Morocco	DemPref	203.00	360.00	36.06	203.00	0.00		20.00
Egypt	TrustGov	155.00	403.00	27.78	155.00	0.00	0	21.00
Jordan	TrustGov	348.00	243.00	58.88	348.00	105.00	0	70.00
Kuwait	TrustGov	171.00	155.00	52.45	171.00	16.00	0	43.00
Morocco	TrustGov	350.00	213.00	62.17	350.00	137.00	0	157.00

Scenario 3: Public Sector Hiring

The regime is recruiting for skilled positions in the public sector. Once again, the regime prefers to recruit citizens with no grievance. However, the regime hopes to prevent an uprising by providing high quality public services. Also discrimination by algorithm is unpopular, people will only tolerate accurate discrimination. Excluding a loyalist would alienate supporters and advertise incompetence in the interior ministry. Also recruiting from the nepotistic pool would give employees greater leverage

to shirk or engage in corruption. Therefore the regime reverses its threshold form the military training; it receives +1 for each non-grievance allowed and -1/3 for each grievance admitted.

Table III. Model Performance when False Positives Cost 1/3 a True Positive

Country	Y	True	False	True_Frac	Max	BaseRate	Logit	RF
Algeria	DemPref	131.00	389.00	25.19	131.00	1.33	0	29.33
Egypt	DemPref	123.00	435.00	22.04	123.00	0.00	0	41.67
Jordan	DemPref	204.00	387.00	34.52	204.00	75.00	0	74.00
Morocco	DemPref	203.00	360.00	36.06	203.00	83.00	0	0.00
Egypt	TrustGov	155.00	403.00	27.78	155.00	20.67	0	64.67
Jordan	TrustGov	348.00	243.00	58.88	348.00	267.00	0	266.67
Kuwait	TrustGov	171.00	155.00	52.45	171.00	119.33	0	117.00
Morocco	TrustGov	350.00	213.00	62.17	350.00	279.00	0	278.67

Scenario Results

The scenarios illustrate the practical limitations of AR at low specificity. In all cases the gains to using RF were modest. On average, the RF models were ¡some value¿ closer to the max score than the base rate and ¡some value¿ closer than the logit model. To the extent that logit models approximate human intelligence targeting, the RF provides a marginal improvement. AR may still have a niche in automating grievance imputation for mass communication attacks such as internet shutdowns. It would be prohibitively expensive to use human intelligence for a citywide internet shutdown.

Furthermore, the models only add value when the desired subjective probability is near the base rate of grievances. This result is best explained with Bayesian

probability theory. The base rate for individuals in each sample ranges from 20 to 65%. From that prior, an ideal reasoner updates off using the evidence assembled. With multiple updates the resulting posterior probability is given by

$$log(\frac{P(A|E)}{P(\neg A|E)} = log(\frac{P(A)}{P(\neg A)}) + \sum_{n=1}^{n(e)} log(\frac{P(A|e_n)}{P(\neg A|e_n)} \quad (1)$$

