# An example Journal of Peace Research paper typeset in LaTeX

Name of Author
University of Author
Word count: 9000

February 21, 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.

# 1 Introduction Version 3

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 ” (China Cables). On the basis of these tags and subsequent interrogation, tens of thousands were sent to reeducation camps in just that month.

Algorithmic 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 even the tracking of individuals locations. Scholars and journalists are now raising the alarm about a new pattern of predictive repression using ”dataveilance” to preemptively repress actors. This new pattern diverges from the more common human analysis of speech and action for repressive targeting (Qiang, 2019). Students of repression speculate that preventative repression is now spreading from Xinjiang across China and to other aurocracies (Feldstein; Frantz et al.; Qiang; Wired). We refer to the automatic detection of potential future dissidents as Predictive Algorithmic Repression (PAR).

The discussion of AR 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 can survey. 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 change the predictive viability of big data. Nonetheless, 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 can influence dataveilance of certain states, through producing countersurveillance software (Bock et al. 2019) and through state capacity development programs. A better understanding of variable importance allows us to target those efforts.

We directly measure the predictive value of a variety of citizen level variable 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 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.

There are serious limitations to our methodology. Most importantly, the dataveillance capacity of repressive states varies widely (Frantz et al. 2020). China has hired hundreds of thousands of security specialists to read internet communication and handcode data from checkpoints (Qiang et al, 2019; Greitens, 2020). In Egypt the state struggles to collect tax and employment data on a majority of citizens

(Zaher, 2020). Our data best approximates an intermediate state between the two. Secondly, survey data is not a perfect proxy for state dataveillance. Given the lack of alternative methods, we believe our technique gives the best possible measure of variable importance.

The main finding is that variables which describe either political behavior or information sources are far more valuable. Geographic data could not be assessed due to weaknesses of the survey data. Tax and service use information had very low predictive value except for income measures. These results suggest that successful AR will depend on internet surveillance, not the direct collection of state data. The highest leverage area for preventing AR is improving access to digital privacy in repressive states, while changes to state capacity have much less impact.

The paper proceeds as follows. The next section explains why leaders impute the preferences of their followers. The following section briefly describes the history of PAR in Xinjiang, the only known use case. Next, we describe our data sources and analysis method. We then summarize the results, focusing on general trends in all 6 country cases. We conclude with comment on the CCP's motivation for using AR in Xinjiang and with policy priorities for preventing AR.

## 1.1 Algorithmic Repression in Xinjiang

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 et al. 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. In response to this mass resistance the CCP rapidly increased local coercive capacity using its customary techniques, such as increased security spending and stationing a trained policemen of local origin in each village. Mass imprisonment and intensive personal data collection were not part of the response. Tibet experienced the same timeline of resistance and response.

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. Greitens et al use CCP public officials discourse to argue "Around 2015-16, just as the CCP observed new evidence of Uyghur participation in Islamic militant groups abroad, it also concluded that as much as a third of Xinjiang's population was vulnerable to extremist influence (...) Officials concluded that existing policies focused on degrading citizens' capacity for terrorism were inadequate" (pp. 44).

In 2016 repressive tactics in Xinjiang diverged from Tibet and the rest of China. Over one million Uighur were involuntarily detained in re-education camps where they are subject to political indoctrination. The state tolerates a high false positive rate in identifying targets for detention, on the basis of eradicating Islamist ideology in China. The primary role of the AR tactics described below is selecting targets for detention and reeducation. Therefore the primary purpose is to identify a particular anti-state ideology rather than general grievances. ¡note speculation that these tactics will eventually spread across China¿

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 is often collected during intrusive home visits which families cannot refuse. Much of this data is collected by hand in a laborious process of checkpoints and home visits by tens of thousands of contract security workers (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 could use much more sophisticated algorithms to analyze this personal data. Leaked documents show that the IJOP releases tags for individuals who are then investigated and a large portion are detained for reeducation (Wilson-Chapman, 2019).

The majority of the monitored behaviors lack any obvious connection to political ideology, dissent or terrorist activities. One explanation is that the security forces are detaining persons almost indiscriminately, while falsely claiming to have a sophisticated detection system (Wang, 2019). That explanation fits Greiten et al.'s narrative that the party is happy to detain many innocents to eradicate Islamist ideology. An alternative explanation is that the CCP can use modern AI techniques to identify ideologies from apparently unrelated data.

