Democratic Values and Institutions*

Timothy Besley
LSE and CIFAR

Torsten Persson
IIES and CIFAR

August 2018

Abstract

This paper builds a model of the two-way interaction between democratic values and institutions to bridge sociological research focusing on values with economics research which studies strategic decisions. Some citizens hold values that make them protest to preserve democracy with the share of such citizens evolving endogenously over time. There is then a natural complementarity between values and institutions creating persistence without assuming any form of commitment. The approach unifies ideas in the literature, explains observed patterns in the data on democratic values and political institutions and suggests new insights into sources of heterogeneity in values.

*We are grateful to the editor Larry Samuelsson and three anonymous referees, as well as Philippe Aghion, Roland Bénabou, Cameron Hepburn, Guido Tabellini, Peyton Young, participants in seminars at CIFAR, Stanford, Kings College, IIES, Bocconi, Tsinghua, Nottingham, Georgetown, Warwick, the EEA Congress, and the conferences on Culture, Institutions and Prosperity, and Political Economy and Climate for helpful comments. We gratefully acknowledge financial support from the Swedish Research Council and the European Research Council.

"(I)f a political system is not characterized by a value system allowing the peaceful 'play' of power ... there can be no stable democracy." Lipset (1959, p. 71)

"During the nineteenth century most Western societies extended voting rights, ... these political reforms can be viewed as strategic decisions by the political elite to prevent widespread social unrest and revolution." Acemoglu and Robinson (2000, p. 1167).

1 Introduction

Looking across today's world and its history, the heterogeneity of democratic experiences is striking. Some polities have made secure transitions into democracy and these institutions are accepted pretty much by everybody. Others have never secured democracy. A third group occupies a middle ground with a history punctuated by protests and institutional reversals, and occasional transitions to the stable groups.

Understanding what drives democratic reforms is important intrinsically, as well as instrumentally – a body of research gives political institutions a central role in explaining cross-country differences in economic growth and development (e.g., North 1983).

The initial quotes illustrate two approaches to democratic reform in the social sciences. Recent research in economics argues that democratic institutions and reforms are the result of strategic, forward-looking decisions by dominant groups. An older body of research in political science and sociology holds that democratic values play a key rule in inducing and supporting democratic institutions. Although both approaches highlight important drivers of democracy, few have attempted to join them and investigate whether this generates new insights.

In this paper, we model the drivers of democratic reforms with dynamic democratic values as well as strategic decisions – including costly decisions to fight – by prospective winners and losers from institutional reform. Neither institutions nor values have an upper hand in the process of democratic change: the two evolve jointly and interdependently.

The now standard model of institutional change from Acemoglu and Robinson (2000, 2006) assumes that decision-makers can commit institutions one or more periods ahead. We dispense with any commitment assumption: institutional reforms are sustainable only if they are incentive compatible for the *current* incumbent. Democratic values is the single slow-moving state variable which sustains persistent change. The model allows us to readily interpret the broad patterns of democratic reforms and democratic values found in the Polity IV (PIV) and World Value Survey (WVS) data. But it also generates new predictions, including the effects on values of foreign occupations, via colonialism or the Cold War. We present some within-country correlations from the WVS consistent with these auxiliary model predictions.

The next section selectively overviews different approaches to democratic institutions and provides background facts about the dispersion of democratic institutions and values over countries and time. Section 3 sketches a simple model of the interplay between democratic institutions and democratic values. Section 4 shows how this model helps us interpret the patterns of institutional dynamics and values in the data, unifies apparently diverse ideas in the existing literature, and pinpoints auxiliary predictions which are consistent with the data. Section 5 concludes. An Online Appendix collects supporting materials.

2 Background

2.1 Related ideas

Cultural, value-based arguments for democracy go back to Aristotle. But the *locus classicus* is Montesquieu (1748) who spells out how geography and climate interact with culture to shape how alternative political institutions work. In modern political science, Lipset (1959) and Almond and Verba (1963) pioneer the argument that political culture and values are vital pre-requisites for democracy.

These ideas has influenced the measurement of values and attitudes, at least since Inglehart (1997). Drivers and consequences of values are subjects of an evolving literature, which has argued that mass attitude as measured in the WVS, gauge the demand for democratic change (Inglehart and Welzel 2005) and demonstrate the willingness to struggle for democracy (Welzel 2007). Fuchs-Schündeln and Schündeln (2015) show that individual experience with democracy raises support for democracy, while Neundorf (2010) exploits political attitudes from eastern Europe to show that such support is considerably weaker for individuals who grew up during the cold war. Gorodnichenko and Roland (2015) emphasize why individualistic rather than collectivist cultures are more likely to underpin democratization.

Almond and Verba (1963, p. 367) discuss how civic culture is shaped by socialization, which "includes training in many social institutions – family, peer group, school, work place, as well as in the political system itself". Our approach builds on models of cultural evolution beginning with Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985).

Research on culture, individual behavior, and institutions has increased among economists in recent years; see the overview in Bisin and Verdier (2011). We model cultural change through the dynamics of preferences or values (rather than behavior or beliefs) following the indirect evolutionary approach of Güth and Yaari (1992).

Acemoglu and Robinson (2000, 2006) suppose that an elite uses the franchise as a commitment device to guarantee the masses more favorable policy treatment. On top of the case studies in these works, Aidt and Jensen (2014) and Aidt and Franck (2015) provide supportive econometric evidence. Our modeling follows Acemoglu and Robinson except in one crucial dimension. In their model, political institutions is a state variable causing persistence, on the argument that they are harder to change than economic policies. In our model, by contrast, democratic values are the only state variable, on the argument that they move slower than institutions.

Closest to our approach is Ticchi, Verdier, and Vindigni (2013) who model the interaction between value formation and political reforms, giving an explicit role to education. Their model has two state variables and assumes, in common with the earlier literature, that political institutions can be committed one period ahead. Studying the coevolution of institutions and culture, Bisin and Verdier (2017) also make this assumption.

Our approach is also akin to Weingast (1997) who shows how rights can emerge as a self-enforcing equilibrium, and Lagunoff (2001) who shows how greater political turnover raises support for civil liberties. However, neither has a role for democratic values.

2.2 Motivating Facts

The model links two sets of facts: the heterogeneity in country-level democratic histories and the covariation of democratic values with democratic histories.

We gauge each country's democratic history from the PIV, classifying a country as democratic if the *polity2* variable – measured on a 20-step scale from -10 to +10 – is greater than zero. When documenting the patterns of democratic reforms, we confine ourselves to the 50 countries that appear

in the PIV data in each year from 1875 onwards. We summarize the heterogeneity of country dynamics as follows:

Institutions *Histories of democratic reforms come in three broad forms: always non-democratic, permanent transition to democracy, or churning between the two, with the churning group the most prevalent one.*

Table 1 illustrates these facts, classifying each country according to its history. The left-most column shows that three of the 50 countries have never been democratic. The top of the right-most column shows a striking institutional longevity in countries with democracy from the outset (or from 1800) although transitions to democracy are more recent in countries at the bottom of the right-most column, except for Costa Rica and Sweden. Countries with transitions in both directions, in the middle column, is the largest group.

If we extend this table to all PIV countries, all columns have more entries. A few countries, like South Korea and Taiwan, have made single transitions to democracy while others, like Gambia or Somalia, have made single transitions in the other direction. However, as in Table 1, most countries fall into the mixed category.

To study democratic values, we use data from WVS waves 5 and 6. V. 140 asks people to rate the importance of democracy on a ten-point scale. We adopt a binary indicator: someone has (strong) democratic values if she rates democracy strictly above 8. This variable, with a global mean of about 0.6, reveals that

Values *Support for democracy varies across individuals and countries, with strongest (weakest) support in countries with long (short) histories of democracy.*

To illustrate these facts, the left panel in Figure 1 shows a positive relation between a country's share of people with democratic values (relative to the global mean) and its fraction of democratic years. The middle panel shows a similar relation, while conditioning on individual gender, education, age, and income (see notes to the figure). The right panel shows that democratic support is about 25% higher in countries with a once-and-for-all entry into democracy (the right column of Table 1) rather than a mixed history (the left and middle columns).[1]

3 Model

Our framework highlights a conflict of interest over democratic institutions between an incumbent group (a "political elite") and its opposition. In each period, the incumbent chooses whether to install a democracy or an autocracy, without being able to commit society to future institutions. The *only* state variable in the model is the proportion of individuals with democratic values, who may stand up for democracy against autocracy.

Groups and payoffs There are two groups of equal size, each normalized to measure 1. Their roles may shift across periods, as indicated by $G \in \{I, O\}$ with I denoting the incumbent and O the opposition.[2] Institutions are denoted by $D_t \in \{0, 1\}$ where $D_t = 1$ is democracy and $D_t = 0$

[1] Fuchs-Schündeln and Schündeln (2015) use a country-fixed-effects regression with WVS data to show that eight more years of exposure to democracy raises individual support for democrcay by the equivalent of secondary (rather than primary) school education.

[2] The assumption of two groups with equal size is for analytical convenience. Other assumptions – e.g allowing for multiple groups, or letting the incumbent elite have neglible size – would produce similar qualitative results.

autocracy. Payoffs depend on this institutional indicator and on the realization of random variable $x_t \in [\underline{x}, \overline{x}]$, with distribution function $H(\cdot)$.