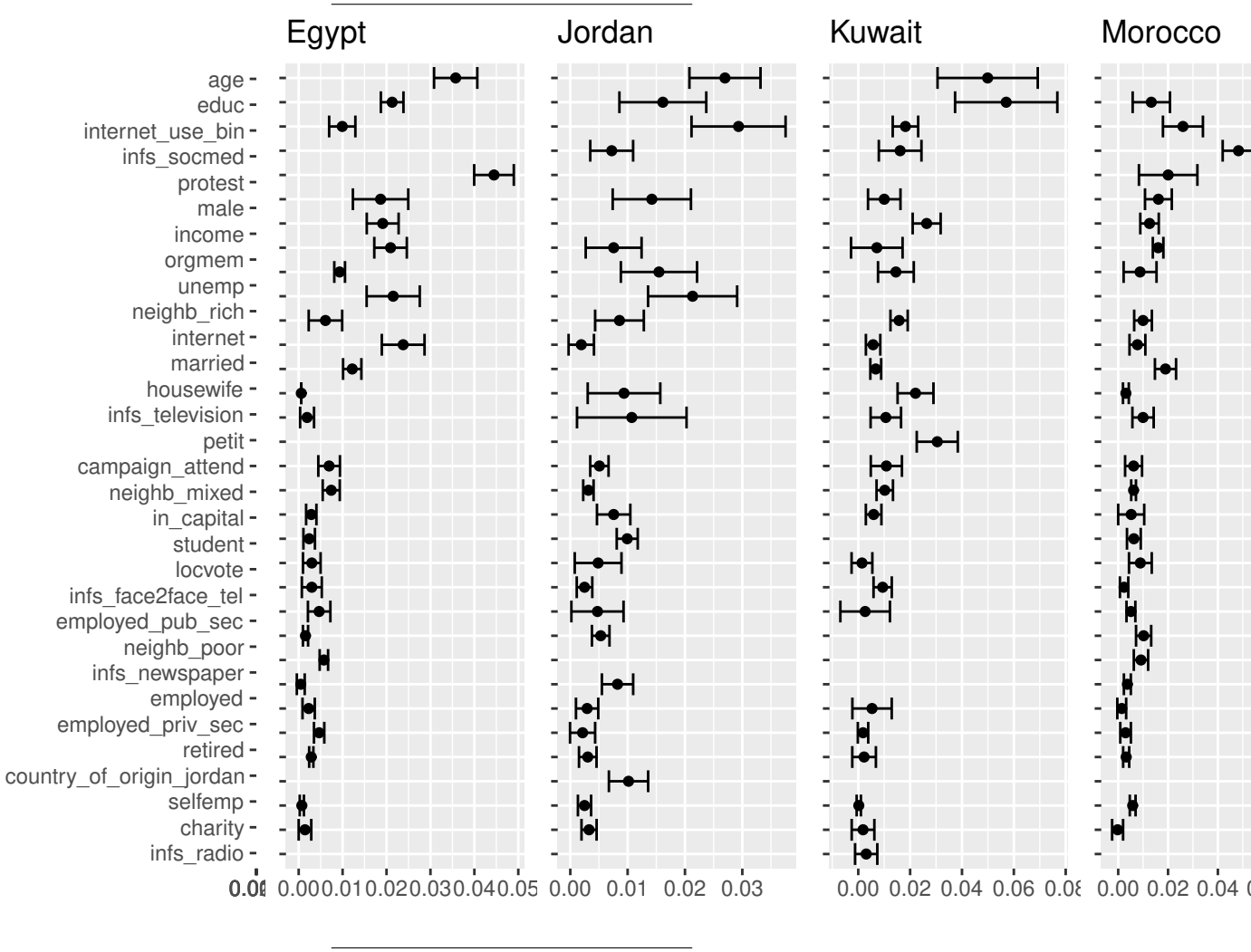
Where A is having a grievance, E is the total observed evidence, e_n is each additional (nonredundant) unit of evidence and $n(e)$ is the number of pieces of evidence. The evidence needed grows linearly with the distance in log odds between the target posterior probability and the prior.

