

A Theory of Illiberal Democracy and Political Transitions with Heterogeneous Policy Preferences*

Chris Bidner[†]

Shirleen Manzur[‡]

January 15, 2026

Abstract

We develop a model of illiberal democracies— a formal democracy within which Elites manage to exert outsized influence. We use the model to better understand (i) the conditions that promote illiberal democracy, (ii) why illiberal democracy is harmful to, yet popular among, Citizens, and (iii) the nature of transitions between liberal democracy, illiberal democracy, and outright non-democracy. In the model, Citizens decide whether to resist Elite attempts at exerting influence. Elite influence has two main consequences: policy is distorted and the risk of transitioning to non-democracy is heightened. In understanding the popularity of illiberal democracy, we emphasize the role of economic and social heterogeneity among Citizens. Specifically, some Citizens benefit from the policy distortion, but all are harmed by the prospect of non-democracy. We show how the risk of democratic erosion leading to illiberal democracy increases as the Elite become weaker, and that the effect of mobility is nuanced. The model explains the relatively frequent transitions between illiberal democracy and non-democracy and shows how the existence of liberal democracy in fact relies upon these dynamics. We also provide some preliminary empirical support for our model from the World Values Survey and Polity IV data.

1 Introduction

Democratic backsliding is no longer happening in the form of sudden transitions from democracy to autocracy, but rather takes a more stepwise form of transitioning to illiberal democracies and eventually non-democracies. Some of the world’s oldest democracies, including USA and India, have experienced the dismantling of democratic institutions and faced threats of authoritarianism (Repucci and Slipowitz, 2021). Democratic backsliding has changed from blatant forms of coups d’etat, executive coups and election fraud to subtle forms like promissory coups, executive aggrandizement and strategic manipulation of elections (Bermeo, 2016). A systematic empirical analysis of such backsliding is provided in Lührmann and Lindberg (2019) where episodes of gradual autocratization are identified using the V-Dem dataset (Coppedge et al., 2018), documenting the “third wave of autocratization” starting in early 1990s. The concept of such form of democratic backsliding has been widely studied in the political science literature. Large scale social and economic inequality, adoption of majoritarian institutions in an ethnically divided society, asymmetry in societal resources, and many such other issues have been attributed as causes for the emergence of hybrid regimes in Merkel et al. (2006). Formal theories of illiberal democracies suggest elites driving regimes towards illiberalism,

*We are grateful for very useful comments and suggestions from participants of the Canadian Economic Association Conference 2021, Midwest Political Science Association Conference 2022 and seminar participants at Simon Fraser University.

[†]Department of Economics, Simon Fraser University. email: cbidner@sfu.ca

[‡]Department of Economics, Simon Fraser University. email: shirleen.manzur@sfu.ca

or lack of economic security leading citizens to vote for illiberal governments, even though they do not favor such regimes. These however fail to explain recent phenomenon where certain groups of citizens prefer illiberal democracies while others do not. Existing frameworks that address this issue by explaining hybrid regimes arising due to majority-minority cleavage however does not speak to the dynamics between the political regimes. To the best of our knowledge, our paper is the first to provide a formal economic theory in a dynamic setting that provide a theory of illiberal democracies and explore the conditions under which they arise and how transitions occur across different regimes.

We construct a general framework with a concrete meaning of illiberal democracy - a regime where elites exert over-sized influence on politics leading to policy distortion. More importantly, this influence creates an environment more conducive to transitioning to an authoritarian regime. Authoritarian regimes are where elites are in complete power, and thus entail the worst possible outcome for citizens. We model citizens as heterogeneous groups with different policy preferences. Certain groups' policy preference align with the elites along some dimension which results in them benefiting from policy distortion. This can arise, for example, due to the citizen group sharing the same social identity as the elite. Thus, they face a trade-off - resisting against elite influence provides protection from the prospect of non-democracy, but leads to lower flow payoffs. This is what determines citizens' optimal actions and provides insight into why, despite the threat of an authoritarian regime with negative consequences, citizens support illiberal democracies. We then allow elites to make investments which determine the likelihood with which they overtake political institutions and non-democracies arise. We find that elites invest less in increasing their chances to take over politics when some citizen groups prefer illiberal democracies and choose to not resist elite influence. These dynamics allow us to unearth interesting trends in the dynamics of transitions across political regimes, particularly, the systematic transitions between illiberal democracies and non-democracies. We now discuss our contributions in the relevant areas of the literature.

Hybrid regime - Illiberal Democracy

The political economy literature has mainly focused on political regimes classified in a binary manner. Our paper models political regimes beyond the binary measure, by allowing for the existence of hybrid regimes. We refer to these regimes as "*illiberal democracy*", where elites' influence distorts policy and creates the scope for elites to take over political institutions resulting in a non-democracy. However, elites are only able to exert this influence through receiving partial citizen support. Recent literature has addressed the existence of such "hybrid" regimes which possess both democratic and authoritarian characteristics. Such regimes have been characterized as having competitive elections but limited inclusiveness and constraints on the executive (Bidner et al., 2014), or by size of population that can collectively select incumbent and level of civil liberties (Persico, 2021). Imperfect democracies have also been modeled as captured democracies where political institutions are democratic, but elites control economic institutions and repress citizens in labor markets (Acemoglu and Robinson, 2008) or where elites' lobbying or clientilism leads to policies favoring elites within formal democracies (Grossman and Helpman, 1994). These models focus mainly on policy distortion and/or suggest that the rise of hybrid regimes is entirely elite driven which poses the question of why citizens do not take any action against the elites.

Citizen driven rise of Illiberal Democracy

Our framework incorporates the idea of citizen driven rise of illiberal democracies by modeling citizens as heterogeneous groups with different policy preferences. Some citizens' policy

preferences align with the elites and thus, they benefit from the policy distortion that occurs when elites enter politics. Hence, our model shows how citizen support for illiberal democracy is inherently rational, and does not arise from irrational actions or distorted beliefs. The political economy literature has some work on why citizens may prefer illiberal democracies. Most relevant to our model is Mukand and Rodrik (2020), which defines political regimes as a combination of property rights, political rights and civil rights, and suggests that electoral democracies often arise due to majority-minority cleavages. Acemoglu et al. (2013) theorizes that citizens vote to dismantle checks and balances on government when institutions are weak and elites are more likely to be able to bribe the government. Our paper contributes to this literature by modeling illiberal support for captured democracies, but in a dynamic setting that allows insight into the transition dynamics that ensue.

While our model allows citizen driven illiberalism to arise in a unidimensional policy space concerning only, say, economic policy, it is also applicable when considering multidimensional policy spaces that include issues such as social policy. This is particularly relevant because in recent times, illiberal democracies have been seen to often arise through citizen support, especially in the presence of identity cleavages. One of the most recent demonstrations is the Capitol Riot in the USA. Resurgence of racial animosity was used in political dialogue to turn majority against ethnic or racial minority (Huq and Ginsburg, 2018) in the US. Starkly put in Kaufman and Haggard (2019), the demonization of racial and ethnic minorities and existing resentments against immigrants was used to reduce support for centrist political parties and allow majoritarian or autocratic electoral campaigns. Bartels (2020) further documents empirically that the best predictor of anti-democratic sentiments among American Republicans is ethnic antagonism, especially concerns about political power and government resources aimed towards immigrants, African-Americans and Latinos. Another prevalent example where ethnic identity created a divide across citizens is Hungary, one of the first countries in Europe moving towards authoritarian rule as documented in Bogaards (2018). From 2011, the incumbent Fidesz government brought changes to the electoral system, constitution and the justice system in ways that gave them a competitive advantage and facilitated an authoritarian regime. Hungary exhibited illiberalism through flawed voting rights of non-resident ethnic Hungarian, the government’s handling of the refugee crisis, a combination of nativism and Christianity, and attacks towards academic freedom. In a more general context, Foa and Mounk (2016) uses the World Values Survey to document the threat of democratic de-consolidation that may arise in Europe and North America as a consequence of younger cohorts in consolidated democracies finding liberal institutions as less essential for democracy, being less politically engaged and having increased support for authoritarian political systems. Latin American countries have also been sliding back on the democratic scale, despite the third democratization wave of late 1990s. Populist leaders like Evo Morales in Bolivia and Hugo Chavez in Venezuela came into power under the rhetoric that they want to free the state from being controlled by elites, and used majority support through referendums to erode checks and balances (Acemoglu et al., 2013). Both leaders’ social reforms involved efforts to eliminate discrimination, which was beneficial for the indigenous groups and led them to support these populist leaders, until the leaders’ authoritarian measures started costing them in terms of economic or environmental policies.

Other forms of identity divide that facilitated the rise of illiberal democracy is religious divide, e.g., the Hindu-Muslim divide in India, Secularism-Islam divide in Turkey and Buddhist-Muslim conflict in Myanmar. For instance, using exogenous Ramadan timing, Colussi et al. (2021) shows empirically that in areas with mosques in Germany, increased salience in religious identity and cultural dissimilarity due to Ramadan tends to increase votes for extreme left and right parties due to worsened attitudes towards Muslims. This empirically highlights

how salience in identity can lead to a deterioration in the liberal aspect of a democracy, while retaining the majoritarian aspect. Our model aims to explain the rise of illiberal democracies as well as persistence of liberal democracies in the presence of identity cleavages, whether or not income class differences between citizens is not dominant.

Overall, our model is relevant in the flourishing literature in identity and is able to explain cases of illiberal democracies emerging with citizen support all over the world. Shayo (2009) introduced identity in the political economy literature by modeling the benefit voters receive from status of belonging to an identity. Bonomi et al. (2020) shows how this can lead to distorted political beliefs, polarization and explain changing political cleavages. While we do not explicitly model manipulation of beliefs about polarization, parameters within the model can be defined as functions of media bias, political dialogues regarding identity differences or negative shocks (as in Gratton and Lee (2020)) which could change how citizens weigh different policy dimensions and can potentially explain how illiberal democracies arise with citizen support. The generality of our framework thus opens up the scope of studying how increasing salience of identity can directly influence the political regime through its effect on collective action and divide among the citizens within the society.

Transitions across regimes

Our model focuses on the heterogeneity among voters, and the interplay with elites' investment in taking over political institutions. This allows us to analyze the trends in transitions between political regimes. We find that transitions out of non-democracy is more likely when elites are weak, and citizens are impatient. Furthermore, both transitions out of non-democracy and out of democracy are more likely when citizens are divided in their resistance. Such transition trends mirror the cyclical movement between illiberal democracies and non-democracies that we observe in data, further discussed in Section 2.