Predictive ideology imputation is just one new tool within digital repression (Feldstein, 2019). Digital repression is "the use of new technologies, primarily the Internet, social media and Artificial Intelligence (AI) - to (...) maintain political control. 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. The costs of repression include the risk of backlash as unpopular punishments and surveillance enlarge the political opposition. Traditional repression also requires the regime pay, train and motivate thousands of party members or spies to surveill citizens. Furthermore, repression prevents citizen 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.

The digital communications world has opened a variety of tactics for repressive actors to dominate citizens by controlling information, punishing dissent and degrading collective action (Frants et al.). Regimes commonly monitor digital communications, often by hand, to identify opposition leaders, activists and dissenters to target. Gohdes has already found that in the Syrian Civil War 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 (the threat failed to dissuade protesters). For a longer treatment of new tactics see Felstein (2019) or Frantz et al. (2020).

Franz et al. 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. 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. Algorithmic repression is just one mechanism of digital repression and is relatively new and untried.

The algorithmic imputation of ideology diverges from existing strategies. First, the intended targets are all members of society, not the first movers of a revolution who tend to be highly ideological and risk-tolerant (Kuran, 1991). Secondly, the regime does not wait to observe revolutionary acts or statements then punish them. Instead it proactively alters beliefs and capacity by marginalizing persons with a particular ideology, in this case through mass arbitrary detention.

Belief Imputation and Repression

Algorithmic repression automates an ancient practice of political leaders: imputing the beliefs and preferences of followers. 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 falsehoods. We use belief and preference interchangeably because believing a regime is beneficial and preferring that regime are not distinguished in 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, but 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 both adopting utility-maximizing (safe) beliefs and hiding their true beliefs. Socially adaptive belief theory (SAB) refers to the strategic adoption of valuable 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). Citizens do also employ direct deceit. Robinson and Tannenberg find that in a list experiment (where citizens have greater anonymity) regime support in China drops by 25 percentage points relative to a direct question.

Once grievances are imputed, autocrats can choose from a limited set of responses. Leaders often respond with repression to challenges to the status quo (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 disincentives such behaviors as association and phone use before personal views. Preventative repression should be particularly difficult to legitimate. The coerced confessions common in Stalin and Saddam Husseins regimes were plausibly intended to legitimate preventative attacks on suspected future challengers or dissidents. Furthermore, the regime pays a cost when revealing grievances because dissidents become aware for their numbers and allies and have less to lose from stating their preferences (Kuran, 1995). The open question of whether and to what extent private grievance information benefits a repressive regime is beyond the scope of this paper.

An authoritarian regime can secure itself by placing loyalists positions of power, while marginalizing the disloyal. They also prefer to increase collective action capacity of loyalists and decrease that of dissenters. However, this goal is much easier stated than achieved. Evolutionary psychologists argue that humans possess innate evolved skills at coalition politics because it was critical to accessing safety and resources in early human tribes (De Waal, 2007). More recently, 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). 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.

Imputing beliefs throughout a 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 Soviet Commissars are am example specialist institution for policing loyalty. 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.

Loyalty imputation has already changed with the move to digital repression. We show that while many regimes take advantage of digital communication, much of the imputation is done by humans.

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 main objective of this study is to determine broadly which types of data are most concerning for algorithmic repression. We perform a rough simulation of algorithmic grievance imputation using publicly available data on observable behaviors and personal regime attitudes from surveys in authoritarian states. We then extract variable importance information from the most predictive models for comparisons

We simulate grievance imputation using 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

In democracies individuals have little incentive to hide their ideological positions and often intentionally signal their loyalty and belonging. In repressive states individuals have strong reasons to hide their ideology. Because ideology imputation relies on public behavior, tools for predicting ideology in democracies should be less effective in autocracy. For example, choice of newspaper should be less predictive when the state closes dissenting newspapers. As Kuran argued "In [repressive states] the very forces that discourage truthful expression also inhibit the collection and dissemination of opinion data" (1995 p. 1538). Therefore we only used data from countries with polity IV scores below 5 (Davenport and Armstrong 1996; Regan and 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 (for classifying grievances), political and economic behavior and information consumption. It has especially detailed information about internet use. The survey questions vary only slightly between countries.

We required 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 ideologies. 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 for democracy. In Algeria in 2019 millions entered the streets in a pro-democracy cascade (Volpi, 2020). Ideological positions toward democracy are also shaped by life experiences. In Egypt, Ketchley and El-Rayyes (2020) find that sustained protests for democracy reduced support for democracy by disrupting daily life.

We did not model democracies, defined by a Polity IV score above 5 (Tunisia and Lebanon). We also did not model countries experiencing civil wars (Iraq, Liby, Yemen) or foreign occupation (Palestine). This 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.

