Shaping Institutions*

William Fuchs† Satoshi Fukuda‡

March 20, 2026

Abstract

We propose a simple model of the evolution of institutional strength, where leaders' actions have a persistent effect by shaping the norms of the institutions they lead. This leads to different long-run behaviors even for institutions with the same formal rules. The early history of leaders plays a crucial role in determining which outcome prevails. Every period, a leader decides to respect or abuse their position. Respect strengthens the institutions; abuse weakens them. Leader's type and institutional strength determine both the benefit/cost of abusing the position and the replacement probability. We elucidate democratic backsliding and corporate-board capturing.

JEL Codes: D02; D7; C73; G3; M14

Keywords: Institutions; Norms; Leadership; Long-Run Dynamics; Democratic Backsliding; Corporate Board Capturing

*We would like to thank Aydogan Alti, Simon Board, Alessandro Bonatti, Wouter Dessein, Matteo Escudé, Scott Gehlbach, Ole Jann, Niccolò Lomys, Ester Manna, Suraj Prasad, Andrea Prat, Marek Pycia, and Thomas Wiseman for their insightful comments and discussions. We would also like to thank seminar participants at various conferences and universities. We thank Francesco Bilotta, Kevin Mei, and Mahyar Sefidgaran for their excellent research assistantship. Fuchs gratefully acknowledges the support from ERC Grant 681575, Grant PGC2018-096159-B-I00 financed by MCIN/AEI/10.13039/501100011033, and Comunidad de Madrid (Spain), Grant EPUC3M11 (V PRICIT) and Grant H2019/HUM-5891.

†Department of Finance, McCombs School of Business, UT Austin, CEPR, and FTG.

‡Department of Economics, Leavey School of Business, Santa Clara University.

There is scarcely any part of my conduct which may not hereafter be drawn into precedent.
George Washington, in a letter to Catharine Macaulay

1 Introduction

Institutions rely on formal rules, but these alone cannot ensure effective governance. In both political and corporate environments, informal norms play a critical complementary role.[1] These norms constrain behavior, guide expectations, and influence whether formal rules are followed or subverted. Crucially, norms are not static. They evolve over time, shaped by the behavior of those in power. A leader who respects institutional boundaries can strengthen norms and raise the cost of future violations; one who abuses power can weaken them, reducing the costs of future transgressions. This paper develops a dynamic model in which leaders endogenously shape the trajectory of institutional norms over time.

Our main contribution is to formalize the idea that the behavior of early leaders can have a persistent effect on long-run institutional quality. As captured in the epigraph, George Washington was well aware that his actions establish the precedents that shape future governance. We model an institution where an incumbent leader chooses whether to respect or abuse its authority. Norms respond to these actions, creating feedback: stronger norms reduce the benefit of abuse and raise the probability of removal after abuse, while weaker norms have the opposite effect. These dynamics generate path dependence: starting from the same formal rules, one institution may converge to a high-norm equilibrium where abuse is rare, while another may drift into persistent abuse and institutional decay. This mechanism offers a simple and tractable explanation for why societies or organizations with identical legal frameworks can diverge in practice. Crucially, this divergence is impossible when norms are held fixed: it is precisely the feedback between leadership behavior and endogenous institutional strength that generates the multiplicity of long-run outcomes. It also allows for the timely and related question as to whether the American institutional framework is "truly" strong, or if Americans have been lucky in the past to have great leaders who in general did not seriously challenge the institutions and help cement those institutions with strong norms.

The model helps explain democratic backsliding, where leaders gradually undermine institutions without a single dramatic break. Cháves in Venezuela, Erdogan in Turkey, Orbán

[1]For example, as Renan (2018) puts it, "The nature of the presidency in American constitutional governance cannot be understood without reference to norms.... Presidential power is both augmented and constrained by these unwritten rules of legitimate or respectworthy behavior."

in Hungary, and Putin in Russia are recent prominent examples. The slow erosion of different democratic safeguards from freedom of the press, independence of the courts, corruption and abuse of state resources are clearly observed in all of these countries (see Figure 5). By interpreting norms broadly as including respect for these institutions, our model captures this important phenomena. Our model helps explain a related phenomena in corporate governance: corporate board capturing and entrenched leadership behavior. Our approach also helps reconcile the mixed empirical performance of formal institutional reforms: copying legal rules from high-functioning states often fails when the underlying norm environment differs. We also consider the restoration of democracy after institutional collapse and show how success hinges on the emergence of unusually committed leaders. Thus it provides a framework to interpret both the difficulty of reversing autocratic drift and the rare but significant role of transformative leadership.