Our paper also speaks to democratic consolidation, which refers to securing new democracies against reverse waves and authoritarianism (Schedler, 1998). The term arose after the democratization wave around 1990s, when pressing concerns about stabilizing newly established democracies emerged. Notable theory of democratic consolidation in the economics literature is Persson and Tabellini (2009), which formalizes the consolidation of democracy as democratic capital arising from historical experience with political regimes in neighboring regions. Fitting into the general idea of slippery slopes leading to institutional persistence (Acemoglu et al., 2020), our setup speaks to the conditions under which democracies consolidate and liberal democracies persist. Our theory resonates with the idea of how the threat of a worse institutional arrangement for groups in control (namely non-democracy) can lead to persistence of a "good" equilibrium (namely liberal democracy). Stochastic shocks within a similar setup was also discussed in Acemoglu et al. (2015) where the direction of a society's institutional path changes only when such shocks occur. Our model differs from these models in the aspect that the selection of the regime next period does not occur by chance alone (stochastic shock) and is not chosen by any particular group in power. There is a nuanced difference in how all the groups' actions within society simultaneously interact and how these dynamics determine the regime next period.

While transitions between democratic and autocratic regimes have been extensively studied in the economics literature (Acemoglu and Robinson, 2000, 2001; Buchheim and Ulbricht, 2020), models of regime transition incorporating hybrid regimes is sparse. A relevant study is Gratton and Lee (2020) which models the rise of illiberal democracies from voters facing a trade-off between economic security and liberty. The theory focuses on possible shocks that voters cannot be protected from unless the government operates beyond the constraints of

a formal democracy. We take a different stance where some citizens whose preferences align with elites choose illiberalism, not to be protected from negative shocks but to exploit the benefit from policy distortion.

We also introduce the possibility of mobility within citizen groups and analyze how mobility affects the likelihood of illiberal democracies. We find that the effect of mobility on the likelihood of the emergence of illiberal democracy is nuanced and is conditional on the relative magnitude of the benefit of democracy as opposed to non-democracy for the citizen groups.

The paper is organized as follows - firstly, Section 2 presents some stylized facts as motivation. Section 3 lays out the general model and analyzes the equilibrium; Section 4 then extends the analysis by introducing mobility between the heterogeneous groups of citizens. Finally, concluding remarks are provided in Section 5.

2 Stylized facts

In this section, we present some stylized facts that motivate our model. First of all, we show the relation between attitudes of individuals towards those of differing identities and their views about aspects of liberalism. We then provide an illustration of trends in Polity score which provides an insight into the transition dynamics of non-democracies, illiberal democracies and liberal democracies.

Citizen driven rise of Illiberal Democracy

Table 1: Relation between attitudes and illiberalism from World Value Survey

Neighbor:	Civil Rights	Strong rule	Army leader	Support for democracy
Different race	-0.041*** (0.011)	0.185*** (0.036)	0.186*** (0.023)	-0.146*** (0.023)
Observations	49303	47891	48414	49175
Immigrants/foreign workers	-0.033*** (0.008)	0.102*** (0.035)	0.126*** (0.017)	-0.108*** (0.023)
Observations	49220	47808	48331	49091
Different religion	-0.037** (0.013)	0.178*** (0.046)	0.174*** (0.038)	-0.126*** (0.025)
Observations	49291	47880	48401	49164
Country Fixed Effects	Yes	Yes	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes

Sample is restricted to 17 countries with a Polity score of at least 9 in 2006 in the WVS study¹. Controls for Polity score is also included. Standard errors are clustered at the country level.

We carry out an empirical analysis to study the relation between salience in identity and attitudes towards the liberal aspects of democracy. Using data from Waves 6 and 7 of the World Value Survey for years 2010-2014 and 2017-2020, Table 1 presents the relation between individuals' attitudes towards neighbors of a different race, different nationality or of a different religion with measures for illiberalism. Individuals disliking having neighbors of different identity is indicative of salience in identity. We use the same measures of illiberalism as in

Gratton and Lee (2020) - beliefs about whether civil rights is an important characteristic of democracy, whether political systems with a strong leader is good, whether political systems where the army rules is good and how good it is to have democracy as a political system. The sample is restricted to 17 countries with a Polity score of at least 9 in 2006 in the WVS study². The results clearly reveal that individuals who mention that they would not like having neighbors of a different race, nationality or religion are significantly more likely to believe that civil rights is not a very important characteristic of democracy, political systems with a strong leader or where the army rules is good, and that democracy is not very good as a political system.

Transitions across regimes

The systematic transitions between non-democracy and illiberal democracy is illustrated in Figure 1. Using Polity score ranging from 1950-2020 of the Polity IV dataset, we plot the fitted values of a fifth order polynomial regression of the 25th, 50th and 75th percentile of ten-year changes in Polity score, at each value of the index, normalized between 0 to 1. In other words, for each value of the Polity score, there is an associated distribution of change in score over ten-year periods and the percentiles describe this distribution. If we only focused on the median value of the change (the 50th percentile), it would seem like there is not much change in the Polity score, and regimes are stable. However, the instability of regimes, especially in nations with Polity scores in the mid-range becomes obvious when considering the huge gap between the 25th and 75th percentile. We can observe that for lower levels of the Polity index, a change in score is more likely to be an increase in the score, whereas countries with scores between 0.6 to 0.9 are more likely to experience a decrease in the Polity score. In summary, data suggests systematic transitions between non-democracy and illiberal democracy while liberal democracies tend to remain more stable.

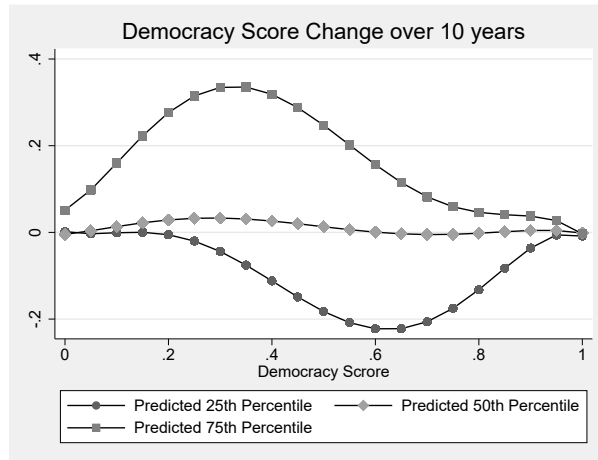


Figure 1: Trends in Changes across Polity V score

3 Model

In this section, we present the basic model that allows us to analyze the equilibrium and provides insight into what leads to illiberal democracies and the transitions dynamics between

²The largest sample of democracies surveyed was in 2006 and thus, the sample countries were selected from 2006. We are fixing the sample of countries using a base year as we do not want the result to be driven by changes in countries, but rather want to identify the relation between regime and attitudes within country over the different survey waves.

regimes.

3.1 Setup

Consider an infinite horizon society in discrete time. The society is populated by two classes of agents: elites and citizens. Elites ($i = 0$) comprise a small fraction of the population while citizens make up the larger portion of the population. Citizens are divided into two groups, $i = \{1, 2\}$ with heterogeneous policy preferences. All agents discount the future with a discount factor of $\beta \in [0, 1)$. Each period starts in one of two possible states, $\omega_t \in \{\mathbf{D}, \mathbf{N}\}$. Political regimes are denoted $\phi \in \{N, D, \tilde{D}\}$. Within the state $\omega_t = \mathbf{N}$, the political regime is non-democracy ($\phi = N$). In the state, $\omega_t = \mathbf{D}$, the political regime can be illiberal democracy ($\phi = \tilde{D}$) or liberal democracy ($\phi = D$). Regimes are determined based on the policies implemented, as discussed below.

Policy and Payoffs

A policy is some $\theta \in \Theta \subset \mathbb{R}^K$ determined from the maximization of a weighted social welfare function

$$W_\eta(\theta_{\eta_\phi}) \equiv \eta_{0\phi} \cdot u_0(\theta_{\eta_\phi}) + \eta_{1\phi} \cdot u_1(\theta_{\eta_\phi}) + \eta_{2\phi} \cdot u_2(\theta_{\eta_\phi}). \quad (1)$$

where $u_i(\theta_{\eta_\phi})$ is the flow payoff of agent i conditional on the regime specific policy, θ_{η_ϕ} . The weights on each group $i = \{0, 1, 2\}$ is denoted as $\eta_\phi = (\eta_{0\phi}, \eta_{1\phi}, \eta_{2\phi})$ and depend on the political regimes.

Assumption 1. *Weights on agents' policy payoffs in each political regime is as follows:*

$$\eta_N = (1, 0, 0) \quad (2)$$

$$\eta_D = (0, \rho, 1 - \rho) \quad (3)$$

$$\eta_{\tilde{D}} = (\varepsilon, (1 - \varepsilon) \cdot \rho, (1 - \varepsilon) \cdot (1 - \rho)). \quad (4)$$

Assumption 1 characterizes the political regimes in our model. When the political regime is non-democracy (N), policy maximizes the elites' payoffs; in a liberal democracy (D), a weighted welfare function of citizen groups is maximized; and in an illiberal democracy ($\tilde{D}$), elites exert influence of $\varepsilon \in (0, 1)$ on policy determination, but they do not entirely seize political power.

The policy preferences for group i is defined as:

$$u_i(\theta_{\eta_\phi}) = - \sum_k (\theta_{ik}^* - \theta_{\eta_\phi k})^2 \quad (5)$$

where θ_{ik}^* is the most preferred outcome in dimension k of the agents' most preferred policy vector θ_i^* . Using these payoffs for the social welfare maximization problem yields the following policy:

$$\theta_{\eta_\phi}^* = \eta_{0\phi} \cdot \theta_0^* + \eta_{1\phi} \cdot \theta_1^* + \eta_{2\phi} \cdot \theta_2^*. \quad (6)$$

We now make some further assumptions on the policy bliss points for each agent to complete the description of the environment.

Assumption 2. Let $\theta_{0k}^* \leq \theta_{2k}^*$ for each k . The preferred policy of the three classes of agents are ordered as follows:

$$\theta_{1k}^* \in \left[\frac{\theta_{2k}^* - \theta_{0k}^*}{2}, \theta_{2k}^* \right] \quad (7)$$

for each k (strict for some k).

With Assumption 2, we first assume that the preferred policy alignment is closer between citizen groups relative to the elites. We further assume that elites have policy preferences that are more aligned with citizen group 1 rather than citizen group 2. Specifically, citizen group 1 has policy preferences that are more aligned with the citizen group 2 than with the elites ($i = 0$) such that

$$\theta_{2k}^* - \theta_{1k}^* \leq \theta_{1k}^* - \theta_{0k}^*$$

and that elites are aligned with citizen 1 more than citizen 2 in the sense that

$$\theta_{1k}^* - \theta_{0k}^* \leq \theta_{2k}^* - \theta_{0k}^*$$

for each k (strict for some k).

Next, we make some assumptions on the extent of influence elites exert in an illiberal democracy (ϵ).

Assumption 3. Elite influence is sufficiently small, but relevant in determining payoffs for citizen 1, that is,

- ϵ is sufficiently small. Specifically, $\epsilon \in [0, \bar{\epsilon}_k]$ where³

$$\bar{\epsilon}_k \equiv \frac{(1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*)}{(1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*) + (\theta_{1k}^* - \theta_{0k}^*)}.$$

- $\theta_{1k}^* - \theta_{0k}^* \neq 0$ for some k .