At realization x and institution D, group G's material payoff is denoted by $u^G(x, D)$,which we assume is (weakly) increasing in x. We make the following assumptions:

$$u^I(x,0) - u^I(x,1) = \Gamma(x) > 0 \text{ is increasing in } x \text{ for all } x \in [\underline{x}, \overline{x}]$$

and

$$u^O(x,1) - u^O(x,0) = \gamma(x) > 0 \text{ is increasing in } x \text{ for all } x \in [\underline{x}, \overline{x}]. \quad (1)$$

A higher value of x implies a greater incentive for the incumbent to maintain $D_t = 0$ and a greater value to the opposition of $D_t = 1$.

Institutional interpretation Why is this a plausible reduced-form model of democracy? Crucially, D_t captures a basic conflict of interest over the private material payoffs under alternative political institutions: incumbents prefer autocracy but oppositions prefer democracy. More specifically, the Web Appendix sketches two examples that both provide a simple microfoundation for the reduced form. Each example focuses on one core element of democratic institutions. Thus, the first highlights constraints on executive power – here, x_t represents some (resource) rents to be split between the two groups at t. The second example instead highlights relatively open access to executive power – here, x_t represents the incumbent's current unpopularity, the probability that the opposition would win an electoral contest at t. However, a similar framework could also be used to model the sustainability of any institutional arrangement that favors one group over another.

Types, democratic values, and fairness Citizens are of two types, the shares of which are endogenous. Fraction $1 - \mu_t$ are *passive* (type P) – if they protest, this is only due to private gains. Their date-t utility is $u^O(x_t, D_t)$. The remaining fraction, μ_t are *concerned* (type C) – a prospective civil society willing to support democracy – who care about the payoffs of society at large.[3] Concerned-citizen payoffs are $u^O(x_t, D_t) + s(x_t, D_t)$ with

$$s(x_t, D_t) = \begin{cases} \gamma(x) & \text{if } D_t = 1 \\ -\chi\gamma(x) & \text{if } D_t = 0, \end{cases} \quad (2)$$

where (2) gives a positive payoff if $D_t = 1$, a negative one if $D_t = 0$,and parameter $\chi \geq 1$ represents loss aversion by concerned citizens. These reference-dependent social preferences (Kahneman and Tversky 1979) capture how citizens value political rights. As we discuss in the Web Appendix, it can be microfounded by concerned citizens judging the outcome as a gain or loss relative to their preferred institution.[4] The formulation makes democratic values distinct from standard preferences, as in the distinction between acquisition utility and transactions utility, which can also reflect a sense of justice (Thaler 1999).

We assume that concerned citizens are equally distributed across the two groups. Democratic values serve two roles. They can motivate concerned citizens to protest. They also affect the "psychological fitness" of such citizens relative to passive citizens, because – beyond material payoffs –

[3] Democratic values are universal rather than particularistic. The complementarity of institutions and values that we emphasise below would be stronger still if concerned citizens had "tribal preferences", i.e. cared only about the payoffs of other concerned citizens.

[4] Our formulation follows Loomes and Sugden (1982) where an individual experiences either regret or rejoices depending on her reference point. This formulation is related to Passarelli and Tabellini (2017), who consider how values underpin citizens' willingness to protest against policies they regard as unfair.

concerned citizens rejoice when they have democratic rights, but despair otherwise.

Concerned citizens and incumbent fighting A successful protest can impose democracy via a successful coup or social pressure.

If a protest involves a fraction ϕ_t of citizens in period t, then the probability of success is $\phi_t p(f_t)$. Here, f_t are the resources that the incumbent devotes to preventing or fighting the protest, at a cost of wf_t.[5] This is consistent with a complementarity in collective action with a greater return to protesting when more citizens join in.[6]

Protests have a random binary cost, which is common to all individuals and denoted by $c_t \in \{\underline{c}, \bar{c}\}$ where ρ is the probability of low protest costs $c_t = \underline{c}$. Draws of c_t are iid over time. Assume that

$$\gamma(x) < \underline{c} < [2+\chi]\gamma(x)p(f) < \bar{c} \text{ for all } x \in [\underline{x}, \overline{x}] \text{ and } f \geq 0, \tag{3}$$

so that material gains are never sufficient to induce protest while democratic values can be. We assume that concerned citizens in the incumbent group never protest in support of democracy.[7] Also, function $p(\cdot)$ is decreasing and log convex, with $p(0) = 1$ and $\lim_{f\to 0} p'(0) = -\infty$ so that it is always worth devoting some resources to fighting a citizen-protest.

Democratic values transmission Over time, values follow an evolutionary dynamic based on a revision protocol (Sandholm 2010). Formally, the protocol is a continuous function $\varsigma^{I,J}(\Delta, \mu_t) \in [0,1]$, which specifies a conditional switching rate from type I to J. Sandholm (2010) suggests a general class of dynamics that yield:

$$\mu_{t+1} - \mu_t = (1-\mu_t)\varsigma^{P,C} - \mu_t \varsigma^{C,P}, \tag{4}$$

where

$$\varsigma^{P,C} > 0 \iff \Delta > 0 \text{ and } \varsigma^{C,P} > 0 \iff \Delta < 0.$$

We call Δ the *relative* (*psychological*) *fitness* – the *expected gain or loss* – of being a concerned citizen. The evolution of values has a "Darwinian" element: if concerned citizens have strictly higher (lower) payoffs than passive citizens, their share in the population increases (decreases) over time. The sign of $\Delta_\mu(\mu)$ affects the equilibrium dynamics (see further below).

The Web Appendix shows that (4) can be given microfoundations, where parents socialize their children (strategically or non-strategically). It can also be derived from replicator-dynamic setting, where the young are influenced by "cultural parents" and imitate more successful types.[8]

Timing The timing within a generation has four steps:

1. A leader in generation t is selected from incumbent group I, and x_t is realized.

2. This leader chooses D_t and f_t.

[5] We do not allow the incumbent to buy off protesters, although this would lead to similar trade-offs.

[6] There could be a further complementarity if the cost of protest (per concerned citizen) would decrease with the number of participants.

[7] This could be rationalized by supposing there is a higher protest cost for such citizens due to within-group peer pressure.

[8] Depending on the exact model, relative fitness can depend either on tomorrow's share of concerned citizens, $\Delta(\mu_{t+1})$, or today's share, $\Delta(\mu_t)$. However, the steady states of the model do not depend on this detail.

3. Under democracy $D_t = 1$, the payoffs are $u^G(x_t, 1)$ for $G \in \{I, O\}$
   Under autocracy $D_t = 0$, c_t is realized and citizens decide whether to protest. With an unsuccessful protest, payoffs are $u^G(x_t, 0)$ for $G \in \{I, O\}$. A successful protest imposes $D_t = 1$ and payoffs $u^G(x_t, 1)$ for $G \in \{I, O\}$.

4. Payoffs are realized, a new generation is born and socialized, changing μ_t to μ_{t+1}. A non-unseated incumbent stays until period $t+1$. With an unseated incumbent (successful protest), the opposition at t becomes the new incumbent at $t+1$.

Preliminaries The Web Appendix analyzes optimal fighting and protesting at stages 2 and 3. Based on these choices, we define two functions $V(x_t, \mu_t)$ and $U(x_t)$ for the incumbent's equilibrium payoffs under autocracy and democracy, respectively, and a survival function $\lambda(x, \mu)$, for the expected probability of successfully enforcing $D_t = 0$ with optimal fighting on both sides. We show that for all $\mu \in [0, 1]$ and $x \in [\underline{x}, \overline{x}]$, a higher x increases $\lambda(x, \mu)$ and $V(x, \mu) - U(x)$ .That is, a higher x raises the incumbent group's gain from remaining in office and its benefit to fighting – it thus makes autocracy more attractive. A larger share of concerned citizens μ has the opposite effects: it decreases expected survival $\lambda(x, \mu)$ and the equilibrium gain from autocracy $V(x, \mu) - U(x)$.

For Proposition 1, we also need

Assumption 1 (i) *The payoff functions satisfy* $V(x, 1) - U(x) < 0$, *and* (ii) *there exists* $\underline{\mu} > 0$ *such that* $V(\underline{x}, \underline{\mu}) - U(\underline{x}) = 0$.

In this assumption, (i) says that it is never worthwhile to maintain autocracy if all citizens are concerned, while (ii) says that μ has a lower bound, which makes the incumbent indifferent between autocracy and democracy at the lowest realization of x. A necessary condition for (ii) is that concerned citizens do protest at $(\underline{x}, \underline{\mu})$.

Equilibrium institutions To choose D_t at step 2, the incumbent compares $V(x_t, \mu_t)$ with $U(x_t)$, given realized x_t, and the share of concerned citizens μ_t. Define value $\hat{x}(\mu)$ that makes the incumbent indifferent between the two: $V(\hat{x}(\mu), \mu) = U(\hat{x}(\mu))$. Then, the choice of democracy D_t satisfies:[9]

Proposition 1 *Under Assumption 1, there are two values* $\mu^L < \mu^H$, *such that for*

1. $\mu \leq \mu^L$, $D(\mu, x) = 0$ *for all* $x \in [\underline{x}, \overline{x}]$;
2. $\mu \geq \mu^H$, $D(\mu, x) = 1$ *for all* $x \in [\underline{x}, \overline{x}]$ *and*
3. $\mu \in [\mu^L, \mu^H]$ *there exists* $\widehat{x}(\mu) \in [\underline{x}, \overline{x}]$ *such that* $D(\mu, x) = 0$ *iff* $x \geq \widehat{x}(\mu)$.