Variable Specification

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. Following conventions in 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. In all AB countries very few respondents ranked democracy as a priority issue, and relative issue ranking was not included in the grievance variables.

The regime-preference variable has greater specificity, as it excludes low-trust citizens who still prefer the current regime. However, respondents who prefer a different nondemocratic regime are excluded from the regime-type preference specification and potentially included in trust specification. Because authoritarian rules face threats from both democratic and nondemocratic challengers, we included results for the trust specification.

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 preference falsification or self-censorship. Robinson and Tanenberg 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, 2019).

### Data Analysis

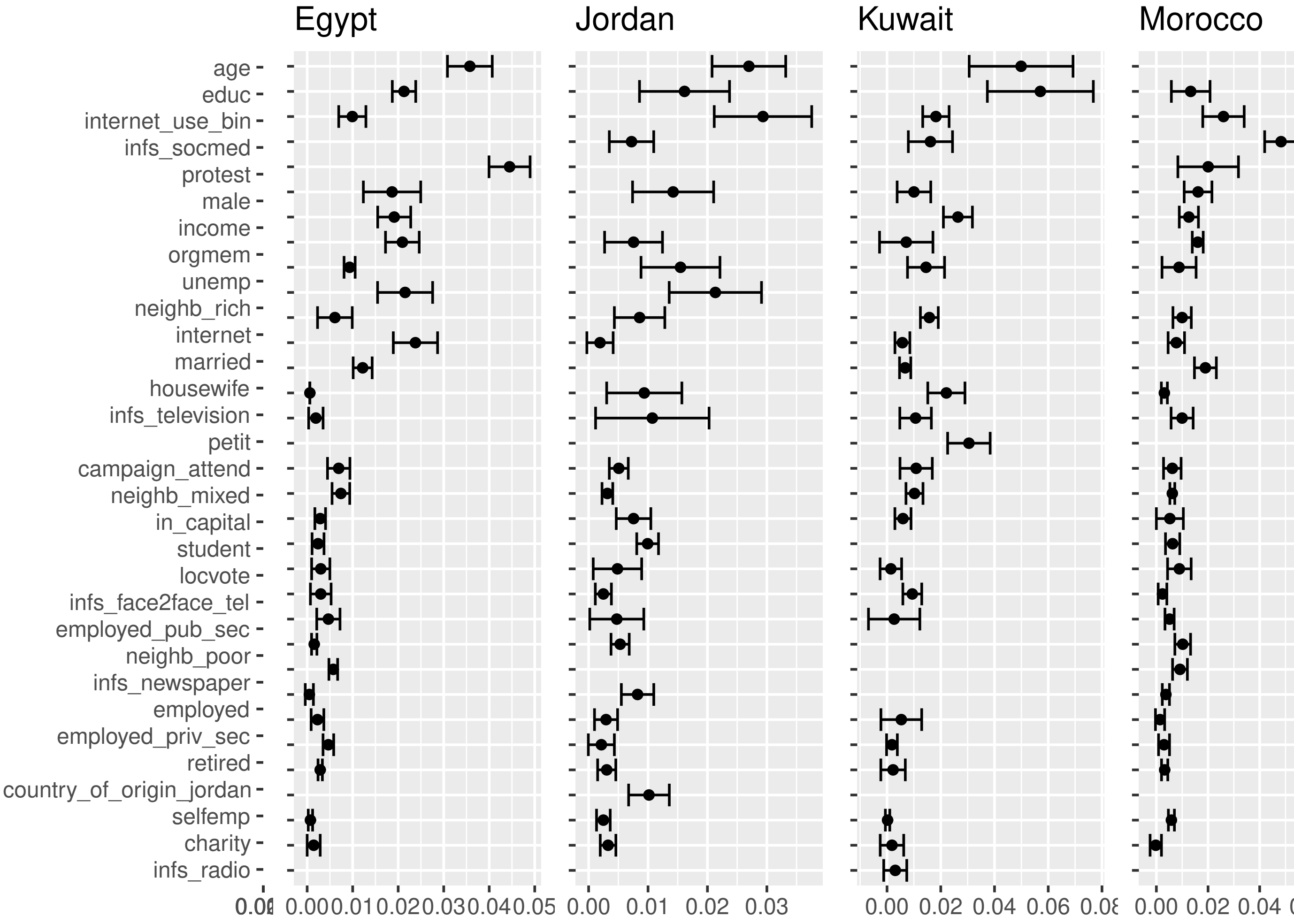
¡Youssef section here¿

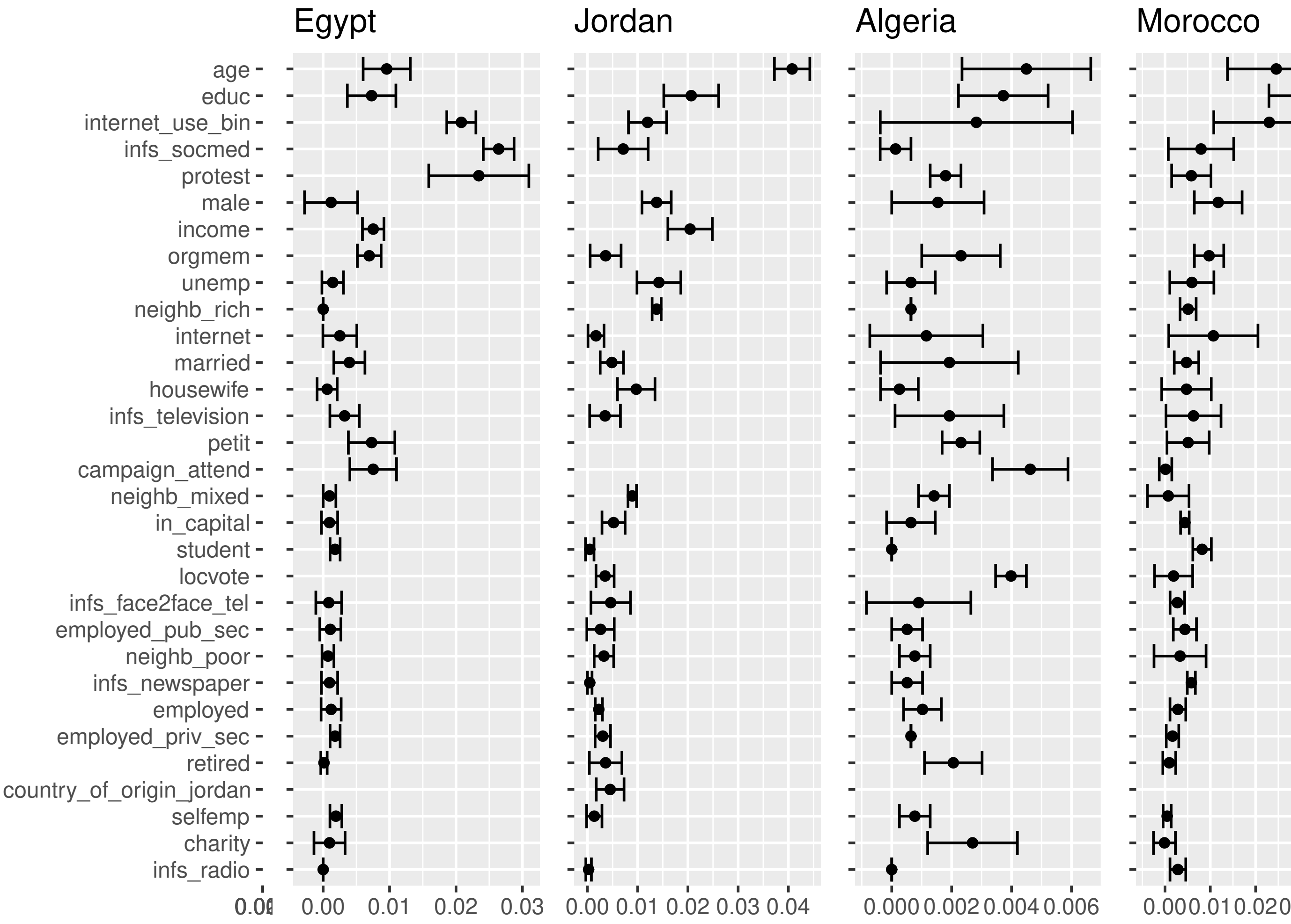
## Results

¡section about the ROC curves of the models¿

### Scenario Results

### Cross Validation


Egypt
Jordan
Kuwait
Morocco
age
educ
internet_use_bin
infs_socmed
protest
male
income
orgmem
unemp
neighb_rich
internet
married
housewife
infs_television
petit
campaign_attend
neighb_mixed
in_capital
student
locvote
infs_face2face_tel
employed_pub_sec
neighb_poor
infs_newspaper
employed
employed_priv_sec
retired
country_of_origin_jordan
selfemp
charity
infs_radio
0.000.010.020.030.040.05
0.00 0.01 0.02 0.03
0.00 0.02 0.04 0.06 0.08
0.00 0.02 0.04



¡in progress¿

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.

## Conclusion

¡in progress¿

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