Assumption 2 and 3 then results in the following ordering of payoffs for each group from each regime⁴:

$$\begin{aligned} \text{Elites: } & u_{0D} \leq u_{0\bar{D}} \leq u_{0N} \\ \text{Citizen 1: } & u_{1N} \leq u_{1D} \leq u_{1\bar{D}} \\ \text{Citizen 2: } & u_{2N} \leq u_{2\bar{D}} \leq u_{2D} \end{aligned} \quad (8)$$

This payoff ordering is a crucial element of our model. Intuitively, elites receive the highest payoffs from non-democracy as they completely control political institutions, second highest payoffs from illiberal democracy as they can somewhat influence policy making and finally, the lowest payoffs from liberal democracy. Citizens always receive the lowest payoffs from a non-democracy as elites capture all political institutions and thus, non-democracy is their least preferred regime. Since citizen group 1 shares interests with the elites, they receive a higher payoff when elites distort policy in elites' favor. Let us now discuss two relevant contexts where these payoff ordering hold.

³Details provided in Appendix A.1

⁴Details provided in Appendix A.1

Redistribution

This payoff ordering is relevant in the widespread discussion on redistribution policies and tax rate. Consider a uni-dimensional policy space with a focus on some economic policy, such as redistribution or the highest marginal tax rate. Assume that elites are the richest, citizen 1 is the middle income group such that their preference for economic policy is more aligned with the elites, relative to citizen 2 who comprises the low income group. This is illustrated in Figure 2.

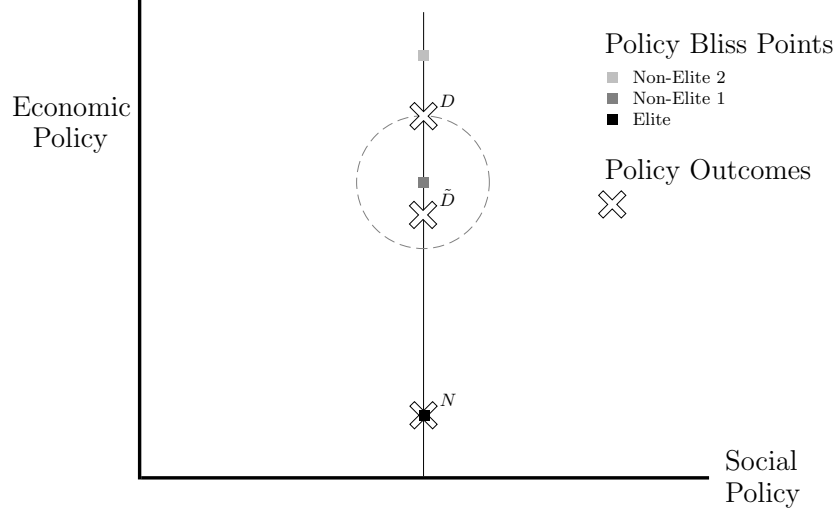


Figure 2: Policy Preference

In a non-democracy, policy would be closest to the elites' bliss point, that is, low levels of redistribution (denoted N). In the state of democracy where elites do not influence politics, the policy outcome would be that of a liberal democracy somewhere in between the citizens' bliss points, where their indifference curves are tangential conditional on the policy weight (denoted by D in Figure 2). However, an illiberal democracy where elites distort tax rate in their favors would benefit citizen 1 as it puts them on an indifference curve closer to their bliss point (policy denoted $\tilde{D}$). Non-democracy would still be the worst regime for both citizen groups as elites completely seize power and undermine both citizen groups' preferences and result in lowest levels of redistribution. Citizen 2 would receive the highest payoff in a liberal democracy where redistribution is highest. Hence, the payoff ordering is the same as assumed in Assumption 2.

The policy implications from this are in line with the empirical findings in Garcia and Von Haldenwang (2016) where the authors identify a U-shapes relation between tax to GDP ratio and the Polity score. Through an extensive literature review of the theoretical considerations, the paper summarizes the reasons behind either high or low taxation in both democracy and autocracy. The reasoning for high taxation in autocracies is because it is a form of appropriation by the rulers. This would imply lower redistribution to citizens. On the other hand, higher taxes lead to higher redistribution in liberal democracies. Using the Polity score, they find that tax collection is lowest in hybrid regimes, which would also lead to lower levels of redistribution relative to a liberal democracy, as illustrated in Figure 2. Note that our model differs here from Mukand and Rodrik (2020) as it allows illiberal democracy to exist even in the absence of identity cleavages and only based on heterogeneity by income class. Next, we consider the scenario where income class difference and identity cleavages co-exist.

Social Identity

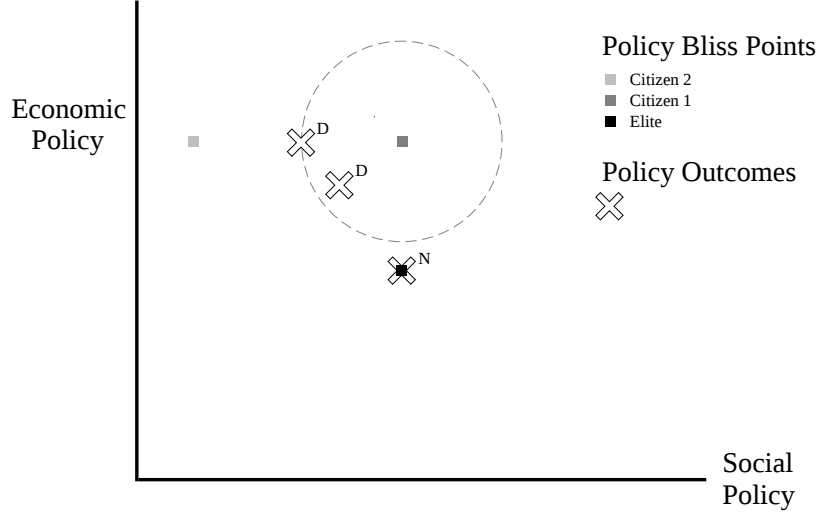


Figure 3: Policy Preference

Illiberalism within democracies in an environment where there is a social divide among citizens is a more commonly observed phenomenon in recent times. Let us consider a multidimensional policy space with economic and social policy (i.e., $k = 2$). For simplicity, assume that citizens have identical preferences about economic policy, but citizen 1 shares a social identity with the elites, whereas citizen 2 has a different social identity. Thus preference of citizen 1 about social policy aligns with that of the elites. The bliss points of each group is as illustrated in Figure 3.

As in the uni-dimensional case, liberal democracy gives the highest payoff to citizen 2. Illiberal democracy gives the highest payoff for citizen 1 due to the policy distortion by elites leading to a policy outcome that puts them on an indifference curve with higher utility⁵. As discussed in Section 1, identity divide in the form of racial divides, religious identity, etc. can lead to divides among citizens that leads to the policy outcomes illustrated in 3 and thus the payoff ordering in Equation 8.

Timing

For the final set up of the model, the agents actions involves citizens choosing to resist (r_{it}) which determines their power (p_{ω_t}). Citizens' power plays a role in the probabilistic determination of flow payoff as well as future state ω_{t+1} . When the citizens are not powerful, elites can enter politics and distort policy. Once in politics, elites can completely seize power with probability δ_{ω} .

The timing of events in each state, $\omega_t \in \{\mathbf{D}, \mathbf{N}\}$ is as follows:

- Citizens choose level of resistance, $r_{it}(\omega_t) \in [0, 1]$ for $i \in \{1, 2\}$.

⁵While not explicitly modeled for simplicity, the flow payoffs of agents could be structured as a weighted combination of payoff from k policies. The differences in weights given on economic vs social policy will then influence the difference between $u_{1\bar{D}}$ and u_{1D} . Changes in external factors (e.g., use of information technology, globalization, etc.) can shift these weights and explain how the increased importance given to social policy can lead to illiberalism.

- Power of citizens is determined as

$$p_{\omega_t} \equiv \alpha_{\omega_t} \cdot [\rho \cdot r_{1t} + (1 - \rho) \cdot r_{2t}]$$

where $0 \leq \alpha_{\mathbf{N}} < \alpha_{\mathbf{D}} \leq 1$. Here, α_{ω_t} can be interpreted as the state dependent effectiveness of citizens' resistance in determining their power. Intuitively, their resistance in democracy is at least as effective as in non-democracy.

The share of citizen group 1 in determining citizens' power is ρ and that of citizen group 2 is $1 - \rho$. This can be interpreted as population share, for instance, $\rho > 0.5$ means citizen group 1 is the majority and thus, their resistance is more significant in making citizens powerful. The interpretation of ρ is however, not restricted to population size, and can be interpreted in many different ways, for example, as citizen group 1's ability relative to group 2 to coordinate and organize.

- Payoffs for $i = \{0, 1, 2\}$ are realized.

When $\omega_t = \mathbf{D}$,

$$u_{it} = \begin{cases} u_{iD} & \text{w/p } p_{\mathbf{D}} \\ u_{i\tilde{D}} & \text{w/p } 1 - p_{\mathbf{D}} \end{cases}$$

Liberal democracies arise when citizens are powerful which occurs with probability $p_{\mathbf{D}}$. Otherwise, elites can distort policy leading to a flow payoff of $u_{i\tilde{D}}$.

When $\omega_t = \mathbf{N}$,

$$u_{it} = u_{iN}$$

The payoff in non-democracy does not depend on the power of the citizens because elites choose policy without taking citizens' preferences into consideration⁶.

- Next state is then determined.

In the state of democracy, $\omega_t = \mathbf{D}$,

$$\omega_{t+1} = \begin{cases} \mathbf{D} & \text{w/p } 1 - (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}} \\ \mathbf{N} & \text{w/p } (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}} \end{cases}$$

When citizens are not powerful and elites completely seize power, the state changes from democracy ($\mathbf{D}$) to non-democracy ($\mathbf{N}$). Otherwise, the state remains in $\mathbf{D}$.

If the current state is non-democracy, $\omega_t = \mathbf{N}$,

$$\omega_{t+1} = \begin{cases} \mathbf{D} & \text{w/p } p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}}) \\ \mathbf{N} & \text{w/p } 1 - p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}}) \end{cases}$$

When citizens are powerful and elites fail to completely seize power, the state changes from non-democracy ($\mathbf{N}$) to democracy ($\mathbf{D}$). Otherwise, the state remains in $\mathbf{N}$.

⁶This is a simplification in the model. In $\omega_t = \mathbf{N}$, the payoffs can still depend on $p_{\mathbf{N}}$. However, since elites choose policy, conditional on citizens' strategy, it would be always optimal to choose a policy that gives the elites maximum payoff.