The result is intuitive. With weak democratic values (low μ), protesters are unlikely to win and the incumbent leader can safely choose autocracy $D_t = 0$ and spend little on fighting. When democratic values are strong, incumbent loss is instead likely, and as fighting is costly citizens get democracy. These polar cases holds independently of x_t. However, for or the sign intermediate democratic values, institutions depend on the realization of x_t – at high (low) x, the leader stays with autocracy (installs democracy).

[9] We prove this proposition in the Web Appendix.

Evolving values Evolving democratic values reflect the relative fitness of being concerned vs. passive, as determined by expected utilities at date $t+1$ (or t). As the material payoffs of passive and concerned citizens are the same, they cancel out. Hence, only (2), the society-wide component of utility for concerned citizens matters. This leads to the following cultural dynamics.

From (4) $\mu_{t+1} - \mu_t$ is positive (negative) when $\Delta(\mu_t)$ is positive (negative). Using (2) and Proposition 1, and recalling that x has c.d.f. H, we can write the expression for $\Delta(\mu_t)$ as:

$$\Delta(\mu) = \begin{cases} \int_{\underline{x}}^{\overline{x}} \gamma(x)\, dH(x) & \mu \geq \mu^H \\ \int_{\underline{x}}^{\widehat{x}(\mu)} \gamma(x)\, dH(x) - \int_{\widehat{x}(\mu)}^{\overline{x}} L(x, \lambda(x,\mu))\, dH(x) & \mu \in \left[\mu^L, \mu^H\right] \\ -\int_{\underline{x}}^{\overline{x}} L(x,\mu)\, dH(x) & \mu \leq \mu^L \end{cases} \tag{5}$$

where $L(x,\lambda) = [\chi - \rho(1-\lambda)(1+\chi)]\gamma(x) + \rho\underline{c}$ is the loss from $D_t = 0$, which is increasing in λ. We focus on the case where $L(x,\lambda) > 0$ for all x, λ, which always holds with sufficient loss aversion χ.

There are three regions for μ. When $\mu \geq \mu^H$, democratic values have reached a point where incumbents always choose democracy $D_t = 1$ and no protests occur. The concerned have an intrinsic gain from this institution, so their share is growing. As $\mu \leq \mu^L$, the incumbent group get its preferred autocracy $D_t = 0$ for any realization of x and the few concerned individuals feel a perpetual sense of injustice, which gives them an intrinsic loss. Hence, democratic values are shrinking. In an intermediate range for democratic values, realized x determines the incumbent's institutional choice. From Proposition 1 and (2), a gain only occurs if $D_t = 1$ which requires $x \leq \widehat{x}(\mu)$. Otherwise, incumbents choose $D_t = 0$, which leads to losses as defined in (2). Democratic values grow (shrink) when expected gains exceed (fall below) expected losses, which in turn requires expected x to fall below (above) threshold $\widehat{x}(\mu)$,according to distribution H. As we show in the Web Appendix, $\partial\widehat{x}(\mu)/\partial\mu > 0$, which implies $\Delta_\mu(\mu) \geq 0$ for all $\mu \in [0,1]$.

From (2), the loss from being a concerned citizen is higher when x is high and the probability of a protest unseating the incumbent is low, which is the case when μ is low, since the survival function $\lambda(x,\mu)$ is then close to one. At the other extreme, the loss is low when the incumbent almost surely loses a rebellion, as $\lambda(x,\mu)$ is close to zero.

Steady states and inertia The possible steady states are described as follows:

Proposition 2 *There exists a critical value $\hat{\mu}$ defined by*

$$\int_{\underline{x}}^{\widehat{x}(\hat{\mu})} \gamma(x)\, dH(x) = \int_{\widehat{x}(\hat{\mu})}^{\overline{x}} L(x, \lambda(x,\hat{\mu}))\, dH(x).$$

Whenever $\mu_0 \geq \hat{\mu}$, the polity converges to $\mu = 1$. However, for $\mu < \hat{\mu}$, the polity converges to $\mu = 0$.

To see why this is true, note that $\Delta(0) < 0$ and $\Delta(1) > 0$. Because $\Delta(\mu)$ is (weakly) monotonically increasing, there must exist a unique level $\hat{\mu}$ such that $\Delta(\hat{\mu}) = 0$. Moreover, this interior point is unstable, meaning that the dynamics described in (4) will converge slowly to either of two extremes (see the Web Appendix for further discussion).

This convergence is associated with a specific path of democratic institutions. Once democratic values on an upward path reach region $\mu \geq \mu^H$, democracy becomes permanently chosen. Equally, once democratic values on a downward path reach the region where $\mu \leq \mu^L$, autocracy becomes permanent. The intermediate region for μ can have reforms in both directions depending on x_t.

To summarize, democratic institutions are persistent without assuming any form of institutional commitments. Institutional inertia reflects slow-moving democratic values which feed back to democratic reform. Democratic institutions also feed back to democratic values.

4 Insights

The model is consistent with the two motivating facts in Subsection 2.2. Its predictions encompass a range of findings discussed in existing research. Moreover, beyond reproducing the two motivating facts, the model makes some auxiliary predictions on democratic values that we may confront with data.

4.1 Motivating Facts Redux

Institutions Table 1 documented three groups of country histories: permanent transitions into democracy, into autocracy, and flip-flopping between the two. These correspond neatly to the predictions from Propositions 1 and 2, namely an upper and lower region for democratic values where democracy and autocracy become absorbing states, and an intermediate range where reforms occur in both directions due to country-specific shocks. The model says that institutional responses to *temporary* shocks to x are heterogeneous, depending on the value of μ. This, together with separate starting values μ_0, implies that countries follow their own paths which reflect an evolving state variable rather than multiple equilibria.

Values Figure 1 documented that people in societies that have never or rarely transitioned into democratic institutions value democracy less than people in long-consolidated democracies. Our model underpins this fact: (4) and (5), together with the complementarity between $D = 1$ and μ, imply that we should observe a larger share of citizens with high democratic values – a higher μ – today, the more time in the past their society had positive and high values of Δ. This, in turn, is associated with more time spent with democratic institutions.

4.2 Relationship to Existing Ideas

Persistence Our model suggests a mechanism behind a long-lived effect of historical political institutions, like the colonial-origins hypothesis of Acemoglu, Johnson, and Robinson (2001). However, it also suggests why cumulated values – like social or democratic capital – may consolidate change, as in Putnam (1993) and Persson and Tabellini (2009). Even though incumbents are free to reform in any period, political institutions become sticky in equilibrium due to slow-moving democratic values.

Varieties of reform The model allows different types of political reforms: "defensive", when ruling elites voluntarily relinquish political control (Acemoglu and Robinson 2000, 2006), and "offensive", when citizens force ruling elites to implement institutional change (Marx and Engels 1848, Kuran 1995).

Critical junctures Except shedding light on the effect of temporary shocks, x_t, and conflicts of interest between ruling elites and opposition groups, the model also shows how *permanent* shocks might matter. Specifically, it underpins how *critical junctures* may shape long-run outcomes, as stressed by Acemoglu and Robinson (2012). Two otherwise similar countries with democratic values just above and below $\hat{\mu}$,the country-specific threshold for the dynamics, can have radically different

trajectories. Moreover, a permanent shock to the *distribution* of x around $\hat{\mu}$, can flip a country to the opposite side of $\hat{\mu}$. Propositions 1 and 2 suggest that such a shift could have long-run consequences for democratic values and institutions. For example, relying on the interpretation of x_t as resource rents, resource discoveries could affects the trajectory of democratic values. This merits further investigation, especially since WVS data show a negative correlation between support for democracy and contemporaneous natural-resource intensity.

Initial conditions The model also highlights the importance of historical processes that change μ or function $Q(\Delta)$. One example is the transformation of political views when the ideas of Locke (1690), Montesquieu (1748), and Paine (1776) influenced the US Founding Fathers, and challenged ruling elites elsewhere. Christian teaching and institutions may also have changed exposure to liberal thought. Our model predicts that once the democratic genie is out of the bottle and μ exceeds $\hat{\mu}$, democratic reform will be sustained.

Reversing this logic, democratic institutions installed before democratic values are built may be hard to sustain. Some post-colonial African states – Nigeria, Sudan, Somalia, and Uganda – began with European-style democratic (parliamentary) regimes, but these broke down within a decade. This could be because lacking democratic values made it hard to support defense of democracy.

Economic growth The model suggests how economic development may sustain democracy. As development raises wages w, the opportunity cost of fighting rises, making it less likely that incumbents will resist democratic rights. If the costs of protests also rise with economic growth, however, this pulls in the opposite direction. But the complementarity at the heart of the model also suggests a coevolution of democratic values and the economy, capturing the predictions of modernization theorists such as Lipset (1959)

Autocracy traps Our model suggests how weak democratic values may create an "autocracy trap". Russia's short democratic history (in PIV) and low democratic values (in WVS) is a case in point. Previous Soviet repression (high f) weakened democratic values and thus undermined later reforms attempts, like that by Boris Yeltsin (upon a low c) – giving democracy little chance of becoming permanent. Changing Russia's trajectory would require different fundamentals or a favorable shock to values μ. Examples could be a weaker repression capacity (raising the influence of given democratic values) or lower resource rents x (cutting the additional rents to power from autocracy).