In our model, the legal framework determines the initial institutional strength. Thereafter, it evolves endogenously as a function of the leader's actions. A leader can either abuse or respect its position. Abuse weakens the institutions by affecting norms. Our use of the term norm is in the spirit of capturing the part of the institutional strength that does not directly derive from formal rules.[2] For example, what used to cause a scandal can become "normal" behavior. Conversely, norms are also strengthened after they have been respected.

In turn, the institutional norms influence the behavior of the leader in two dimensions. Firstly, the weaker the institution, the larger the payoff the leader can reap from abusing its position. Secondly, abusing power can affect the possibility of staying in office in two ways. First, misbehavior can be scandalous and increase the likelihood of being replaced. Second, in the opposite direction, abuse can allow for more political patronage or election meddling favoring the incumbent.

The leader's behavior also depends on her type. The leader's type determines the relative flow benefit of being in office under both actions. In our model, higher types are either more moral or less skilled at cheating and reap less benefit from abusing the position.

This feedback loop between leader behavior and endogenous institutional strength allows us to establish our main result: institutions that begin with similar formal rules can diverge sharply depending on the character of early leaders. A string of norm-respecting leaders can entrench accountability and create a high-norm regime. Conversely, a series of abuses can degrade institutional quality, shift expectations and reduce accountability. Over time, norms become self-reinforcing: strong norms make abuse rare, while weak norms enable its persistence. The possibility of divergence underscores the fragility of institutional integrity

[2]In turn, institutional strength is similar to the notion of "de facto power of civil society" as described in Yuchtman (2024).

in early stages of development.

This feature of the model is essential for understanding why several countries, such as Argentina, even though they modeled their constitutions after the United States, seem to be in a very different steady state (see, for instance, Alston and Gallo, 2010). Of course, there can be many factors explaining such long-term differences, but the US might also have been somewhat lucky with the leaders it has had. Remarking on President Trump's damage to American democracy, Kamarck (2021) points out "Fortunately, we haven't had many of those in our 200-plus years of history."

This endogenous evolution of norms also helps rationalize the concerns about the long-term effects of President Trump's disregard for several institutional traditions. This widely-held sentiment was captured by Foran (2016): "Growing tolerance for conflicts of interest in government, limitations on media access and accountability, and harsh treatment of minority groups can accumulate.... Each norm that falls is one fewer safeguard against executive over-reach than we had before. Even if we never become an authoritarian state, our governance will suffer as a result. For now, we should recognize the precedents that are already being set and try to prevent them from becoming the new normal." Looking forward, Pfiffner (2021) points out: "The broader impact of President Trump's behavior will depend crucially on the character of future presidents." Our model captures such long-lasting effect of a leader through the interaction between taking an abusive action such as undermining the independence of media and the evolution of norms.

Several empirical papers provide additional support to these concerns by pointing out the importance of path dependence in shaping institutions.[3] La Porta et al. (1999) demonstrate the role of exogenous political historical factors in explaining government performance. Acemoglu et al. (2008) argue that cross-sectional relationship between democracy and income today is the result of societies embarking on divergent development paths at certain historical critical junctures. Papers such as Acemoglu et al. (2001) and Glaeser et al. (2004) demonstrate persistence of institutional outcomes. Our paper suggests that, when evaluating the quality of governance, it is important to condition on the history of past leaders, as the behavior of past leaders may have a persistent effect on the behavior of a current leader through the evolution of institutional strength.

Consistently with this evidence, our paper shows (i) how countries or corporations with similar formal rules may diverge based on the early leaders' behavior; and (ii) that there is a threshold of institutional strength, where strength becomes either persistently self-eroding or self-strengthening. Our paper provides a dynamic micro-foundation for the transition

[3]See, for instance, North (1990), Pierson (2000), Currie et al. (2016), and Acemoglu et al. (2021) for an overview.

between multiple equilibria.