3.2 Equilibrium Analysis

In this section, we analyze the Markov perfect equilibria (MPE). The MPE is such that strategies are not conditional on the history of the game beyond the effect the history has on the payoff relevant state $\omega_t \in \{\mathbf{D}, \mathbf{N}\}$. Furthermore, we restrict our analysis to symmetric MPE such that the equilibrium strategies do not depend on individual identity of citizens or elites, and are identical across agents within each class - elites, citizen group 1 and group 2.

The equilibrium analysis can be simplified through some observations of the state dependent payoff ordering. First of all, given that non-democracy is the least preferred regime for citizens, both groups of citizens always resist when state $\omega_t = \mathbf{N}$. Furthermore, since citizen group 2 has the highest payoff when elites are not in politics, they also always choose to resist when state $\omega_t = \mathbf{D}$. Hence, the analysis boils down to citizen 1's resistance in the state of democracy. The main trade-off they face is that resisting reduces the likelihood of elites influencing policy which reduces their flow payoff, but it benefits them through the reduced probability of a transition to non-democracy.

Formalizing citizen 1's problem, let V_ω be the value for citizen 1 in state ω . The above observations imply

$$V_{\mathbf{N}} = u_{N1} + \beta \cdot [(1 - \pi_{\mathbf{N}}) \cdot V_{\mathbf{D}} + \pi_{\mathbf{N}} \cdot V_{\mathbf{N}}] \quad (9)$$

where $\pi_{\mathbf{N}} \equiv 1 - \alpha_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})$ is the equilibrium probability of remaining in non-democracy. We also have

$$V_{\mathbf{D}} = p^* \cdot u_{D1} + (1 - p_{\mathbf{D}}^*) \cdot u_{\tilde{D}1} + \beta \cdot [(1 - (1 - p_{\mathbf{D}}^*) \cdot \delta_{\mathbf{D}}) \cdot V_{\mathbf{D}} + (1 - p_{\mathbf{D}}^*) \cdot \delta_{\mathbf{D}} \cdot V_{\mathbf{N}}] \quad (10)$$

where $p^* \equiv \alpha_{\mathbf{D}} \cdot [\rho \cdot r_1^* + (1 - \rho)]$ is equilibrium power in democracy. These two can be solved to get $\{V_{\mathbf{D}}, V_{\mathbf{N}}\}$ as a function of r_1^* . There is generically two possible Markov-Perfect equilibria (details in Appendix A.2.1). One is where citizen 1 always resists with $r_1^* = 1$ which we call the *Liberal Democracy equilibrium* since citizens act in sync to ensure that elites are not in power. The other equilibrium is where citizen 1 does not resist with $r_1^* = 0$ in the state of democracy which leads to a higher probability of elites entering politics and the political regime becoming an illiberal democracy ($\tilde{D}$).

Proposition 1. *A symmetric Markov Perfect Equilibrium exists for $r_1^* \in \{0, 1\}$.*

An Illiberal Democracy equilibrium exists if and only if

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \leq \Lambda(\varepsilon, \rho) \quad (11)$$

A Liberal Democracy equilibrium exists iff

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \geq \Lambda(\varepsilon, \rho) \quad (12)$$

where

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \equiv \frac{\beta \cdot \delta_{\mathbf{D}}}{\beta \cdot \delta_{\mathbf{D}} + (1 - \beta) + \beta \cdot \alpha_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})} \quad (13)$$

and

$$\Lambda \equiv \frac{u_{1\tilde{D}} - u_{1D}}{u_{1\tilde{D}} - u_{1N}} \equiv \frac{\sum_k \Theta_k(0, \rho)^2 - \Theta_k(\varepsilon, \rho)^2}{\sum_k \Theta_k(1, \rho)^2 - \Theta_k(\varepsilon, \rho)^2} \quad (14)$$

where

$$\begin{aligned}
u_{1N} &\equiv \sum_k \Theta_k(1, \rho)^2 \equiv \sum_k (\varepsilon \cdot (\theta_{1k}^* - \theta_{0k}^*))^2 \\
u_{1\tilde{D}} &\equiv \sum_k \Theta_k(\varepsilon, \rho)^2 \equiv \sum_k (\varepsilon \cdot (\theta_{1k}^* - \theta_{0k}^*) - (1 - \varepsilon) \cdot (1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*))^2 \\
u_{1D} &\equiv \sum_k \Theta_k(0, \rho)^2 \equiv \sum_k ((1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*))^2
\end{aligned}$$

Here, $\Theta_k(1, \rho)$ represents the utility of citizen 1 in non-democracy, $\Theta_k(\varepsilon, \rho)$ is utility in an illiberal democracy and $\Theta_k(0, \rho)$ is the utility in a full democracy (detailed workings in Equation 24 in Appendix A.1). In Proposition 1, Equation 14 represents the added benefit of an illiberal democracy compared to a liberal democracy to citizen 1 as a ratio of the added benefit of an illiberal democracy compared to a non-democracy. If this ratio is high enough, then the benefit of elites influencing policy outweigh the threat of non-democracy for citizen 1. The threshold that determines this decision is given by Equation 13 which depends on the discount factor (β), the probability of elites completely taking over (δ_ω) and the effectiveness of citizens' resistance (α_{ω_t}). Thus, Proposition 1 suggests that if the benefit of elite involvement in politics is high enough for citizen 1, they choose not to resist, which results in lower power p_D of the citizens making it more likely for elites to enter politics. On the other hand, if the potential future costs of being in a non-democracy outweighs the benefit from higher payoffs from elite involvement, citizen 1 chooses $r_1^* = 1$, making a liberal democracy more likely and thus, leading to a liberal democracy equilibrium.

An illiberal democracy equilibrium is thus supported by citizens (at least, certain citizen groups). This is unlike the common explanation that “*Democracy’s erosion is, for many, almost imperceptible*” (Levitsky and Ziblatt, 2018), that is, citizens do not resist illiberalism because they do not notice it. Instead, our model suggests that rational optimization of citizens make them support democratic erosion since they benefit from policy distortion by the elites.

However, the non-resistance also raises the probability of transitioning to an authoritarian regime in the illiberal democracy equilibrium as opposed to the liberal democracy equilibrium. This can be seen clearly from the fact that the illiberal democracy equilibrium leads to citizens having lower power, p_D which increases the probability of elites entering politics. Given the exogenous probability, δ , with which elites completely seize power once they enter politics, the probability of transition from state of democracy is higher in the illiberal democracy equilibrium. Mathematically, the probability of transitioning from state $\omega_t = D$ to $\omega_{t+1} = N$ in the illiberal democracy equilibrium is $(1 - (1 - \rho) \cdot \alpha_D) \cdot \delta_D$ which is greater than the probability, $(1 - \alpha_D) \cdot \delta_D$, in the liberal democracy equilibrium. The probability of transitioning from non-democracy to democracy is however equal across the two equilibria since both groups of citizens always resist in the state of non-democracy under both equilibrium.

Proposition 1 allows us to make meaningful comparative statics that provide insight into the likelihood of the rise of the illiberal democracy equilibrium (Proof in Appendix A.2.2).

Proposition 2. *The illiberal democracy equilibrium more readily arises when*

- *The elites are weak (as $\Gamma_{\delta_D} > 0$, $\Gamma_{\delta_N} > 0$ and $\Gamma_{\alpha_N} < 0$)*
- *Citizens are impatient (as $\Gamma_\beta > 0$)*
- *Citizen 1 plays a smaller role in determining citizens’ power, that is, for a lower ρ (since $\Lambda_\rho \leq 0$)*

Intuitively, when there is a low probability of elites completely seizing power once they enter politics in the state of democracy (δ_D), or when there is lower persistence within non-democracy (δ_N), the threat and potential cost of transitioning to a non-democratic regime is also lower for citizen 1. Thus, they choose not to resist when in a democracy, leading to the illiberal democracy equilibrium. Furthermore, if the effectiveness of resistance in a non-democracy (α_N) is high enough, confidence in the ability to transition back to democracy where they receive the illiberal democracy payoff is also high. This leads to the higher likelihood of the illiberal democracy equilibrium.

Another interesting implication of Proposition 1 is that patience promotes the liberal democracy equilibrium. This is because the trade-off involving higher probability of transitioning to a non-democracy is a threat of the future and the costs are internalized by citizen 1 only when they care sufficiently about the future, thus leading them to resist when in a democracy.

We also have $\Lambda_\rho \leq 0$ which suggests that when citizen 1 has a higher weight in policy and citizens' power determination (higher ρ), the relative benefit of illiberal democracy is lower which makes it less likely for the illiberal democracy equilibrium to arise. If we interpret ρ as the population share of citizen 1, a higher population share would deter an illiberal democracy equilibrium. Consider the context of social identity cleavage. When it is the minority who share a social identity with elites (smaller ρ), an illiberal democracy equilibrium is more likely to arise. This result is in contrast with Mukand and Rodrik (2020) where they use the example of South Africa as a case of liberal democracy arising because elites share the same identity as the minority. The reasoning behind this in their model was the importance of civil rights to the elites, since their identity is the same as that of the minority. Our model can also predict a liberal democracy regime being prevalent despite being in an illiberal democracy equilibrium when citizen group 1 is a minority. The mechanism is however very different - liberal democracy as a regime is more likely to arise simply because the majority (citizen 2) always resists as they do not prefer policy distortion. This makes the citizens powerful (higher p_{ω_t}), rarely creating the opportunity for elites to influence policy and thus, making it less likely for an illiberal democracy to arise. The fact that the nation is in an illiberal democracy equilibrium does not hinder the existence of liberal democracy as long as ρ is sufficiently small. This can therefore potentially explain how rise in globalization and integration of different nationalities within communities could potentially be threatening the majority populations as ρ increases leading to the emergence of illiberal democracies.