Democratic capital Section 2 showed democratic support to be strongest in countries that made once-and-for-all democratic transitions. Persson and Tabellini (2009) interpreted institutional persistence in terms of "democratic capital". This is a classic case where state dependence and unobserved heterogeneity provide competing interpretations. Our model suggests that democratic capital may reflect an unobserved omitted variable – democratic values – rather than state dependence, i.e. past experience with democracy directly causing future democracy. Moreover, our model suggests that causality runs both ways.

4.3 Auxiliary Predictions and Data

The model makes some auxiliary predictions about values.

Foreign occupation World history is replete of examples, such as colonization or Soviet occupation of Eastern Europe, where foreign powers dictate domestic political institutions. Our framework can interpret these as foreign imposition of institutions $D_t = 0$ via repressive use of force f_t.

Such historical episodes should have *persistent* effects via evolving democratic values. The dynamic complementarity between institutions and values implies that a country where a democratic regime is interrupted by a foreign-imposed autocracy may have weaker democratic values in *future* periods.

What if foreign occupation simply replaces an existing domestic autocracy? Under the plausible assumption that a major power is more likely to enforce autocracy through repression compared to a domestic autocrat, such an occupied country will have lower future democratic values compared to spending the same amount of time in homegrown autocracy. To see why, let $\Lambda(x, \mu)$ be the probability that autocracy persists under foreign occupation and, as before, $\lambda(x, \mu)$ the same probability under domestic autocracy. If $\Lambda(x, \mu) > \lambda(x, \mu)$, (4) and (5) imply that today's μ must be lower in an occupied country, *ceteris paribus*, for the same number of years spent in autocracy.[10]

Colonialism Colonial powers mostly established autocratic regimes, though some colonies – e.g., Australia, Canada, New Zealand and South Africa – got elements of democracy. Acemoglu, Johnson and Robinson (2001) distinguish extractive and inclusive institutions, which we could capture as different values of D. The empirical findings in Acemoglu, Johnson and Robinson (2001) are then readily interpretable in our model. Maintaining $D_t = 0$ ($D_t = 1$), colonialism may have permanently affected post-colonial democratic institutions by inhibiting (promoting) emerging democratic values.[11] Countries with repressed values would then face long-run effects of colonialism, beyond any initial efforts to bring in democratic reforms.

To shed light on this prediction, we exploit within-country cross-cohort variation. Taken literally, the model's generational structure translates the predicted variation in democratic values across time into variation across cohorts. Empirically, this requires that democratic values are formed relatively early and become sticky over an individual's lifetime. Then, the model predicts individuals with their formative years under colonization to have lower democratic values than those growing up post independence. We check this against WVS data in post-colonial countries, comparing individuals who had, or had not, turned 16 (results are similar for other cutoffs) by the country-specific independence year. Thus we follow a similar approach as earlier studies of age-dependent political preferences (Alesina and Fuchs-Schündeln 2007, Kaplan and Mukand 2014).

Specifically, we estimate the following linear probability model:

$$v_{b,c,w} = \alpha_b + \alpha_c + \alpha_w + \delta_{b,c,w} + \gamma x_{b,c,w} + \varepsilon_{b,c,w}, \tag{6}$$

where $v_{b,c,w}$ is a dummy variable for democratic support in the WVS (as in Section 2.2), for an individual born in year b in country c answering the question in survey wave w. We include a full set of birth-year, wave and country dummies $\{\alpha_b, \alpha_w, \alpha_c\}$, as well as a set of individual controls $x_{b,c,w}$ as detailed in the note to Table 2 (results are similar with 10-year cohort dummies replacing birth-year dummies). The individual treatment variable $\delta_{b,c,w}$ is a binary indicator set equal to one if the individual was aged sixteen or older at the end of colonialism.

Table 2 column (1) shows that a smaller share of cohorts with early-life exposure to colonialism holds strong democratic values. The cross-cohort difference is about 10% of the overall (world) sample mean. Moreover, column (2) shows that the result holds up when we estimate the same regression

[10] This follows since loss function $L(x, \cdot)$ is increasing.

[11] Acemoglu, Johnson and Robinson use strong executive constraints – a component of the *polity2* democracy index – as a dependent variable in the post-colonial era.

on the sub-sample of countries that were ever colonies. This adds further credibility to the idea that democratic values reflect past political regimes as posited by the theoretical model.

Communism We can apply a similar logic to cold-war occupation, when the USSR absorbed some independent countries – such as the Baltic ones – and made others satellites. Among countries with WVS data, we code 16 (see the Table 2 note) as subject to Soviet occupation. The proportion of the population who nowadays strongly support democracy in these countries is 0.54, vs. 0.61 in non-USSR influenced countries.

Column (3) estimates a version of (6) where the treatment indicator $\delta_{b,c,w}$ is now set at one for those who turn 16 before the end of USSR occupation, which we set at 1990 in all countries. Like in columns (1) and (2), we thus only exploit within-country cross-cohort variation in values. We find a negative and significant correlation between democratic values and formative years spent under Soviet influence – the same effect as for colonialism both qualitatively and quantitatively. This result echoes the finding of Neundorf (2010) how within-country intergenerational preferences for democracy in ten East-European countries depend on Soviet influence. Column (4) estimates this on the sub-sample of countries, which were ever subject to Soviet influence. Although the point estimate is the same as in column (3) the loss of power in a much smaller sample means that the coefficient is no longer statistically significant.

5 Conclusion

We model the two-way interaction between democratic values and institutions with a single state variable: the proportion of citizens holding strong enough values to defend democracy. Rejoice or despair about political institutions among these citizens help propagate democratic values via a dynamic complementarity. Institutional change becomes a gradual process, not because incumbents can commit future incumbents, but because these pay close attention to gradually evolving democratic values. Shocks along this path create the kinds of episodic change seen in the data.

Our model bridges the cultural and strategic approaches to institutional change: democratic values and democratic reforms reinforce each other. These joint dynamics help us better understand persistence and change in political institutions across countries and time. The model can cast light on the heterogeneity in country experience with democratic reform – it also allows us to be precise about critical junctures and the role of initial conditions. Finally, we present some within-country correlations consistent with the auxiliary predictions we get when applying the model to the effect of foreign occupation on domestic democratic values.

The paper suggests a wider agenda. On the empirical side, our model has a number of implications, which could be explored beyond simple correlations. On the theoretical side, little research has been devoted to the co-determination of values and institutional rules. Models like ours can be deployed to study related phenomena, such as the joint dynamics of organizational cultures and organizational designs (Besley and Persson 2018).

References

[1] Acemoglu, Daron, Simon Johnson, and James Robinson [2001], "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review* 91, 1369-1401.

[2] Acemoglu, Daron and James Robinson [2000], "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective, " *Quarterly Journal of Economics* 115, 1167-1199.

[3] Acemoglu Daron and James Robinson [2006], *Economic Origins of Dictatorship and Democracy*, Cambridge, MA: Cambridge University Press.

[4] Acemoglu Daron and James Robinson [2012], *Why Nations Fail*, New York, NY: Crown Publishers.

[5] Aidt Toke and Raphaël Franck [2015], "Democratization Under the Threat of Revolution: Evidence From the Great Reform Act of 1832," *Econometrica* 83, 505-547.

[6] Aidt, Toke and Peter Jensen [2014], "Workers of the World, Unite! Franchise Extensions and the Threat of Revolution in Europe, 1820-1938," *European Economic Review* 72, 52-75.

[7] Alesina, Alberto and Nicola Fuchs-Schündeln [2007], "Goodbye Lenin (or Not?): The Effect of Communism on People," *American Economic Review* 97, 1507–1528.

[8] Almond, Gabriel and Sidney Verba [1963], *The Civic Culture: Political Attitudes and Democracy in Five Nations*, New York: Sage Publications.

[9] Bandiera, Oriana, Myra Mohnen, Imran Rasul, and Martina Viarengo [2016], "Nation-Building through Compulsory Schooling During the Age of Mass Migration," mimeo, LSE.

[10] Besley, Timothy and Torsten Persson [2009], "The Origins of State Capacity: Property Rights, Taxation and Politics", *American Economic Review* 99, 1218-1244.

[11] Besley Timothy and Torsten Persson [2011], *Pillars of Prosperity: The Political Economics of Development Clusters*, Princeton, NJ: Princeton University Press.

[12] Besley, Timothy and Torsten Persson [2018], "Organizational Dynamics: Culture, Design and Performance", Mimeo, LSE and IIES.

[13] Bisin, Alberto and Thierry Verdier [2001], "The Economics of Cultural Transmission and the Dynamics of Preferences," *Journal of Economic Theory* 97, 298–319.

[14] Bisin, Alberto and Thierry Verdier [2011], "The Economics of Cultural Transmission and Socialization," Chapter 9 in Benhabib, Jess (ed.), *Handbook of Social Economics*, Volume 1A, Elsevier.

[15] Bisin, Alberto and Thierry Verdier [2017], "On the Joint Evolution of Culture and Institutions," NBER Working Paper No. 23375.

[16] Boyd, Robert and Peter Richerson [1985], *Culture and the Evolutionary Process*, Chicago, IL: University of Chicago Press.

[17] Cavalli-Sforza, Luigi L. and Marcus Feldman [1981], *Cultural Transmission and Evolution*, Princeton, NJ: Princeton University Press.

[18] Fuchs-Schündeln, Nicola and Matthias Schündeln [2015], "On the Endogeneity of Political Preferences: Evidence from Individual Experience with Democracy," *Science* 347, 1145-1148.