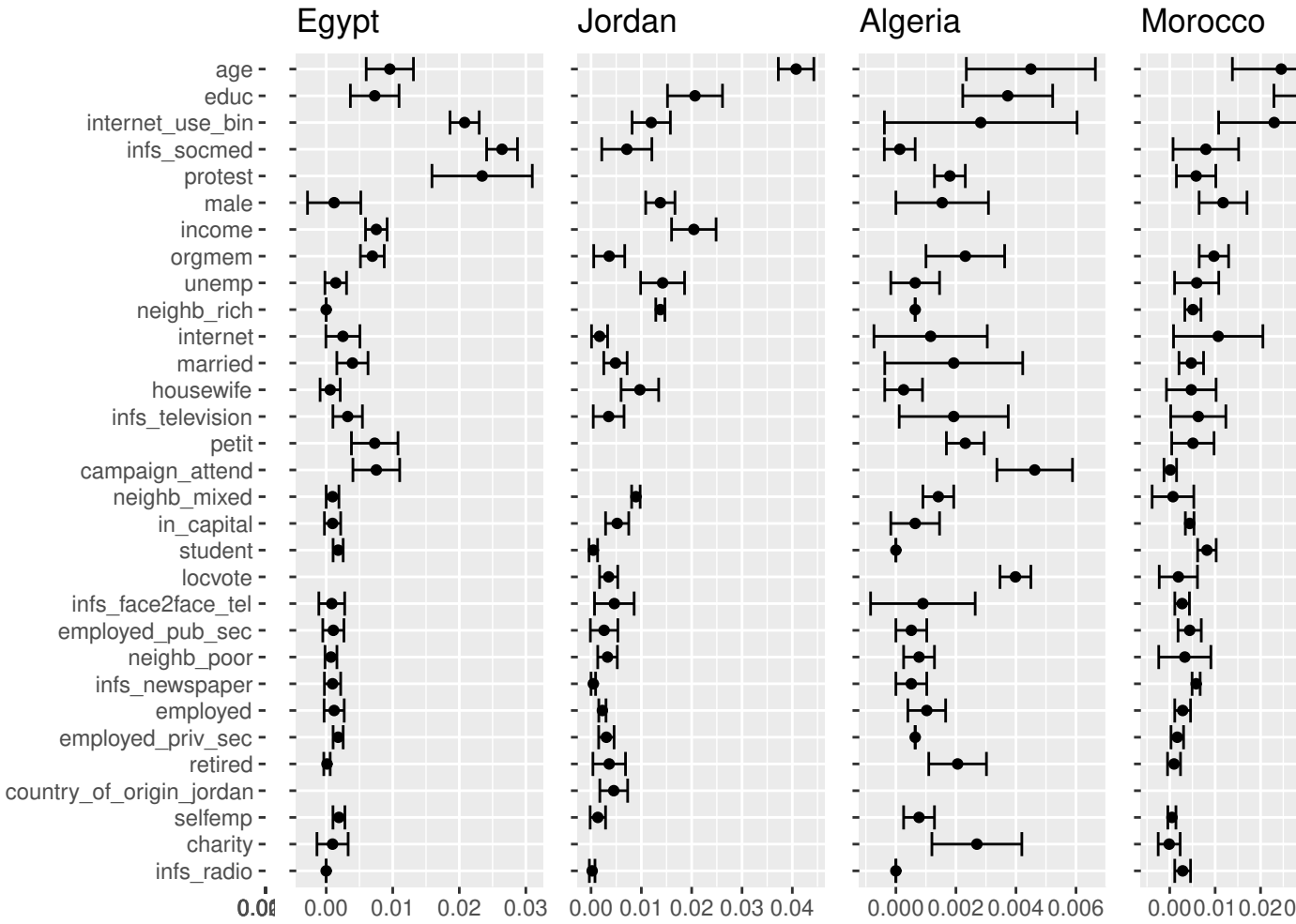
In other words, extraordinary claims require extraordinary evidence. Extraordinary here means a high likelihood ratio (something grievance bearers do frequently but non bearers do rarely). In a society where in general 25% of persons bear a grievance, claiming that a 75% of some subset do list an extraordinary claim that requires strong evidence. Regardless of model performance, few such individuals can be found without extraordinary evidence. Likewise in a society where 75% of persons bear a grievance, strong evidence is needed to identify a subset of whom 25% bear a grievance.

As a result, the sophisticated algorithms only provide added value when the payoff ratio of the repressive act is close to the base rate, unless strong evidence is available. In a society with a high base rate of grievances, the algorithm can identify few "safe" loyalists so the regime withdraws civil liberties from all. In a society with a low base rate, few certain grievance-bearers can be identified and the rational choices is to repress little. Absent stronger evidence and subject to legitimacy costs, AR has

limited added value.

Cross Validation





Both models rely on internet and social media use data. The variable "internet_use" refers to frequency of internet use. It is the third most useful in both models. The "infs" variables are dummy codes for "What is your primary source of information to follow the breaking news as events unfold?". Social media as primary source is the fourth most significant variable.