3.3 Endogenous δ

The model presented establishes the different regimes that may arise as a result of citizens' actions and the probability of transition in states under two different equilibrium. However, so far, the elites do not take any action in the game, and thus has no role in determining the equilibrium. The emergence of different regimes and transition between non-democracy and democracy are both likely to be influenced by elites. In this section, we check the robustness of the model's prediction by enriching the model and allowing elites to invest in influencing the probability with which they completely seize power, δ while incurring a cost $c(\delta_\omega)$ where $c'(\delta_\omega) > 0$ and $c''(\delta_\omega) > 0$. In democracy, the elites choose δ_D in the event that they get the opportunity to enter politics, that is, citizens are not powerful. In non-democracy, the elite choose δ_N which determines the likelihood that the elites retain their power and the state remains a non-democracy. In either case, the elite's choice of δ_ω influences the determination of the state in the following period. The problem of the elite in state ω is thus:

$$\max_{\delta_\omega \in [0,1]} \{ \beta \cdot [\delta_\omega \cdot E_N + (1 - \delta_\omega) \cdot E_D] - c(\delta_\omega) \} \quad (15)$$

where $E_{\mathbf{N}}$ and $E_{\mathbf{D}}$ are the equilibrium values for the elite in non-democracy and democracy respectively, defined as follows:

$$E_{\mathbf{D}} = p_{\mathbf{D}}^* u_{D0} + (1 - p_{\mathbf{D}}^*) u_{\bar{D}0} + \beta \cdot [(1 - (1 - p_{\mathbf{D}}^*) \cdot \delta^*) \cdot E_{\mathbf{D}} \dots \dots + (1 - p_{\mathbf{D}}^*) \cdot \delta^* \cdot E_{\mathbf{N}}] - (1 - p_{\mathbf{D}}^*) \cdot c(\delta^*) \quad (16)$$

$$E_{\mathbf{N}} = u_{N0} + \beta \cdot [p_{\mathbf{N}}^* \cdot (1 - \delta^*) \cdot E_{\mathbf{D}} + (1 - p_{\mathbf{N}}^* \cdot (1 - \delta^*)) \cdot E_{\mathbf{N}}] - p_{\mathbf{N}}^* \cdot c(\delta^*) \quad (17)$$

Firstly, solving the elites' problem shows that the optimal δ_{ω} is independent of state, that is, $\delta_{\mathbf{D}}^* = \delta_{\mathbf{N}}^* = \delta^*$. This is due to the timing of events - elites choose their action after the citizens' have chosen whether to resist and payoffs for that period are realized. This is because in the state of democracy, citizens' actions determine whether elites can enter politics in the first place, and it is only if they do that elites can invest in changing the probability of transitioning to a state of non-democracy. In the state of non-democracy, citizens choose to resist, payoffs are realized and then the elites can invest in changing the probability of retaining complete control over policy and remaining in non-democracy. Secondly, from the optimization problem, we can deduce that δ^* is larger in the liberal democracy equilibrium since the elites invest more in taking over power when they benefit relatively more from non-democracy (Proof in Appendix A.2.3).

Proposition 3. *The elite invest more in the Liberal Democracy equilibrium*

This gives interesting predictions concerning the transition dynamics across the states and political regimes. First, a transition from the state of non-democracy to democracy is likely even if, and more so, in the illiberal democracy equilibrium. The reasoning, while counter-intuitive at first glance, is quite straightforward. If citizen 1 does not resist in democracy, citizens' power is lower which makes it more probable for elites to enter politics. Given the payoff in state $\omega_t = \mathbf{D}$ is that of an illiberal democracy ($u_{0\bar{D}}$), transitions to democracy is not too costly for the elites which makes them invest less in remaining in non-democracy ($\delta_{\mathbf{N}}^*$) or seizing power in democracy ($\delta_{\mathbf{D}}^*$). Hence, we are likely to observe transitions from non-democratic to democratic state, but the regime in the democratic state is more likely to be an illiberal democracy.

On the other hand, transitions out of democracy to non-democracy may be higher in either equilibrium. In liberal democracy equilibrium, the probability of transitioning out of democracy is $(1 - \alpha_{\mathbf{D}}) \cdot \delta^*(r^* = 1)$ whereas in illiberal democracy equilibrium it is $(1 - (1 - \rho) \cdot \alpha_{\mathbf{D}}) \cdot \delta^*(r^* = 0)$. Thus, transitions out of democracy are faster in the illiberal democracy equilibrium if

$$\frac{\delta^*(r^* = 0)}{\delta^*(r^* = 1)} \geq \frac{1 - \alpha_{\mathbf{D}}}{1 - \alpha_{\mathbf{D}} + \rho \cdot \alpha_{\mathbf{D}}}.$$

This holds for $\alpha_{\mathbf{D}}$ sufficiently close to one. This is because for a small $\alpha_{\mathbf{D}}$, resistance of both citizens in the liberal democracy equilibrium is less effective, but δ^* is higher which increases the possibility of transitioning to a non-democracy. Thus, for sufficiently high $\alpha_{\mathbf{D}}$, our model predicts frequent transitions out of illiberal democracy to non-democracy and vice versa. Finally, for a sufficiently high $\alpha_{\mathbf{D}}$, as resistance makes citizens powerful and keeps elites out of politics, transitions out of democracy will be rare in the liberal democracy equilibrium.

These transition dynamics implied by Proposition 3 align with the transition in regime scores illustrated in Figure 1. As discussed in Section 2, we expect to observe frequent transitions to illiberal democracies from non-democracy and vice versa when in the illiberal democracy equilibrium. Thus, our model resonates the systematic transitions between non-democracies

and illiberal democracies that we observe in the real world. Our model also predicts that in the liberal democracy equilibrium, regimes with very high Polity scores are relatively stable and rarely experience decreases in score as shown in Figure 1. Hence, our model allows us to understand the underlying mechanisms that give rise to the transition dynamics across political regimes that we observe in the real world. We highlight that the motive of citizens for protecting liberal democracies is to avoid the worst outcome of an authoritarian regime, even if flow payoffs are slightly lower. This resonates with the rightmost part of Figure 1 where, for very high scores on the Polity index (≥ 0.9), regimes are stable with very little change/decrease in the scores. This can also be observed from the transition matrix in Appendix B.1 where countries with very high scores on the Polity index (≥ 0.9), generally liberal democracies, tend to be stable as most transitions across the period 1900-2020 occurred between non-democracies and illiberal democracies.

While the transition dynamics above are based on the idea that the parameters are fixed across time, illiberalism in established democracies could also potentially arise from parameter changes that result in a switch between the two equilibria. For instance, increased levels of immigration or melting pots of culture could lead to a change in ρ which could lead to certain citizen groups not choosing to resist elites, thus resulting in a switch from the liberal democracy to the illiberal democracy equilibrium. Events of war and conflict could also change demographic structure in ways that lead to a switch in equilibrium. Another interesting possibility is that illiberal democracies arise because of changes in α_N . Suppose α_N starts off low but increases in an unanticipated manner, for instance, due to superior communication technology or strengthening of national solidarity. With this higher α_N , citizen 1 are less fearful of transition to non-democracy, thus leading to a switch to the illiberal democracy equilibrium. On the other hand, changes in citizen's discount factor, for example, increased life expectancy making them more patient could lead to a switch towards the liberal democracy equilibrium. Sudden increase in investment by elites in δ^* due to, say, green movements threatening capitalists could also result in a switch to the liberal democracy equilibrium. The general framework of our model therefore allows exploration of a wide range of scenarios that we leave for future research.

4 Extensions

4.1 Mobility

So far, we have assumed that group affiliation is determined exogenously and there is no movement across groups. However, unlike in the context of say, racial identity, it is often possible for identity to change, the most common example being across income classes. To introduce this idea of mobility between groups, suppose that citizens are mobile in terms of their group affiliation. Let π_{ij} denote the (exogenous) probability of transitioning from group i to group j , where $i, j \in \{1, 2\}$. Then the value functions for citizens satisfy:

$$\mathbf{V} = A\mathbf{u} + \beta \cdot B\mathbf{V}$$

where $\mathbf{V}$ is a vector of state and citizen group specific value functions and A is a matrix of the power distribution that determines the regime and the payoff from the vector $\mathbf{u}$. Each element in matrix B denotes the associated probability of receiving the relevant equilibrium values of vector $\mathbf{V}$. This probability depends on whether a citizen remains in their current group or becomes a member of the other group (that is, on π_{ij}), and the state in the next

period (details in Appendix A.3). Denote $\mathbf{V} = \mathbf{V}(r_{D1}^*, r_{D2}^*, r_{N1}^*, r_{N2}^*)$ to explicitly recognize the dependence of values on equilibrium actions.

To derive equilibrium strategies, first note that it will remain optimal for both groups to resist in non-democracy as long as the payoff ordering based on the assumptions made in Section 3 holds. Given Equation 8⁷, the incentive constraint of citizen group 2 suggests that they will still always prefer to resist in all states. As before, the analysis then boils down to determining the strategy of citizen 1 in the democracy state. To simplify, using the results from Section 3.3, we set $\delta_D = \delta_N = \delta$. Then, citizen 1's preference for resisting as opposed to not resisting, that is, the incentive constraint simplifies to the following, the sign of which determines their decision:

$$-\{u_{\bar{D}1} - u_{D1}\} + \frac{\beta \cdot \delta}{1 - \tilde{\beta}} \cdot [\varphi \cdot \Delta_{u1} + (1 - \varphi) \cdot \Delta_{u2}] \quad (18)$$

where, letting $\tilde{\beta} \equiv \beta \cdot [1 - (1 - p_D) \cdot \delta - \alpha_N \cdot (1 - \delta)]$, we have

$$\varphi \equiv \frac{\pi_{11} - (\pi_{11} + \pi_{22} - 1) \cdot \tilde{\beta}}{1 - (\pi_{11} + \pi_{22} - 1) \cdot \tilde{\beta}} \quad (19)$$

and

$$\Delta_{u1} \equiv (p_D \cdot u_{D1} + (1 - p_D) \cdot u_{\bar{D}1}) - u_{N1} \quad (20)$$

$$\Delta_{u2} \equiv (p_D \cdot u_{D2} + (1 - p_D) \cdot u_{\bar{D}2}) - u_{N2} \quad (21)$$

The first term of Equation 18 is the cost of resisting, that is, a lower flow payoff and the second term is the benefit of resisting, that is, avoiding non-democracy. If the expression is positive (when $p_D = \alpha_D$) then a *liberal democracy equilibrium* arises. If it is negative (when $p_D = (1 - \rho) \cdot \alpha_D$) then an *illiberal democracy equilibrium* arises.

The mobility parameters (π_{11}, π_{22}) have an effect on the decision only through φ . In particular, φ is increasing in π_{11} but decreasing in π_{22} . Thus the effect of ‘mobility’ depends on the specifics. If we assume $\Delta_{u1} < \Delta_{u2}$ ⁸, a higher π_{11} or lower π_{22} will lower the term $[\varphi \cdot \Delta_{u1} + (1 - \varphi) \cdot \Delta_{u2}]$. For instance, consider the case of “general” mobility where $\pi_{11} = \pi_{22} = \pi \in [0.5, 1]$. Here we have

$$\varphi = \left(\frac{\pi - (2\pi - 1) \cdot \tilde{\beta}}{[1 - (2\pi - 1) \cdot \tilde{\beta}]} \right), \quad (22)$$

which is increasing in π . Thus, a higher π (lower general mobility) would lower the benefit of resisting making an illiberal democracy equilibrium more likely.

Consider another case where the proportion of citizens in either group does not change regardless of mobility. Let $\pi_{12} = m$ and $\pi_{21} = \frac{\rho}{1-\rho} \cdot m$. We then get

$$\varphi = \left(\frac{1 - m - (1 - \frac{m}{1-\rho}) \cdot \tilde{\beta}}{1 - (1 - \frac{m}{1-\rho}) \cdot \tilde{\beta}} \right), \quad (23)$$

⁷For the purpose of notational convenience and simplicity, we do not delve into the micro-foundations of payoff determination in this section.