[19] Gorodnichenko, Yuriy and Gerard Roland [2015], "Culture, Institutions, and Democratization," Mimeo, UC Berkeley.

[20] Güth, Werner and Menahem E. Yaari [1992], "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in Witt, Ulrich (ed.) *Explaining Process and Change: Approaches to Evolutionary Economics*, Ann Arbor, MI: University of Michigan Press.

[21] Inglehart, Ronald [1997], *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*, Princeton, NJ: Princeton University Press.

[22] Inglehart, Ronald and Christian Welzel [2005], *Modernization, Cultural Change, and Democracy: The Human Development Sequence*, Cambridge, UK: Cambridge University Press.

[23] Kahneman, Daniel and Amos Tversky [1979], "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263–291.

[24] Kaplan, Ethan and Shurun Mukand [2014], "The Persistence of Political Partisanship: Evidence from 9/11", Mimeo, University of Maryland.

[25] Kuran, Timur [1995], *Private Truths, Public Lies: The Social Consequences of Preference Falsification,* Cambridge, MA: Harvard University Press.

[26] Lagunoff, Roger [2001], "A Theory of Constitutional Standards and Civil Liberty," *Review of Economic Studies* 68, 109-132.

[27] Lipset, Seymour Martin [1959], "Some Social Requisites of Democracy: Economic Development and Political Legitimacy," *American Political Science Review* 53, 69-105.

[28] Locke, John [1988, 1690], *Two Treatises on Governmen*t, Cambridge Texts in the History of Political Thought.

[29] Loomes, Graham and Robert Sugden [1982], "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *Economic Journal* 92, 805-824

[30] Marshall, Monty G. and Keith Jaggers [2005], "POLITY IV Project: Political Regime Characteristics and Transitions, 1800-2004," <www.cidcm.umd.edu/polity>.

[31] Marx, Karl and Friedrich Engels [2004, 1848], *Manifesto of the Communist Party*, Marxist Internet Archive, <www.marxists.org/archive/marx/works/1848/communist-manifesto/index.htm>.

[32] Montesquieu, Charles de [1989, 1748], *The Spirit of the Laws*, Cambridge, UK: Cambridge University Press.

[33] Neundorf, Anja [2010], "Democracy in Transition: A Micro Perspective on System Change in Post-Socialist Societies," *Journal of Politics* 72, 1096-1108.

[34] Passarelli, Francesco and Guido Tabellini [2017], "Emotions and Political Unrest," *Journal of Political Economy* 125, 903-946.

[35] Persson, Torsten and Guido Tabellini [2008], "Political Regimes and Economic Growth," in Helpman, Elhanan (ed.), *Institutions and Economic Performance*, Cambridge, MA: Harvard University Press.

[36] Persson, Torsten and Guido Tabellini [2009], "Democratic Capital: The Nexus of Economic and Political Change," *American Economic Journal Macroeconomics* 1, 88-126.

[37] Putnam, Robert [1993], *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton, NJ: Princeton University Press.

[38] Sandholm, William [2010], *Population Games and Evolutionary Dynamics*, Cambridge, MA: MIT Press.

[39] Siedentop, Larry [2014], *Inventing the Individual: The Origins of Western Liberalism*, Cambridge, MA: Harvard University Press.

[40] Ticchi, Davide, Thierry Verdier, and Andrea Vindigni [2013], "Democracy, Dictatorship and the Cultural Transmission of Political Values," IZA Discussion Paper, No. 7441.

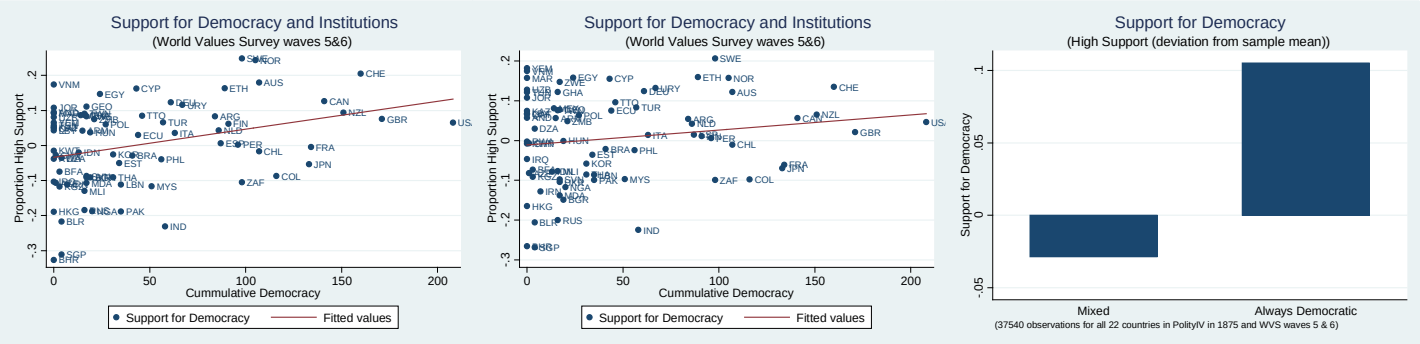
[41] Thaler, Richard, H. [1999], "Mental Accounting Matters", *Journal of Behavioral Decision Making* 12, 183-206.

[42] Weber, Max, [1968, 1922], *Economy and Society*, Berkeley, CA: University of California Press

[43] Weingast, Barry [1997], "The Political Foundations of Democracy and the Rule of Law," *American Political Science Review* 91, 245-263.

[44] Welzel, Christian [2007], "Are Levels of Democracy Affected by Mass Attitudes? Testing Attainment and Sustainment Effects on Democracy," *International Political Science Review* 28, 397-424.

Figure 1: Democratic Values and Democratic History



Notes: The data on institutions come from the Polity IV website http://www.systemicpeace.org/polityproject.html . For democracy, we use the variable *polity2* (on a -10,+10 scale) to create a dummy variable which is equal to 1 if *polity2* takes a positive value in a given country-year. The horizontal axes in the left and middle panel display the number of years for which a country has had a 1 for this democracy dummy. Support for democracy is an individual dummy variable from the World Values Survey http://www.worldvaluessurvey.org/wvs.jsp waves 5 and 6 which equals 1 if the individual expresses Support for Democracy (on a 10 point Scale) at 9 or 10. The vertical axis gives the average value of the dummy variable for each country across both waves. The left panel plots the raw data. The middle panel holds constant each individual's gender, education, age and income: we estimate an individual-level linear probability model with the dummy for democratic support on the left-hand side including on the right-hand side controls for gender, ten dummies for income groups, three for education groups, and three age bands. To construct the figure, we average the residuals at the country level. The right panel compares the values in countries (in the top right panel of Table 1 along with Sweden, the UK and Uruguay) that have one long-standing transition into democracy with those with a recent, multiple or no transition into democracy (in the left and middle panels of Table 1 along with Hungary, Italy, Mexico and Russia).

Table 1: Classification of Countries by Democratic History

<table>
<tr><th>Weak</th><th colspan="2">Mixed</th><th>Strong</th></tr>
<tr><td>Always Non-democratic</td><td colspan="2">Multiple Changes
(Number Upward, Number Downward)</td><td>Always Democratic</td></tr>
<tr><td>Afghanistan
Morocco*†
Oman</td><td rowspan="3">Argentina*† (7,6)
Austria (3,2)
Belgium (3,2)
Bolivia (2,1)
Brazil*† (2,1)
Chile*† (3,2)
China*† (1,1)
Colombia*† (3,2)
Denmark (3,2)
Dominican Republic (2,1)
Ecuador† (3,2)
Ethiopia*† (1,1)
France* (3,2)
Germany*† (2,1)
Greece (5,4)
Guatemala (6,5)</td><td rowspan="3">Haiti (4,4)
Honduras (3,2)
Iran*† (1,1)
Japan*† (2,1)
Liberia (1,1)
Nepal (3,2)
Netherlands*† (2,1)
Norway* (2,1)
Peru*† (8,7)
Portugal (3,2)
Paraguay (2,1)
Serbia* (4,3)
Spain*† (4,3)
Thailand*† (5,4)
Turkey*† (3,2)
Venezuela (1,1)</td><td>Canada*†
New Zealand†
Switzerland*
United States*†</td></tr>
<tr><td>Permanent Switch to Non-democracy (Year of Switch)</td><td>Permanent Switch to Democracy (Year of Switch)</td></tr>
<tr><td></td><td>Costa Rica (1841)
El Salvador (1982)
Hungary* (1989)
Italy* (1945)
Mexico*† (1994)
Nicaragua (1990)
Romania (1990)
Russia*† (1992)
Sweden*† (1910)
United Kingdom*† (1837)
Uruguay* (1910)</td></tr>
</table>

Notes: Sample is 50 countries which appear in the PolityIV data base as independent countries in 1875. The data set covers the period 1800 to 2011 and Table 3 displays when each country first entered the data. Data for Germany are for unified Germany; West Germany had strong executive constraints from 1950 onwards. A * denotes a country in wave 5 and a † denotes a country in the wave 6 of World Values Survey.