Unsurprisingly, having attended one or more protests in the past 3 years is the

5th strongest predictor. In Egypt it is the single most useful variable, despite only 14 % of Egyptians having attended a protest. Protest attendance is the only behavior directly caused by attitudes toward the government and regime type. Also, protest attendance is a fixture of regime surveillance long predating algorithmic repression. This result is worrying given the CCP's program to integrate cameras in public venues with facial recognition software.

However, campaign attendance is not an effective predictor, though positive for 20% of respondents. One explanation is that opposition, independent and pro-regime campaigns are all included in a single variable. In Egypt, where opposition parties are banned, it had a moderate positive affect. Alternatively, people may attend campaigns for social reasons unrelated to the target variables.

The most optimistic news is the many variables which did not improve imputation. Living in the capital, employment sector, school enrollment, radio listening, petitioning and self-assessed neighborhood wealth had little or no influence. This suggests much of the data the CCP is currently gathering in Xinjiang is of little use. Darren Blyder warned that integrated platform data is being "correlated to ethnicity, employment, gender, age, foreign travel history, household registration, individual and family criminal history, and religious practice" (Kuo, 2017). Employment, financial information and neighborhood had little value across all 5 countries. The value of criminal history, foreign travel and government service use should be investigated elsewhere.

Variables describing sources of information and political acts are most valuable to imputation. Variables describing non-political and non-informative actions are

of low value. ¡talk about declining marginal utility of data¿ Reports from China stressing the spread data collection into non-political activities overstate the danger, while internet surveillance remains a high risk. We discuss the policy implications below.

Discussion

This section begins by discussing the viability of AR with the data presented here. We argue that the legitimacy costs of false positives will severely limit AR unless much more predictive data can be found. We then speculate on possible alternative data sources. We conclude by arguing that AR repression is preventable. Facilitating digital privacy in authoritarian states should be a top priority.

Viability

The model performance suggests that legitimacy costs of false positives will constrain the viability of algorithmic repression. Civil liberties restrictions have lower legitimacy costs that political terror (Davenport), and will represent the earliest applications. False positives risk driving supporters away from the regime, such as when loyalists lose privileges aligning them with the regime. The west's intractable debate about fairness in ML bodes ill for legitimating false positives among loyalists (Binns, 2018; Chouldechova and Roth, 2018). Fairness determines legitimacy more than economic performance in 9 of 14 countries sampled by the Asia Barometer (Nathan, 2020). Furthermore false positives are compelling news stories which

demonstrate the information poverty of the regime. This decreases the perceived risks of opposition.

In certain applications false positives can simply be hidden. Western companies widely rely on poor quality natural language processing algorithms to sort applicants, creating a market for counter strategies (Jobscan, 2020). The underperforming algorithms are popular because rejected candidates are unaware of the reason. Low transparency could mitigate the legitimacy costs in hiring, university admissions and housing discrimination. Selective internet shutdowns cannot be hidden likewise.

Targeting isolated ethnic minorities can also lower legitimacy costs (Rozenas; 2014). Non-targeted repression of a well-defined identity group helps the regime by signalling the outgroups isolation to prevent collective action with anti-regime ingroups. Rorbaek and Knudsen find that excluded ethnic minorities do increase state political terror. Improving the repression of isolated groups is unlikely to impact regime survivals. Hale's big four revolutionary cascades all had strong support in the dominant ethnicity (Hale, 2013).

In most applications, AR must compete with existing methods for controlling political preferences. Military recruiting by algorithm will compete with existing strategies of recruiting favored minorities, localities and lineages (Makara, 2013; McLauchlin, 2010). In the public sector, it will compete with nepotism and patronage networks for securing loyalty. Totally new acts such as selective internet shutdowns will have no competition and should appear earlier.

How strong can AR get?

Either deep regime preferences and attitudes are difficult to impute, or this dataset is missing pieces which at otherwise necessary.

Subconscious rational control over belief formation is one explanation. Citizens in autocracies are under no obligation to form an opinion about democracy or regime change, and face costs for doing so. As Timur Kuran points out, the peoples of Eastern Europe went decades without piecing together the economic, political and governance failures of communist regimes into a preference for liberal democracy. Most citizens can simple delay considering the question until the critical moment of revolutionary action. Imputing the top-of-the-mind reaction to democracy would therefore depend less on life experiences or personalities. The actual response to revolutionary organizing would be more predictable then, but training data would be unavailable until too late.