⁸An example of when $\Delta_{u1} < \Delta_{u2}$ holds true - first, assume $u_{N1} = u_{N2}$, that is, in non-democracy, elites treat all citizens similarly. Given this, $\Delta_{u1} \leq \Delta_{u2}$ is satisfied if $\frac{p_D}{(1-p_D)} \geq \frac{u_{\bar{D}1} - u_{\bar{D}2}}{u_{D2} - u_{D1}}$. As p_D is endogenous to citizen 1's decision to resist, a stricter assumption would be $\frac{(1-\rho)\alpha_D}{(1-(1-\rho)\alpha_D)} \geq \frac{u_{\bar{D}1} - u_{\bar{D}2}}{u_{D2} - u_{D1}}$ which would ensure $\Delta_{u1} < \Delta_{u2}$. This assumption thus implies that the benefit from illiberal democracy for citizen 1 relative to citizen 2 compared to the benefit of liberal democracy for citizen 2 relative to citizen 1 is sufficiently small. This holds true for $\frac{(1-\rho)\alpha_D}{(1-(1-\rho)\alpha_D)}$ sufficiently large. However, if we also assume $u_{D1} = u_{D2}$, then we always have $\Delta_{u1} > \Delta_{u2}$. In that case, the predictions from Proposition 4 goes in the opposite direction.

which is decreasing in m as long as $\rho < 1 - \tilde{\beta}$. Thus, a lower m (lower mobility) would lead to a higher φ which again makes an illiberal democracy equilibrium more likely. These examples along with the assumptions allow us insight into the relation between mobility and the emergence of the two possible equilibrium.

Proposition 4. *Assuming $\Delta_{u_1} < \Delta_{u_2}$,*

1. *Greater ‘general’ mobility tends to promote the liberal democracy equilibrium*
2. *Illiberal democracy equilibrium is more likely if there is*
 - (a) *Smaller 1-to-2 mobility (higher π_{11})*
 - (b) *Greater 2-to-1 mobility (lower π_{22})*

Proposition 4 suggests that the effect of “social mobility” is nuanced - as mobility eases group boundaries, it is more likely to promote liberal democracy equilibrium as suggested in de Tocqueville (1835), but it also depends on whether group 1 is more likely to enter group 2 or the reverse. This is because the benefits of illiberal democracy is more pronounced for citizen 1, if they are less likely to become a member of citizen 2 in the future; or, even if they do become citizen 2, if there is a high possibility of reverting back to being citizen 1.

If we consider mobility as economic/social mobility, where elites rank highest, and citizen 2 ranks lowest, then our model also shows that under certain restrictions on payoffs ($\Delta_{u_1} < \Delta_{u_2}$), greater upward mobility promotes illiberal democracies. Qualitatively similar results regarding upward mobility and instability of democracy is found in Acemoglu et al. (2018) where the median voter gives more voice to the poor if there is more downward mobility and to the rich if there is higher upward mobility. However, the mechanism in our paper differs from Acemoglu et al. (2018) since the transitions are determined probabilistically (where the probabilities depend on the strategies) as opposed to the current group in power choosing which group comes to power in the future. The critical distinction in the predictions arise from the fact that our model still predicts transitions across political regimes in an environment with zero mobility (as in the baseline model in Section 3.2).

Proposition 4 is also somewhat in accordance with the POUM (Prospect of Upward Mobility) hypothesis formalized in Benabou and Ok (2001) which suggests that upward mobility leads to policies that do not support the poor. However, in their model ⁹, the poor prefer lower redistribution as they expect to move up the income ladder in the future. The result in our paper is more subtle in the sense that the poor (citizen 2) will support liberal democracy equilibrium regardless of upward mobility. However, those in the middle group status will support an illiberal democracy as long as there is sufficient upward mobility such that they do not get stuck with the low income group in the event that their group affiliation changes in the future. The intuition here is that in the future, if the likeliness of being stuck as citizen 2 is low enough, members of citizen 1 are willing to undermine the other group not only because they will receive a higher flow payoff, but also because the prospect of lower payoffs from being in citizen group 2 is sufficiently low. Another interesting extension to study would be introducing mobility between citizens and elites, but it is a relatively less likely scenario that we leave for future research.

⁹The setup of their model also differs significantly from ours. One of the main distinction is the timing such that individuals actions today affect policy tomorrow. This is unlike our model where today’s action determines today’s regime and thus, policy.

5 Conclusion

Our paper presents a formal economic theory of illiberal democracies in a dynamic setting highlighting how rational actions of citizens can lead to illiberalism. In our paper, illiberalism arises from elites exerting over-sized influence over policy making, facilitated by citizen support that stems from heterogeneity in preferences among citizens. The model suggests that if the benefit from elite involvement is sufficiently high for some citizen groups, an illiberal democracy equilibrium arises where these groups support the illiberal democracy regime despite an increased threat of transitioning to non-democracy. Otherwise, a liberal democracy equilibrium arises where citizens resist elite involvement. The illiberal democracy equilibrium is more probable when elites are weak such that their ability to completely seize power is lower, citizens are impatient. It is also more likely to arise when citizens whose preference align more with the elites plays a smaller role in determining citizens overall power.

We then introduce elites' investment in the probability of seizing complete power and find that they invest more in the liberal democracy equilibrium. This provides interesting insight into the transition dynamics across regimes. Transition from a state of non-democracy to democracy is more likely in the illiberal democracy equilibrium. Furthermore, in the state of democracy, when the collective power of citizens is sufficiently high, transitions out of illiberal democracy to autocracy is also more likely in the illiberal democracy equilibrium relative to the liberal democracy equilibrium. Our model can hence explain the systematic transitions between non-democracies and illiberal democracies observed in the real world.

We also discuss how the framework is applicable to scenarios of differences among citizen groups due to income class and social identity leading to support for democratic, yet illiberal institutions. We further extend our model to allow mobility between groups of citizens which provides nuanced insight into the conditions under which mobility can enhance or restrict possibility of illiberal democracies. Identifying potential causes of illiberal democracies is the first step to safeguard the political institutions of democracies. This is vital as studies suggest that democracies do lead to economic growth (Acemoglu et al., 2019). Potential future research involve constructing valid measures for the extent of illiberalism within democracies and empirically decomposing how much of it is citizen driven as opposed to being driven by elite investment. These findings can inform policy decisions regarding strengthening checks and balances on the government, as well as nation building and unity among communities of different identities.

References

- ACEMOGLU, D., G. EGOROV, AND K. SONIN (2015): “Political economy in a changing world,” *Journal of Political Economy*, 123, 1038–1086.
- (2018): “Social mobility and stability of democracy: Reevaluating de Tocqueville,” *The Quarterly Journal of Economics*, 133, 1041–1105.
- (2020): “Institutional Change and Institutional Persistence,” Tech. rep., National Bureau of Economic Research.
- ACEMOGLU, D., S. NAIDU, P. RESTREPO, AND J. A. ROBINSON (2019): “Democracy does cause growth,” *Journal of political economy*, 127, 47–100.
- ACEMOGLU, D. AND J. A. ROBINSON (2000): “Why did the West extend the franchise? Democracy, inequality, and growth in historical perspective,” *The quarterly journal of economics*, 115, 1167–1199.
- (2001): “A theory of political transitions,” *American Economic Review*, 91, 938–963.
- (2008): “Persistence of power, elites, and institutions,” *American Economic Review*, 98, 267–93.
- ACEMOGLU, D., J. A. ROBINSON, AND R. TORVIK (2013): “Why do voters dismantle checks and balances?” *Review of Economic Studies*, 80, 845–875.
- BARTELS, L. M. (2020): “Ethnic antagonism erodes Republicans’ commitment to democracy,” *Proceedings of the National Academy of Sciences*, 117, 22752–22759.
- BENABOU, R. AND E. A. OK (2001): “Social mobility and the demand for redistribution: the POUM hypothesis,” *The Quarterly Journal of Economics*, 116, 447–487.
- BERMEO, N. (2016): “On democratic backsliding,” *Journal of Democracy*, 27, 5–19.
- BIDNER, C., P. FRANCOIS, AND F. TREBBI (2014): “A theory of minimalist democracy,” Tech. rep., National Bureau of Economic Research.
- BOGAARDS, M. (2018): “De-democratization in Hungary: diffusely defective democracy,” *Democratization*, 25, 1481–1499.
- BONOMI, G., N. GENNAIOLI, AND G. TABELLINI (2020): “Identity, Beliefs, and Political Conflict,” Tech. rep., Working Paper.
- BUCHHEIM, L. AND R. ULBRICHT (2020): “A quantitative theory of political 9 transitions,” *The Review of Economic Studies*, 87, 1726–1756.
- COLUSSI, T., I. E. ISPHORDING, AND N. PESTEL (2021): “Minority salience and political extremism,” *American Economic Journal: Applied Economics*, 13, 237–71.
- COPPEDGE, M., J. GERRING, C. H. KNUTSEN, S. I. LINDBERG, S.-E. SKAANING, J. TEORELL, D. ALTMAN, M. BERNHARD, M. S. FISH, A. CORNELL, ET AL. (2018): “V-Dem dataset v8,” .
- DE TOCQUEVILLE, A. (1835): “Government of the Democracy in America,” *Democracy in America*.

- FOA, R. S. AND Y. MOUNK (2016): “The danger of deconsolidation: the democratic disconnect,” *Journal of democracy*, 27, 5–17.
- GARCIA, M. M. AND C. VON HALDENWANG (2016): “Do democracies tax more? Political regime type and taxation,” *Journal of International Development*, 28, 485–506.
- GRATTON, G. AND B. E. LEE (2020): “Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies,” .
- GROSSMAN, G. M. AND E. HELPMAN (1994): “Protection for Sale,” *The American Economic Review*, 84, 833–850.
- HUQ, A. AND T. GINSBURG (2018): “How to lose a constitutional democracy,” *UCLA L. Rev.*, 65, 78.
- KAUFMAN, R. R. AND S. HAGGARD (2019): “Democratic decline in the United States: What can we learn from middle-income backsliding?” *Perspectives on Politics*, 17, 417–432.
- LEVITSKY, S. AND D. ZIBLATT (2018): *How democracies die*, Broadway Books.
- LÜHRMANN, A. AND S. I. LINDBERG (2019): “A third wave of autocratization is here: what is new about it?” *Democratization*, 26, 1095–1113.
- MERKEL, W., H.-J. PUHLE, A. CROISSANT, AND P. THIERY (2006): “Defekte Demokratie. Band 2: Regionalanalysen,” .
- MUKAND, S. W. AND D. RODRIK (2020): “The political economy of liberal democracy,” *The Economic Journal*, 130, 765–792.
- PERSICO, N. (2021): “A Theory of Non&Democratic Redistribution and Public Good Provision,” .
- PERSSON, T. AND G. TABELLINI (2009): “Democratic capital: The nexus of political and economic change,” *American Economic Journal: Macroeconomics*, 1, 88–126.
- REPUCCI, S. AND A. SLIPOWITZ (2021): *Freedom in the World 2021: Democracy under Siege*, Freedom House.
- SCHEDLER, A. (1998): “What is democratic consolidation?” *Journal of democracy*, 9, 91–107.
- SHAYO, M. (2009): “A model of social identity with an application to political economy: Nation, class, and redistribution,” *American Political science review*, 147–174.