Table 2: External Influence on Individual Democratic Values

	(1)	(2)	(3)	(4)
Colonial rule at 16	- 0.062***	- 0.058***		
	(0.015)	(0.016)		
USSR occupation at 16			- 0.069***	- 0.067
			(0.018)	(0.088)
Individual controls	Yes	Yes	Yes	Yes
Birth-year dummies	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes
Countries in sample	All	Past Colonies	All	Post USSR Block
Number of Countries				
Observations	140311	103776	140311	25952
R^2	0.075	0.074	0.075	0.056

Notes: All the estimates in Table 2 come from individual-level, linear-probability models, where the left-hand side variable is our dummy for a score of 9 or 10 of democratic support. We control for a wave dummy, country dummies, dummies for birth year, gender, ten dummies for income, three dummies for education, and three age bands. For the end of colonialism, we use the following list of countries from the WVS with their dates of decolonization in parentheses: Algeria (1963), Argentina (1853), Australia (1901), Bahrain (1971), Brazil (1822), Burkina Faso (1960), Canada (1867), Chile (1818), Colombia (1810), Cyprus (1960), Ecuador (1822), Egypt (1922), Finland (1917), Ghana (1957), India (1947), Indonesia (1949), Iraq (1932), Jordan (1946), South Korea (1948), Kuwait (1962), Lebanon (1941), Libya (1951), Malaysia (1957), Mali (1958), Mexico (1810), Morocco (1955), New Zealand (1907), Nigeria (1960), Norway (1905), Pakistan (1947), Peru (1821), Philippines (1898), Qatar (1971), Rwanda (1962), Singapore (1965), South Africa (1910), Taiwan (1949), Trinidad and Tobago (1962), Tunisia (1956), Uruguay (1825), USA (1776), Vietnam (1945), Yemen (1967), Zambia (1964), Zimbabwe (1980). For Soviet influence, we use data for Armenia, Azerbaijan, Bulgaria, Belarus, Estonia, Georgia, Hungary, Kazakhstan, Kyrgyzstan, Moldova, Poland, Romania, Russia, Slovenia, Serbia, Ukraine and Uzbekistan. In columns (1) and (2), "Colonial rule at 16" is a dummy variable equal to one if the individual was aged 16 or older in the year her country gained independence. In columns (3) and (4), "USSR occupation at 16" is a dummy variable equal to one if the individual was 16 or older when Soviet occupation ended, which we set to 1990 for all countries. Standard errors are adjusted for clustering at the country level. A "*" denotes significant at 10%, a "**" significant at 5% and "***" significant at 1%.

Web Appendix
Democratic Values and Institutions
Tim Besley and Torsten Persson

A Micro-Foundations for Political Institutions

We begin by discussing two examples that outline possible microfoundations for interpreting our framework in Section 3 of the text as a model of democracy. Each example focuses on one of the two main aspects of democratic institutions, namely open and free elections of the executive, on the one hand, and constraints on the executive (once in power), on the other. Both examples are highly stylized and can be considerably generalized.

Checks and balances The first example more nearly captures *constraints on the executive.* Here, we imagine that (a representative of) the incumbent group has proposal power over how to split some (resource) rents x_t across the two groups. This proposal will always allocate all the rents to the incumbent group. Under autocracy – i.e., with $D_t = 0$ – this proposal just goes through and we have

$$u^I(x_t, 0) = x_t \text{ and } u^O(x_t, 0) = 0.$$

Under democracy, $D_t = 1$, then instead with some exogenously given probability $2q < 1$, the opposition group can reject the proposal and impose an equal split of the rents with $x_t/2$ to each group. The expected rent allocation is thus

$$u^I(x_t, 1) = (1-q)\, x_t \text{ and } u^O(x_t, 1) = qx_t.$$

Altogether, we have

$$\Gamma(x) = qx = \gamma(x).$$

Open elections The second example more nearly captures *open recruitment of the executive.* Under autocracy, $D_t = 0$, a representative of the incumbent group faces no challenge for power (but there may still be costly protests) and safely remains to the next period. But under democracy, $D_t = 1$, this representative runs against a representative of the opposition group in a stochastic electoral contest. The incumbent candidate wins this contest with probability $1 - x_t$. Thus $x_t \in [0, 1]$ is the (relative) unpopularity of the incumbent leader. We normalize the value of winning (which captures some unmodeled policy advantage) to 1. With $D_t = 0$, we have

$$u^I(x_t, 0) = 1 \text{ and } u^O(x_t, 0) = 0.$$

With $D_t = 1$,we instead have

$$u^I(x_t, 1) = 1 - x_t \text{ and } u^O(x_t, 1) = x_t.$$

Altogether, we have

$$\Gamma(x) = x = \gamma(x).$$

B Democratic Values

In this section, we discuss a possible microfoundation for the democratic values that appear in (2) of Section 3 in the text.

The expression in (2) assigns an additional positive payoff if $D_t = 1$ and a negative one if $D_t = 0$. It also assumes that democratic values are universal rather than particularistic. That is, concerned citizens care about society-wide gains and losses from democratic rights, and not only those which accrue to other concerned citizens. Assuming the latter would be an alternative way to formulate the model and would tend to strengthen the main results.

The formulation in (2) can be derived from a reference-dependent social preference, with one reference point for gains r_g and one for losses r_l

$$S\left(r_g, r_l, D, x\right) = \chi \min\left\{u^O\left(x, D\right) - u^O\left(x, r_l\right), 0\right\} + \max\left\{u^O\left(x, D\right) - u^O\left(x, r_g\right), 0\right\}. \tag{B.1}$$

We set $r_g = 0$ and $r_l = 1$ so gains are measured relative to the worst institution and losses relative to the best – i.e., concerned citizens evaluate social affairs against an institutional benchmark. The idea of reference-dependent preferences is well-established, following Kahneman and Tversky (1979) and a range of psychological studies. Specifically, our formulation follows Loomes and Sugden (1982), where an individual experiences either regret or rejoice depending on her reference point for an outcome.

Applications of reference-dependent preferences to concrete phenomena are discussed, e.g., in Kahneman et al (1991). (Koszegi and Rabin 2006 give a more recent theoretical treatment of reference-dependent preferences.) The payoffs in (B.1) can be thought as reflecting a feeling of (in)justice among citizens, based on *societal* gains/losses relative to the outcomes under the alternative institution, which embody their views about the right kind of society. Democratic values are thus distinct from standard preferences, analogous to the distinction between acquisition utility and transactions utility, which can also reflect a sense of justice (Thaler 1999).

C Socialization

In this section, we show three possible microfoundations for the evolutionary model stated in Section 3 of the main text.

Basic socialization model We first consider a model with successive generations, which overlap only in so far as parents endow their children with values, as modeled in Besley (2015). Children have two parents and – to keep the population balanced – all pairs have two children. We also assume that all marriage matching is random.[12]

Children are socialized into having democratic values. For simplicity, we model socialization as resulting from a form of osmosis rather than strategic behavior by parents.[13] Two parents of the same type simply pass along the values associated with their common type. However, children whose parents have different types get their type depending on the expected utilities of being concerned with democratic values rather than passive. Let $\Delta(\mu)$ be the expected utility difference between these types – their relative fitness – when the proportion concerned in the population is μ. Moreover, let $\eta \in (-\infty, \infty)$ be a couple-specific idiosyncratic negative shock to this utility difference. Then, a child with mixed parentage becomes concerned with democratic values, if and only if $\eta \leq \Delta$.

[12] For the results to go through, we only require that there is at least some element of random matching. With full assortative matching, there would be no socialization as all offspring would have parents of the same type.

[13] This makes the model simpler and does not fundamentally affect the insights compared to the strategic socialization model of Bisin and Verdier (2001). .

We assume that η has a symmetric single-peaked distribution with c.d.f. K and p.d.f. k. This implies that a mixed-marriage child holds democratic values with probability $K(\Delta(\mu))$ at utility difference $\Delta(\mu)$. By the law of large numbers, this is also the proportion among those with mixed parentage. By definition, $K(\cdot)$ is monotonically increasing, and by symmetry $K(0) = 1/2$.

The evolution of democratic values becomes :

$$\mu_{t+1} = \mu_t + 2\mu_t(1-\mu_t)\left[K\left(\Delta\left(\mu_{t+1}\right)\right) - 1/2\right]. \tag{C.2}$$

This corresponds to (4) with $\varsigma^{P,C} = \mu\left[K\left(\Delta\left(\mu_{t+1}\right)\right) - 1/2\right]$ and $\varsigma^{C,P} = -(1-\mu)\left[K\left(\Delta\left(\mu_{t+1}\right)\right) - 1/2\right]$.

Strategic socialization We now show that the key equation (4) can be derived from a model, in which matching is assortative and socialization is purposeful. This follows the approach of Cavalli-Sforza and Feldman (1981) as adapted by Bisin and Verdier (2001). Socialization would then have two parts:

1. Direct Socialization: A parent may directly socialize a child into being a concerned citizen, depending on parental effort.

2. Oblique Socialization: If this is unsuccessful, the child may become socialized by society at large becoming a concerned citizen with probability μ_t.

We focus on the case where marriages are perfectly assortative and each pair of parents has two kids. Let $e \in \{0,1\}$ be the effort put into socializing kids as concerned at cost C. Also, let the probability of successful socialization depend on $e + \varphi$ where φ is a stochastic socialization shock uniformly distributed on $\left[-\frac{1}{L}, \frac{1}{L}\right]$. Then, we have:

$$\text{Prob[concerned: } e] = \frac{1}{2} + Le.$$

Finally, as in our canonical model, let η be an idiosyncratic shock to parental preferences. They now choose socialization effort:

$$e^* = \arg\max\left\{\left(\frac{1}{2} + Le\right)[\Delta(\mu) + \eta] - Ce\right\}.$$

This defines a threshold

$$\hat{\eta} = \nu - \Delta(\mu),$$

where $\nu = C/L$ such that $e^* = 1$ if and only $\eta \geq \hat{\eta}$.