Norms and their evolution are tricky objects to model, and there is no consensus on an appropriate way to pin down their evolution. From a modeling point of view, we consider our approach to be complementary to the existing ones. An advantage of our modeling strategy is that our assumed law of motion is simple yet rich enough to capture what might be considered desirable features of a fully micro-funded model without needing to take a stance on a particular model or dealing with the additional modeling complexities it would naturally entail.

Our baseline model is purposely crafted to offer a transparent framework highlighting the mechanism by which for given formal rules, institutional trajectories can diverge, based on the behavioral feedback between leadership and institutional strength. Clearly illustrating how small differences in leadership behavior—especially in early periods—can have lasting consequences. This tractability allows us to isolate core forces behind institutional resilience and decay, while still accommodating heterogeneity in leader types and outcomes.

As we detail in Sections 5 and 6, the model can be enriched in several dimensions. In Section 5 we introduce a voting model that microfounds the replacement probability used in the main analysis. We also explain how institutional safeguards such as judicial independence, political competition, or media scrutiny can be interpreted via the lens of our model. Accountability operates through two margins: an extensive margin, governing how many voters the incumbent can capture through patronage or electoral manipulation, and an intensive margin, governing how well-informed the remaining voters are—itself a function of media independence and political competition. Leaders who dismantle checks and balances not only lower the cost of abuse but also reduce the likelihood of being replaced through norm destruction. This piecemeal erosion of accountability allows for gradual yet persistent institutional decay and helps rationalize empirical patterns of democratic backsliding, even in settings where formal rules remain unchanged.

In Section 6, we examine how term limits shape a leader's incentives. On one hand, shorter terms can trigger end-of-term myopia and increase abuse. On the other hand, since institutional erosion requires time to yield "benefits," shorter limits can also deter long-term abuse.

The paper is structured as follows. The rest of this section discusses the related literature. Section 2 lays out the model. Section 3 contains our main analysis: Section 3.1 characterizes a leader's decision, and Section 3.2 studies dynamics of institutional strength. Section 4 addresses democratic backsliding. Section 5 endogenizes the replacement of a leader. Section 6 provide discussions such as term limits. Section 7 provides concluding remarks. Proofs are in Appendix A. Appendix B, available online, discusses additional extensions.

Related Literature

The legal and political-science literature has long emphasized the roles of informal rules and norms on the quality of governance, as early as Bryce (1888 [1995]). Renan (2018) and Ahmed (2022) study how "presidential norms" augment and constrain presidential powers. O'Donnell (1996) and Linz (1978, 1990) discuss the role of informal rules and leaders' behavior on democratic consolidation, a process through which democracies consolidate lowering the risk of reverting to authoritarianism (e.g., O'Donnell and Schmitter, 1986). Azari and Smith (2012) and Levitsky and Ziblatt (2018) study the roles of informal rules and norms on democratization and autocratization. Levitsky and Way (2015), Huq and Ginsburg (2018), and Diamond (2021) point out that democratic backsliding in the world has been caused not by coups but by elected governments, suggesting the importance of constitutional norms. Our contribution in these strands of literature is to provide a micro-founded process in which institutions are gradually strengthened or weakened as a result of past actions. This allows us to obtain endogenous and possibly differing long-run configurations of institutional strength.

There is now emerging literature on democratic backsliding.[4] Helmke et al. (2022) study two parties trying to gradually tilt the electoral rules (e.g., gerrymandering). Grillo and Prato (2023) show in their static model that democratic backsliding can occur when minorities are willing to accept violations of democratic norms and politicians value popular support.

Howell et al. (2023) model an executive who, despite judicial review, incrementally undermines checks and balances, leading to a persistent accumulation of authority. While we also link current behavior to future outcomes, their framework lacks path dependence and divergent long-run steady states. In Luo and Przeworski (2023), incumbents strategically erode institutions to increase survival probabilities—occurring when voters prioritize policy over democratic integrity or when unpopular leaders see no other path to retention. Unlike our model, however, they assume past leadership does not shift the incentives or survival of successors. To the best of our knowledge, our paper provides the first formal model that elucidates the role of the evolution of norms on democratic backsliding. We also make a novel contribution by characterizing long-run dynamics and path dependence in democratic backsliding.