A Appendix A: Mathematical Appendix

A.1 Payoff Ordering

Given Assumption 1, we can see clearly from η_N that since no weight is given to their preferences, citizens receive the lowest payoff in regime N compared to either D or $\tilde{D}$. Elites on the other hand benefit most as they receive the maximum utility possible. Furthermore, by Assumption 2, we have that any influence by the elites will make citizen group 2 worse off relative to when $\epsilon = 0$. Hence, in a straightforward manner, we have

$$\begin{aligned} \text{Elites: } & u_{0D} \leq u_{0\tilde{D}} \leq u_{0N} \\ \text{Citizen 2: } & u_{2N} \leq u_{2\tilde{D}} \leq u_{2D} \end{aligned}$$

Finally, we are left to determine the payoff ordering of citizen 1. Due to the preference alignment with both citizen 2 and elites, let us first define the payoff for citizen 1 in the illiberal democracy regime by plugging the optimal policy in Equation 5.

$$\begin{aligned} u_{1\tilde{D}}(\varepsilon, \rho) &= - \sum_k (\theta_{1k}^* - \epsilon \cdot \theta_{0k}^* - \rho \cdot (1 - \varepsilon) \cdot \theta_{1k}^* - (1 - \rho) \cdot (1 - \varepsilon) \cdot \theta_{2k}^*)^2 \\ &= - \sum_k (\varepsilon \cdot (\theta_{1k}^* - \theta_{\eta_N k}^*) + (1 - \varepsilon) \cdot (\theta_{1k}^* - \theta_{\eta_D k}^*))^2 \\ &= - \sum_k (\varepsilon \cdot (\theta_{1k}^* - \theta_{0k}^*) - (1 - \varepsilon) \cdot (1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*))^2 \\ &= - \sum_k \Theta_k(\varepsilon, \rho)^2. \end{aligned} \tag{24}$$

Here, we have $\frac{du_{1\tilde{D}}}{d\varepsilon} > 0$ for $\varepsilon \in [0, \bar{\varepsilon}_k]$ where

$$\bar{\varepsilon}_k \equiv \frac{(1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*)}{(1 - \rho) \cdot (\theta_{2k}^* - \theta_{1k}^*) + (\theta_{1k}^* - \theta_{0k}^*)}. \tag{25}$$

This shows that citizen 1 benefits from *some* level of elite involvement. However, if elite's influence is too large, then citizen 1 would no longer benefit from elite involvement as policy would resemble that of a non-democracy. Thus, from Assumption 3, we have that ϵ lies within the above range. Furthermore, we get $\sum_k \Theta_k(0, \rho)^2 > \sum_k \Theta_k(\varepsilon, \rho)^2$ as long as $\theta_{1k}^* - \theta_{0k}^* \neq 0$ for some k . Thus, the payoff ordering for citizen 1 is:

$$\text{Citizen 1: } u_{1N} \leq u_{1D} \leq u_{1\tilde{D}} \tag{26}$$

A.2 Equilibrium

A.2.1 Proof of Proposition 1

Given $\{V_D, V_N\}$, r_1^* is optimal if for all $p_D \in [\alpha_D \cdot (1 - \rho), \alpha_D]$ we have:

$$V_D \geq p_D \cdot u_{D1} + (1 - p_D) \cdot u_{\tilde{D}1} + \beta \cdot [(1 - (1 - p_D) \cdot \delta_D) \cdot V_D + (1 - p_D) \cdot \delta_D \cdot V_N]$$

which is

$$[\alpha_D \cdot \rho \cdot (r_1^* - r_1)] \cdot [-(u_{\tilde{D}1} - u_{D1}) + \beta \delta_D \cdot (V_D - V_N)] \geq 0 \quad \forall r_1 \in [0, 1].$$

Let $\pi_{\mathbf{N}} = 1 - \alpha_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})$. Since $\alpha_{\mathbf{D}} \cdot \rho$ is positive, the incentive constraint simplifies to¹⁰:

$$(r_1^* - r_1) \cdot [\beta \cdot \delta_{\mathbf{D}} \cdot (u_{\tilde{D}1} - u_{N1}) - (1 + \beta \cdot [\delta_{\mathbf{D}} - \pi_{\mathbf{N}}]) \cdot (u_{\tilde{D}1} - u_{D1})] \geq 0 \quad \forall r_1 \in [0, 1].$$

This gives two Markov Perfect equilibria - one with $r_1^* = 1$ and one with $r_1^* = 0$.

Hence, a symmetric Markov Perfect Equilibrium exists for $r_1^* \in \{0, 1\}$. An illiberal democracy equilibrium exists if and only if

$$-(u_{i\tilde{D}} - u_{iD}) + \Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \cdot (u_{i\tilde{D}} - u_{iN}) \leq 0 \quad (27)$$

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \cdot \leq \frac{u_{1\tilde{D}} - u_{1D}}{u_{1\tilde{D}} - u_{1N}} \quad (28)$$

A Liberal Democracy equilibrium exists iff

$$-(u_{i\tilde{D}} - u_{iD}) + \Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \cdot (u_{i\tilde{D}} - u_{iN}) \geq 0 \quad (29)$$

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \cdot \geq \frac{u_{1\tilde{D}} - u_{1D}}{u_{1\tilde{D}} - u_{1N}} \quad (30)$$

Plugging in the optimal policy and simplifying from 24, we have

$$\Lambda \equiv \frac{u_{1\tilde{D}} - u_{1D}}{u_{1\tilde{D}} - u_{1N}} \quad (31)$$

$$\Lambda(\varepsilon, \rho) = \frac{\sum_k \Theta_k(0, \rho)^2 - \Theta_k(\varepsilon, \rho)^2}{\sum_k \Theta_k(1, \rho)^2 - \Theta_k(\varepsilon, \rho)^2} \quad (32)$$

and from solving 9 and 10 we have

$$\Gamma(\delta_{\mathbf{D}}, \delta_{\mathbf{N}}, \alpha_{\mathbf{N}}, \beta) \equiv \frac{\beta \cdot \delta_{\mathbf{D}}}{\beta \cdot \delta_{\mathbf{D}} + (1 - \beta) + \beta \cdot \alpha_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})} \quad (33)$$

Together, these give us Proposition 1.

A.2.2 Proof of Proposition 2

The illiberal democracy equilibrium more readily arises as mentioned in Proposition 2 when

- The elites are weak since $\Gamma_{\delta_{\mathbf{D}}} > 0$, $\Gamma_{\delta_{\mathbf{N}}} > 0$, and $\Gamma_{\alpha_{\mathbf{N}}} < 0$.
- Citizens are impatient as $\Gamma_{\beta} > 0$
- Citizen 1 plays a smaller role in determining citizens' power, that is, for a lower ρ as $\Lambda_{\rho} \leq 0$. To see this, let $D_{21k} = \sum_k (\theta_{2k}^* - \theta_{1k}^*)^2$, $D_{10k} = \sum_k (\theta_{1k}^* - \theta_{0k}^*)^2$ and $d_k = \sum_k (\theta_{2k}^* - \theta_{1k}^*)(\theta_{1k}^* - \theta_{0k}^*)$. For any k, we have,

$$\frac{d\Lambda}{d\rho} = - \frac{\left(4\varepsilon(1 - \rho)D_{21k}D_{10k} + 2\varepsilon D_{10k}d_k + 2\varepsilon(1 - \rho)^2 D_{21k}d_k \right)}{(1 - \varepsilon) \left((1 + \varepsilon)D_{10k} + 2\varepsilon(1 - \rho)d_k - (1 - \varepsilon)(1 - \rho)^2 D_{21k} \right)^2}$$

Thus, $\frac{d\Lambda}{d\rho} \leq 0$ since all the terms in the numerator and the denominator are positive given the parameters.

¹⁰Since the denominator $1 + \beta \cdot [(1 - p_{\mathbf{D}}^*) \cdot \delta_{\mathbf{D}} - \pi_{\mathbf{N}}]$ is positive, it can be factored out.

A.2.3 Proof of Proposition 3

First of all, the FOC from the elites' optimization problem 15 is:

$$\beta \cdot [E_{\mathbf{N}} - E_{\mathbf{D}}] = c'(\delta^*) \quad (34)$$

This clearly shows that optimal investment in δ^* is independent of state.

The difference in $E_{\mathbf{D}}$ and $E_{\mathbf{N}}$ from 16 and 17 simplifies to

$$E_{\mathbf{N}} - E_{\mathbf{D}} = \frac{v_0 + [1 - p_{\mathbf{D}}^* - p_{\mathbf{N}}^*] \cdot c(\delta^*)}{1 - \beta \cdot (1 - p_{\mathbf{N}}^*) + \beta \cdot [1 - p_{\mathbf{D}}^* - p_{\mathbf{N}}^*] \cdot \delta^*}$$

Plugging this into the above FOC yields:

$$\beta \cdot v_0 = [1 - \beta \cdot (1 - p_{\mathbf{N}}^*)] \cdot c'(\delta^*) + \beta \cdot [1 - p_{\mathbf{D}}^* - p_{\mathbf{N}}^*] \cdot [c'(\delta^*) \cdot \delta^* - c(\delta^*)]. \quad (35)$$

where $v_0 \equiv u_{N0} - (p_{\mathbf{D}}^* \cdot u_{D0} + (1 - p_{\mathbf{D}}^*) \cdot u_{\tilde{D}0})$ which is independent of δ^* . The right side starts at zero and is strictly increasing and goes to infinity as δ^* goes to one. Thus, there exists a unique value of δ^* which solves this condition and is interior. To make meaningful connections to the equilibrium discussed in Proposition 1, Equation 35 clearly shows that higher values of $p_{\mathbf{D}}^*$ lower the right hand side and raise the left hand side and therefore raise δ^* .