For the children of concerned parents, the probability of a child being socialized as concerned is $K\left(\Delta\left(\mu_{t+1}\right) - \nu\right)$. For those who are not directly socialized, the probability of oblique socialization into being concerned is $\left(1 - K\left(\Delta\left(\mu_{t+1}\right) - \nu\right)\right)\mu_t$.

Adding these expressions, the overall probability that the kids of concerned parents are concerned is:

$$K\left(\Delta\left(\mu_{t+1}\right) - \nu\right) + \left(1 - K\left(\Delta\left(\mu_{t+1}\right) - \nu\right)\right)\mu_t. \tag{C.3}$$

If a child is born to passive parents, we assume she is never directly socialized into being concerned. However, she is socialized as passive with probability $\left(1 - K\left(\Delta\left(\mu_{t+1}\right) - \nu\right)\right)$. The fraction of such children who are obliquely socialized as concerned is therefore:

$$K\left(\Delta\left(\mu_{t+1}\right) - \nu\right)\mu_t. \tag{C.4}$$

The overall fraction of concerned citizens in the next generation becomes

$$\begin{aligned}\mu_{t+1} &= \mu_t\left[K\left(\Delta\left(\mu_{t+1}\right)-\nu\right)+\left(1-K\left(\Delta\left(\mu_{t+1}\right)-\nu\right)\right)\mu_t\right]+\left(1-\mu_t\right)\left[K\left(\Delta\left(\mu_{t+1}\right)-\nu\right)\mu_t\right] \\ &= \left(\mu_t\right)^2+2\left(1-\mu_t\right)\mu_t K\left(\Delta\left(\mu_{t+1}\right)-\nu\right),\end{aligned}$$

which corresponds to (4) with $\varsigma^{P,C}=\mu_t\left[K\left(\Delta\left(\mu_{t+1}\right)-\nu\right)-\frac{1}{2}\right]$ and $\varsigma^{C,P}=-\left(1-\mu_t\right)\left[K\left(\Delta\left(\mu_{t+1}\right)-\nu\right)-\frac{1}{2}\right]$. The only difference is that costly effort of being socialized as passive reduces the probability of concerned citizens in the population relative to our basic model, which has $\nu=0$. This is the special case when $C=0$ – i.e., when the effort by parents into socializing their child is costless.

In this setting, the candidate for an interior steady state is:

$$\Delta\left(\hat{\mu}\right)=\nu,$$

but when $\Delta_\mu(\mu)\geq 0$ this is unstable and the basic thrust of the basic-model analysis goes through unscathed.

A replicator dynamic Suppose that concerned and passive citizens are two behavioral types in the population and that members of each young generation adopts their types to the relative success of the "cultural parents" they encounter. This kind of imitation will give rise to a standard replicator dynamics:

$$\begin{aligned}\mu_{t+1}-\mu_t &= \mu_t\frac{\left[\left(\text{Utility Concerned:}\mu_t\right)-\left(\text{Average Utility:}\mu_t\right)\right]}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}} \\ &= \mu_t\left(1-\mu_t\right)\frac{\left[\left(\text{Utility Concerned:}\mu_t\right)-\left(\text{Utility Passive:}\mu_t\right)\right]}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}},\end{aligned}$$

where we have chosen to normalize by the maximum utility gain from democratic institutions so that the relevant expressions is bounded in the unit interval. Let $\pi\left(x,\mu\right)$ be the probability that $D=1$ given $\{x,\mu\}$. This expression boils down to

$$\begin{aligned}\mu_{t+1}-\mu_t &= \mu_t\left(1-\mu_t\right)\frac{\int_{\underline{x}}^{\bar{x}}\left[\pi\left(x,\mu_t\right)\gamma\left(x\right)-\left(1-\pi\left(x,\mu_t\right)L\left(x,\lambda\left(x,\mu\right)\right)\right)\right]dH\left(x\right)}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}} \\ &= \mu_t\left(1-\mu_t\right)\frac{\Delta\left(\mu_t\right)}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}}.\end{aligned}$$

This is a special case of (4) if

$$\begin{aligned}\varsigma^{P,C} &= \frac{\mu_t\max\left\{\Delta,0\right\}}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}} \\ &\text{and} \\ \varsigma^{C,P} &= \frac{\left(1-\mu_t\right)\max\left\{-\Delta,0\right\}}{\left(1+\chi\right)\gamma\left(\bar{x}\right)+\rho\underline{c}}.\end{aligned}$$

Then the tipping point for the dynamics would be $\Delta\left(\hat{\mu}\right)=0$, which would be similar to our analysis. Moreover, as long as $\Delta_\mu(\mu)\geq 0$, the dynamics would be qualitatively the same as in the canonical model.

D Steps 2 and 3 and Proposition 1

In this section, we analyze the optimal fighting decisions by the incumbent and the opposition, define the equilibrium functions $V(x_t, \mu_t)$, $U(x_t)$ and $\lambda(x, \mu)$ mentioned in the text, analyze their properties, and prove Proposition 1.

Protests and payoffs – step 3 All citizens observe the level of fighting f chosen at step 2 and protest if the benefit exceeds the cost. Given (3), passive citizens never protest as their private benefit is always lower than the cost. Therefore, the only issue is whether concerned citizens find it worthwhile to protest, given the realization of c_t. To determine this, define a threshold $\hat{c}(\mu, f, x)$ from the condition

$$\mu p(f) \left[u^O(x, 1) - u^O(x, 0) + s(x_t, 1) - s(x_t, 0)\right] = \hat{c},$$

i.e., the expected benefit from protesting equals the cost of protesting. Using (1) and (2) in the text, we can rewrite this condition as:

$$\hat{c}(\mu, f, x) = \mu p(f) [2 + \chi] \gamma(x).$$

Note that $\bar{c} > \hat{c}(\mu, f, x)$ for all $x \in [\underline{x}, \overline{x}]$ by (3). If $\underline{c} \leq \hat{c}(\mu, f, x)$, there is an equilibrium where all concerned citizens protest when $c_t = \underline{c}$ and the probability of a protest is therefore ρ. It is straightforward to see that a larger share of concerned citizens, μ, and/or a higher gain to democracy, x, increases the incidence of protests, while more incumbent fighting, f, reduces it.[14]

Now consider what happens when $D_t = 0$. The expected payoff to the incumbent leader with his preferred institution is $u^I(x_t, 0) + \widehat{\lambda}(\mu, f)\Gamma(x) - wf_t$, where $\widehat{\lambda}(\mu, f) = [1 - \rho\mu p(f)]$ is the probability of successfully enforcing $D_t = 0$ when devoting f units of labor to fighting.[15]

With democracy $D_t = 1$, we can write the leader's payoff as

$$\widetilde{U}(x_t, f_t) = u^I(x_t, 1) - wf_t, \tag{D.5}$$

which takes into account the fact that no protest occurs in this case.

Choice of f – step 2 There is no incentive to fight when $D_t = 1$ and hence the payoff function under democracy is

$$U(x_t) = \text{Max}_f \widetilde{U}(x_t, f) = u^I(x_t, 1). \tag{D.6}$$

With autocracy, i.e. $D_t = 0$, fighting increases (via $p(f)$) the probability that an occurring protest is successfully defeated. The maximized expected payoff of an incumbent under autocracy ($D_t = 0$) is

$$V(x_t, \mu_t) = u^I(x_t, 0) + \max_{f \geq 0} \left\{\widehat{\lambda}(\mu_t, f)\Gamma(x) - wf\right\}. \tag{D.7}$$

Let $f^*(x, \mu)$ denote the optimal choice of fighting by the incumbent at stage 2 and define the survival function $\lambda(x, \mu) = \widehat{\lambda}(\mu, f^*(x, \mu))$.

[14] There is always an equilibrium where nobody protests. This has the possibility that protests can occur as "sunspot" phenomenon. Here, we assume that the concerned citizens can coordinate on the protest equilibrium when it exists.

[15] This objective function supposes that a passive incumbent-group citizen chooses the level of fighting. If we instead supposed that the decisions were made to maximize the average payoff in the incumbent group, then this would weaken their willingness to fight. Moreover, it would add an additional complementarity between democratic values and institutions, since a larger group of concerned citizens in the incumbent group would imply fewer resources devoted to fighting.