Some academics argue that individual preferences play a small role in revolutionary cascades (Gallopin, 2019). As dissent breaks out both citizens and state actors have strong incentives to place themselves on the right side of history. Early protest joiners gain lasting status as brave national heroes, while late joiners signal cowardice and dishonesty to potential allies. Repressive actors may be killed for joining an unsuccessful mutiny, but also for refusing to mutiny.These may be higher stakes than abstract beliefs about the relative merits of leader-selecting institutions. The importance of strategic interactions implies the imputability of political behavior is limited, even if beliefs are not.

Alternatively, this model simply lacks achievable data for high accuracy. Data

on phone and internet use is a likely source. Stachl et al. predict big five personality traits from phone usage data alone. Social media offers a granular account of information intake impossible with newspaper, radio and television. Regimes are currently investing heavily in this area (Feldsetin; Frantz; Qiang). Future work and the predictive value of social media data in autocracies would be worthwhile.

Information on factional affiliations could be predictive. In many sultanistic states there exist preferred tribes or regions. Loyal tribes exchange regime support for privileges through a personalistic patronage system. Such factional data should also be investigated. Data on public service use is unlikely to prove useful given our results, but also worth investigating.

Our results are ambiguous toward location and facial recognition tracking systems. The vast majority of this data should be redundant, as students visit other students and private sector workers commute. Occupational and location data had little relevance. However, movement could correlate with face to face information sources, such as markets, community gatherings and even political meetings otherwise unsurveiled.

Conclusion

Coalition leaders and joiners remain locked in an arms race to guess beliefs and hide beliefs. Earths most brutal coalition leaders, repressive regimes, are slowly building new tools to impute beliefs of the average citizen. AR thus represents a serious risk for humanities long term future.

Fortunately, most states, perhaps all, are far from winning the arms race. The

data currently available to most autocrats provides weak gains from discriminatory repression. Old data from taxes, registries, services and political rituals (rallies and petitions) is of limited use. This suggests that collecting lots of information with no political or information source content, as the social credit system promises, will have diminishing marginal returns due to repetition.

For the near future, AR will be restricted to applications where the legitimacy costs of false positives are low. This includes civil liberties restrictions, discrimination and political terror against isolated minorities. For AR to solve the false positive problems, states must capture new information sources. Regimes are currently exploring nationwide camera surveillance with facial recognition and, most importantly, phone and internet use data.

The highest leverage intervention to stop AR is to facilitate more digital privacy in authoritarian states. Successful AR likely depends on the legibility of phone and internet use to the state. A full review of privacy in the internet of things is beyond the scope of this paper, but we highlight two low cost but high impact interventions.

The more diverse and secure a country's cell phone stock is, the harder it is to monitor. Large, high-capacity states like China have the resources to monitor many phone brands and operating systems. For poorer autocracies the costs of making every new Nokia, Huawei and Apple phone intelligible to the security apparatus will be prohibitive. International consulting firms distributing those costs is therefore worrying. The FBI has advocated for installing backdoors in phones (Leswing, 2020). Whatever minor benefits these provide law enforcement in democracies, they impose larger costs on future generations by facilitating the digital surveillance of autocrats

(Just Security, 2018).

The banning of popular sites such as YouTube in China has had an unforeseen upside. Users can evade the great firewall using virtual private networks (VPNs) which redirect traffic through external servers (French, 2008). It forces users seeking non-political entertainment into the VPN market, hiding the signatures of users consuming political news (Yang and Liu). Even if the state can detect an individual's use of VPNs, the algorithms will struggle to differentiate entertainment-consuming from information0consuming users. If AR spreads outside China liberal actors could cheaply be subsidise VPN access to citizens in at-risk autocracies. Grants for independent radio stations has long been a cost-effective strategy, and is still used in South Sudan today (Internews, 2021). VPN subsidies could become the Radio Free Europe of the 21st century.

¡what do people write in concluding paragraphs. dead ass idk. Maybe a call for more research...¿

Biographical statement