A.3 Mobility

The citizens' value function is as follows:

$$\mathbf{V} = \mathbf{A}\mathbf{u} + \beta \cdot \mathbf{B}\mathbf{V}$$

where

$$\mathbf{V} \equiv \begin{bmatrix} V_{D1} \\ V_{D2} \\ V_{N1} \\ V_{N2} \end{bmatrix}, \mathbf{A} \equiv \begin{bmatrix} p_{\mathbf{D}} & 0 & 1 - p_{\mathbf{D}} & 0 & 0 & 0 \\ 0 & p_{\mathbf{D}} & 0 & 1 - p_{\mathbf{D}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{u} \equiv \begin{bmatrix} u_{D1} \\ u_{D2} \\ u_{\tilde{D}1} \\ u_{\tilde{D}2} \\ u_{N1} \\ u_{N2} \end{bmatrix}$$

and

$$\mathbf{B} \equiv \begin{bmatrix} \pi_{11} \cdot [1 - (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{12} \cdot [1 - (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{11} \cdot [(1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{12} \cdot [(1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] \\ \pi_{21} \cdot [1 - (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{22} \cdot [1 - (1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{21} \cdot [(1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] & \pi_{22} \cdot [(1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}] \\ \pi_{11} \cdot [p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{12} \cdot [p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{11} \cdot [1 - p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{12} \cdot [1 - p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] \\ \pi_{21} \cdot [p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{22} \cdot [p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{21} \cdot [1 - p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] & \pi_{22} \cdot [1 - p_{\mathbf{N}} \cdot (1 - \delta_{\mathbf{N}})] \end{bmatrix}$$

where $p_{\mathbf{D}}$ and $p_{\mathbf{N}}$ are defined as before. As mentioned earlier, each element in matrix $\mathbf{B}$ denotes the associated probability of receiving the relevant equilibrium values of matrix $\mathbf{V}$. For instance, the first element in row 1 of matrix $\mathbf{B}$ is the probability that non-elite 1 remains so in the next period, and the state continues to be a democracy; the fourth element of the first row, $\pi_{12} \cdot [(1 - p_{\mathbf{D}}) \cdot \delta_{\mathbf{D}}]$, is the probability that an individual in non-elite 1 becomes affiliated to non-elite 2, and the state becomes a non-democracy, and so on.

Thus, the value functions are given by

$$\mathbf{V} = [\mathbf{I} - \beta \cdot \mathbf{B}]^{-1} \mathbf{A}\mathbf{u}$$

where I is the identity matrix where $\mathbf{V} = \mathbf{V}(r_{D1}^*, r_{D2}^*, r_{N1}^*, r_{N2}^*)$. Let $X_{[i]}$ denote the i^{th} row of matrix X .

Strategies (r_{D1}^*, r_{D2}^*) are part of an equilibrium if

$$\mathbf{V}_{[i]}(r_{D1}^*, r_{D2}^*) \geq A_{[i]}(r_{Di}, r_{D,-i}^*)\mathbf{u} + \beta \cdot B_{[i]}(r_{Di}, r_{D,-i}^*)\mathbf{V}(r_{D1}^*, r_{D2}^*, 1, 1).$$

Incentive constraint for strategies (r_{D1}^*, r_{D2}^*) :

$$[A_{[i]}(r_{Di}^*, r_{D,-i}^*) - A_{[i]}(r_{Di}, r_{D,-i}^*)]\mathbf{u} + \beta \cdot [B_{[i]}(r_{Di}^*, r_{D,-i}^*) - B_{[i]}(r_{Di}, r_{D,-i}^*)]\mathbf{V}(r_{D1}^*, r_{D2}^*, 1, 1) \geq 0.$$

Notice that for $i = 1$ this is:

$$[A_{[1]}(r_{D1}^*, r_{D,2}^*) - A_{[1]}(r_{D1}, r_{D,2}^*)] = \alpha_{\mathbf{D}} \cdot \rho \cdot (r_{D1}^* - r_{D1}) \cdot [1, 0, -1, 0, 0, 0]$$

and

$$[B_{[1]}(r_{D1}^*, r_{D,2}^*) - B_{[1]}(r_{D1}, r_{D,2}^*)] = \alpha_{\mathbf{D}} \cdot \rho \cdot (r_{D1}^* - r_{D1}) \cdot \delta_{\mathbf{D}} \cdot [\pi_{11}, \pi_{12}, -\pi_{11}, -\pi_{12}]$$

Thus $\alpha_{\mathbf{D}} \cdot \rho$ can be factored out and dropped (since $\alpha_{\mathbf{D}} \cdot \rho > 0$). Hence, for $i = 1$, we have the incentive constraint:

$$(r_{D1}^* - r_{D1}) \cdot \{u_{D1} - u_{\tilde{D}1} + \beta \cdot \delta_{\mathbf{D}} \cdot [\pi_{11} \cdot (V_{D1} - V_{N1}) + \pi_{12} \cdot (V_{D2} - V_{N2})]\} \geq 0. \quad (36)$$

and for $i = 2$, it is:

$$(r_{D2}^* - r_{D2}) \cdot \{u_{D2} - u_{\tilde{D}2} + \beta \cdot \delta_{\mathbf{D}} \cdot [\pi_{21} \cdot (V_{D1} - V_{N1}) + \pi_{22} \cdot (V_{D2} - V_{N2})]\} \geq 0. \quad (37)$$

Given Equation 8, in Equation 37, the term in braces is always positive which implies that citizen 2 still always prefers to resist in all states. As before, the analysis boils down to determining the strategy of citizen 1 in the democracy state.

The main object of interest in 36 is the sign of the following:

$$-\{u_{\tilde{D}1} - u_{D1}\} + \beta \cdot \delta_{\mathbf{D}} \cdot [\pi_{11} \cdot (V_{D1} - V_{N1}) + \pi_{12} \cdot (V_{D2} - V_{N2})] \quad (38)$$

To simplify, set $\delta_{\mathbf{D}} = \delta_{\mathbf{N}} = \delta$:

$$V_{D1} - V_{N1} = \xi_1 \cdot \Delta_{u_1} + \xi_2 \cdot \Delta_{u_2} \quad (39)$$

$$V_{D2} - V_{N2} = \xi_3 \cdot \Delta_{u_1} + \xi_4 \cdot \Delta_{u_2} \quad (40)$$

where

$$\xi_1 \equiv \frac{1 - \tilde{\beta} \cdot \pi_{22}}{Z} \quad (41)$$

$$\xi_2 \equiv \frac{\tilde{\beta} \cdot \pi_{12}}{Z} \quad (42)$$

$$\xi_3 \equiv \frac{\tilde{\beta} \cdot \pi_{21}}{Z} \quad (43)$$

$$\xi_4 \equiv \frac{1 - \tilde{\beta} \cdot \pi_{11}}{Z} \quad (44)$$

$$Z \equiv [\pi_{11} + \pi_{22} - 1] \cdot \tilde{\beta}^2 - [\pi_{11} + \pi_{22}] \cdot \tilde{\beta} + 1 \quad (45)$$

$$= (1 - \tilde{\beta}) \cdot (1 - (\pi_{11} + \pi_{22} - 1) \cdot \tilde{\beta}) \quad (46)$$

$$\tilde{\beta} \equiv \beta \cdot [1 - (1 - p_{\mathbf{D}}) \cdot \delta - \alpha_{\mathbf{N}} \cdot (1 - \delta)]. \quad (47)$$

The main condition (38) thus becomes

$$-\{u_{\tilde{D}1} - u_{D1}\} + \beta \cdot \delta \cdot [\pi_{11} \cdot [\xi_1 \cdot \Delta_{u_1} + \xi_2 \cdot \Delta_{u_2}] + \pi_{12} \cdot [\xi_3 \cdot \Delta_{u_1} + \xi_4 \cdot \Delta_{u_2}]] \quad (48)$$

Use $\pi_{12} = 1 - \pi_{11}$ and simplify to get

$$-\{u_{\tilde{D}1} - u_{D1}\} + \beta \cdot \delta \cdot [\pi_{11} \cdot [[\xi_1 - \xi_3] \cdot \Delta_{u_1} + [\xi_2 - \xi_4] \cdot \Delta_{u_2}] + [\xi_2 \cdot \Delta_{u_1} + \xi_4 \cdot \Delta_{u_2}]] \quad (49)$$

which is

$$-\{u_{\tilde{D}1} - u_{D1}\} + \beta \cdot \delta \cdot \left[\left(\frac{\pi_{11} \cdot (1 - \tilde{\beta}) + \tilde{\beta} \cdot (1 - \pi_{22})}{Z} \right) \cdot \Delta_{u_1} + \left(\frac{1 - \pi_{11}}{Z} \right) \cdot \Delta_{u_2} \right] \quad (50)$$

B Appendix B

B.1 Transitions across Political Regimes

Using Regimes of the World (RoW) classification of the VDem dataset (Coppedge et al., 2018) and regimes constructed from PolityV concept variables¹¹, Table 2 shows the transition probability matrix across regimes, along with frequencies. The first block provides the transition matrix for the period 1900-2020 and the next three blocks are constructed for different time periods in history. The rows reflect the initial regime, and the columns represent the final regime. For instance, from the first row in the first block, the probabilities show that each year, starting from a non-democracy, the probability of remaining a non-democracy is 99%, the probability of transitioning to an electoral democracy is 1% and there is 0 probability of transitioning to a liberal democracy. The frequencies show the total number of transitions within regimes across the relevant sample.

Table 2: Transition probability and frequency matrix with time period breakdown

VDem Measure												
1900-2020				1918-1945			1946-1991			1992-2020		
Probabilities	ND	ED	LD	ND	ED	LD	ND	ED	LD	ND	ED	LD
ND	0.99	0.01	0.00	0.99	0.01	0.00	1.00	0.00	0.00	0.99	0.01	0.00
ED	0.04	0.95	0.01	0.05	0.94	0.01	0.13	0.87	0.00	0.02	0.95	0.02
LD	0.00	0.01	0.99	0.00	0.00	1.00	0.04	0.02	0.93	0.00	0.00	1.00
Frequencies	ND	ED	LD	ND	ED	LD	ND	ED	LD	ND	ED	LD
ND	13719	161	14	2431	15	3	800	0	0	5426	57	5
ED	90	2331	36	11	207	2	4	26	0	16	618	14
LD	3	18	2262	0	0	173	2	1	42	0	0	885

Polity Measure												
1900-2020				1918-1945			1946-1991			1992-2020		
Probabilities	ND	ED	LD	ND	ED	LD	ND	ED	LD	ND	ED	LD
ND	0.98	0.02	0.00	0.98	0.01	0.00	0.98	0.02	0.00	0.96	0.04	0.00
ED	0.03	0.96	0.01	0.04	0.94	0.01	0.06	0.93	0.01	0.02	0.98	0.00
LD	0.00	0.01	0.99	0.01	0.01	0.98	0.00	0.00	0.99	0.00	0.01	0.99
Frequencies	ND	ED	LD	ND	ED	LD	ND	ED	LD	ND	ED	LD
ND	6144	137	16	984	11	4	2986	64	7	1398	52	0
ED	109	3361	23	15	340	5	62	986	10	29	1799	8
LD	6	14	2319	2	4	345	2	4	948	1	5	891

¹¹From the Polity dataset, regimes are classified based on combination of executive selection score (exrec) and political competition (polcomp). A regimes is defined as high in electoral category if they score at least second highest score in exrec, and highest in polcomp. A country is defined as liberal democracy if a nation sores highest in both category and electoral democracy if high only in electoral category and non democracy if low in both categories. This classification allows for the most overlap with VDem RoW classification.