Properties of the equilibrium payoff and survival functions If none of the concerned citizens protest then $f^*(x,\mu)=0$. Given (3) there exists $\tilde{\mu}$ such that

$$\gamma(x)\,\tilde{\mu}(x)\,p(0)\,[2+\chi]=\underline{c}.$$

For $\mu\geq\tilde{\mu}(x)$ all concerned citizens protest when $c=\underline{c}$ and given the condition on $p(f)$ as f goes to zero. In this case, $f^*(x,\mu)$ solves

$$-\rho\mu p'(f^*(x,\mu))\Gamma(x)-w=0$$

The implicit-function theorem implies that

$$\frac{\partial f^*(x,\mu)}{\partial\mu}=\frac{-p'(f^*(x,\mu))}{p''(f^*(x,\mu))\,\mu}>0 \tag{D.8}$$

and

$$\frac{\partial f^*(x,\mu)}{\partial x}=\frac{-p'(f)\,\Gamma'(x)}{p''(f)\,\Gamma(x)}>0. \tag{D.9}$$

Now, we can substitute $f^*(x,\mu)$ into $\hat{\lambda}(x,\mu,\cdot)$ to define the incumbent's expected probability of successful enforcing $D_t=0$ when fighting optimally:

$$\lambda(x,\mu)=[1-\rho\mu p(f^*(x,\mu))].$$

It follows that

$$\lambda_x(x,\mu)=\begin{cases}-\rho\mu p'(f^*(x,\mu))\frac{\partial f^*(x,\mu)}{\partial x}>0 & \text{if } \mu\geq\tilde{\mu}(x)\\ 0 & \text{otherwise.}\end{cases}$$

Assume that

$$\begin{aligned}\lambda_\mu(x,\mu) &= -\rho\mu p'(f^*(x,\mu))\frac{\partial f^*(x,\mu)}{\partial\mu}-\rho p(f^*(x,\mu))\\ &= -\rho\left[\mu p'(f^*(x,\mu))\frac{\partial f^*(x,\mu)}{\partial\mu}+p(f^*(x,\mu))\right]\\ &= -\rho\left[\frac{-\left[p'(f^*(x,\mu))\right]^2}{p''(f^*(x,\mu))}+p(f^*(x,\mu))\right],\end{aligned}$$

which is negative if $\log(p(f))$ is convex. Thus

$$\lambda_\mu(x,\mu)=\begin{cases}-\rho\left[\frac{-[p'(f^*(x,\mu))]^2}{p''(f^*(x,\mu))}+p(f^*(x,\mu))\right]<0 & \text{if } \mu\geq\tilde{\mu}(x)\\ 0 & \text{otherwise.}\end{cases} \tag{D.10}$$

Moreover, we can write

$$V(x,\mu)-U(x)=\Gamma(x)\,\lambda(x,\mu)-wf^*(x,\mu). \tag{D.11}$$

We can use this expression to derive

$$\frac{\partial\left[V(x,\mu)-U(x)\right]}{\partial x}=\Gamma'(x)\,\lambda(x,\mu)+\lambda_x(x,\mu)\Gamma(x)>0 \tag{D.12}$$

and

$$\frac{\partial\left[V\left(x,\mu\right)-U\left(x\right)\right]}{\partial\mu}=\begin{cases}\lambda_{\mu}(x,\mu)\Gamma\left(x\right)<0 & \text{if }\mu\geq\tilde{\mu}\left(x\right)\\ 0 & \text{otherwise.}\end{cases}$$

Hence we have shown that, as stated in the main text of Section 3, for all $\mu\in[0,1]$ and $x\in[\underline{x},\overline{x}]$

1. *A higher x increases $\lambda\left(x,\mu\right)$ and $V\left(x,\mu\right)-U\left(x\right)$.*

2. *A higher μ decreases $\lambda\left(x,\mu\right)$ and $V\left(x,\mu\right)-U\left(x\right)$.*

Proof of Proposition 1 Assumption 1 stated in the text requires that

$$\Gamma\left(\underline{x}\right)\lambda\left(\underline{x},\underline{\mu}\right)-wf^{*}(\underline{x},\underline{\mu})=0,$$

which will hold only if

$$\gamma\left(\underline{x}\right)\underline{\mu}p\left(f^{*}(\underline{x},\underline{\mu})\right)\left[2+\chi\right]\geq\underline{c}.$$

Hence there is both citizen protest by all concerned citizens when $c_t=\underline{c}$ and $f^{*}(\underline{x},\underline{\mu})>0$. Moreover, since $\gamma\left(x\right)$ is increasing then citizens protest for all $\mu\geq\underline{\mu}$ and $x\geq\underline{x}$. The decision rule used by the incumbent is

$$D_t=\begin{cases}0 & \text{if }V\left(x,\mu\right)-U\left(x\right)\geq0\\ 1 & \text{otherwise.}\end{cases}\tag{D.13}$$

Let $\mu^{L}=\underline{\mu}$ as defined in Assumption 1. Then, for all $\mu\leq\mu^{L}$ and $x\in[\underline{x},\overline{x}]$, we will have $D\left(\mu,x\right)=0$. Since $V\left(\overline{x},1\right)-U\left(x\right)<0$ for all $x\in[\underline{x},\overline{x}]$, there exists

$$V\left(\bar{x},\mu^{H}\right)-U\left(\bar{x}\right)=0.$$

Since $V\left(x,\mu\right)-U\left(x\right)$ is increasing, it follows that $\mu^{H}>\mu^{L}$. And for for all $\mu\geq\mu^{H}$, and $x\in[\underline{x},\overline{x}]$, $D\left(x,\mu\right)=1$. Given that $f^{*}(\underline{x},\underline{\mu})>0$, $V\left(x,\mu\right)-U\left(x\right)$ is a continuous function of μ and x for all $\mu\in\left[\mu^{L},1\right]$ and $x\in[\underline{x},\overline{x}]$. Thus, for $\mu\in\left[\mu^{L},\mu^{H}\right]$, the intermediate value theorem implies that there must be a value $\hat{x}\left(\mu\right)\in[\underline{x},\overline{x}]$ such that

$$V\left(\hat{x}\left(\mu\right),\mu\right)-U\left(\hat{x}\left(\mu\right)\right)=0.\tag{D.14}$$

E Dynamic Stability

This final section discusses the dynamic stability of the model.

The signs of $\Delta_{\mu}\left(\mu\right)$ and $d\hat{x}\left(\mu\right)/d\mu$ To rule out a stable interior steady state below it is sufficient that $\Delta_{\mu}\left(\mu\right)\geq0$. This, in turn, is the case if $d\hat{x}\left(\mu\right)/d\mu>0$. To see this, use (5) to compute:

$$\Delta_{\mu}\left(\mu\right)=\begin{cases}\int_{\underline{x}}^{\overline{x}}\gamma\left(x\right)dH\left(x\right) & \mu\geq\mu^{H}\\ -\int_{\widehat{x}(\mu)}^{\overline{x}}L_{\lambda}\left(x,\lambda\left(x,\mu\right)\right)\lambda_{\mu}\left(x,\mu\right)dH\left(x\right)+ & \\ \left[\gamma\left(\hat{x}\left(\mu\right)\right)+L\left(\hat{x}\left(\mu\right),\lambda\left(\hat{x}\left(\mu\right),\mu\right)\right)\right]h\left(\widehat{x}\left(\mu\right)\right)\frac{\partial\widehat{x}(\mu)}{\partial\mu} & \mu\in\left[\mu^{L},\mu^{H}\right]\\ -\int_{\underline{x}}^{\overline{x}}L_{\lambda}\left(x,\lambda\left(x,\mu\right)\right)\lambda_{\mu}\left(x,\mu\right)dH\left(x\right) & \mu\leq\mu^{L}.\end{cases}\tag{E.15}$$

Because $L_{\lambda}>0$ and $\lambda_{\mu}<0$, a sufficient condition for $\Delta_{\mu}\left(\mu\right)\geq0$ for all $\mu\in[0,1]$, is $\partial\widehat{x}\left(\mu\right)/\partial\mu>0$.

Using the definition of $\hat{x}(\mu)$, we can show that this condition is satisfied, because

$$\frac{\partial \widehat{x}(\mu)}{\partial \mu} = -\frac{\partial V/\partial \mu}{\frac{\partial [V(x,\mu)-U(x)]}{\partial x}} = -\frac{\lambda_\mu}{\frac{\partial [V(x,\mu)-U(x)]}{\partial x}} > 0. \tag{E.16}$$

The sign follows from the results in section D, which say that the numerator is negative while the denominator is positive.

Stability We now provide the basic argument as to why only the corner solutions for μ can be stable steady states of the model.

We require that any steady state, $\hat{\mu}$, has to be stable following a small perturbation to $\hat{\mu} \pm \nu$. To prove that only the extremal steady states are stable, we start from

$$\mu_{t+1} - \mu_t = (1-\mu_t)\varsigma^{P,C} - \mu_t \varsigma^{C,P}. \tag{E.17}$$

Note that if $\Delta > 0$ for all $\mu \in [0,1]$ then $\varsigma^{P,C} > 0$ and $\varsigma^{C,P} \leq 0$ and (E.17) is positive so μ converges to one globally. The opposite is true if $\Delta < 0$ for all $\mu \in [0,1]$. Now consider the case where there exists $\hat{\mu}(\sigma)$ such that $\Delta(\hat{\mu}) = 0$. Then since $\Delta(\mu)$ is globally increasing for $\mu \in [0,1]$, then at $\Delta(\hat{\mu}) = 0$, we must have $\mu_{t+1} - \mu_t \geq 0$ for all $1 \geq \mu \geq \hat{\mu}$, while $\mu_{t+1} - \mu_t < 0$ for all $0 \leq \mu < \hat{\mu}$. The interior steady state is therefore unstable. Moreover as $\Delta(\mu)$ is globally increasing, we must have $\Delta(1) \geq 0 \geq \Delta(0)$. Hence

$$\mu_{t+1} - 1 + \nu = (1-\nu)\varsigma^{P,C} - \nu\varsigma^{C,P} > 0$$
$$\mu_{t+1} - \nu = \nu\varsigma^{P,C} - (1-\nu)\varsigma^{C,P} < 0$$

for small enough $\nu > 0$. This implies that the steady states at $\mu = 0$ and $\mu = 1$ are stable as required. ■