Our model also connects to work on corruption as a self-reinforcing process. Andvig and Moene (1990) present a static model of corruption with multiple equilibria which tries to explain why the same socio-economic structure can give rise to different levels of corruption. Shleifer and Vishny (1993) study a static model in which economic and political competition can reduce the level of corruption. In the empirical literature, Tanzi (1998) points out the role

[4]See, for instance, Lust and Waldner (2015) and Grillo et al. (2024) for surveys on democratic backsliding in the political science literature.

of the example provided by the political leadership. Paldman (2002) point out that countries with similar backgrounds can drift into very different corruption regimes (e.g., Argentina and Chile). Our paper formalizes the role that the current political leadership plays on the behavior of the future leaders and thus a rationale for the persistence of corruption.

There is also broad literature on leadership (e.g., Jones and Olken, 2005; Myerson, 2011). The role of leadership in our model diverges from the existing literature by focusing on how an incumbent's behavior exerts a permanent effect on the actions of their successors. In this way we also distinguish ourselves from the broad literature that discusses leadership and culture (e.g., Ashforth and Anand, 2003; Biggerstaff et al., 2015; Guiso et al., 2015). Our paper also suggests the importance of conditioning on the history of past leaders in evaluating the quality of governance.

There are strands of literature that view norms as equilibrium objects. Focusing on institutional norms, Invernizzi and Ting (2024) view norms as the expected play in the efficient subgame perfect equilibrium of a policy game, in the spirit of Dixit et al. (2000). Although they can study the interaction of formal or informal rules, they do not really have a focus on the norm dynamics. Unlike our model, there are no changing leader types and thus on path norms do not weaken or strengthen. Hence, they cannot address democratic backsliding nor the importance of path dependence for different steady states.

Using a game-theoretic approach, Bidner and Francois (2013) focus on the role of norms on democratic consolidation. In their setting, voters' willingness to punish transgressions depends on the expectation of future accountability, which increases the value of selecting a new leader. Similarly, Svolik (2013) highlights how a "trap of pessimistic expectations" can lead voters to abandon accountability altogether after poor economic shocks. While leaders in these models may lose the incentive to perform, they do not strategically dismantle the system. In contrast, our model centers on the leader's intertemporal decision-making. This allows us to capture the strategic logic of democratic backsliding: how a leader deliberately erodes institutions to cement her own power and capture future rents.

Although it is theoretically appealing to think of norms emerging as equilibrium objects of complex dynamic games, we think our approach has important advantages. As is well-understood, dynamic games naturally have multiple equilibria. Thus, we will ultimately be forced to make some sort of ad-hoc assumptions to refine the set of equilibria. Furthermore, if we are interested in the dynamics over different equilibria the applied refinement must "react" to past play. Our model ultimately achieves the same without the unnecessary burden and notational complexity that would be required in a fully-fledged dynamic game. The parsimonious nature of the model also allows for a clear insight into the forces responsible for the equilibrium dynamics.

Our choice to instead capture norms by their resulting impact on institutional strength brings us closer to the literature on the role of "social capital" on the functioning of governments (e.g., Putnam, 1993; Guiso et al., 2016).[5] Persson and Tabellini (2009) study "democratic capital," as measured by a nation's historical experience with democracy and the incidence of democracy in its neighborhood. They demonstrate that democratic capital reduces the exit rates from democracy and raise the exit rates from autocracy. Besley and Persson (2019) model democratic values, defined as the proportion of citizens who may fight for democracy against autocracy. Yet, there are a few important differences with respect to our work. Chiefly, we highlight the importance of leader types in determining future outcomes. In particular, our main point is that leaders can have a permanent effect. In contrast, in their model the long-run outcomes are not history dependent. Their model does not lend itself to analyze (nor do they discuss) democratic backsliding.

In a broader context, Dessein and Prat (2022) model "organizational capital," an intangible asset that has to be maintained by a leader. The leader faces whether to increase organizational capital or boost short-term profit. They characterize a steady state distribution of organizational capital in which otherwise similar firms may have persistent performance differences. Although similar long-run dynamics can arise in our model, mechanisms are very different. In particular, in our setting, a more patient leader may have less incentives to improve norms. In contrast, in their model, a more patient leader has stronger incentives to invest in organizational capital.[6]

2 Model

Every period $t \in \{1, 2, \dots\}$, the incumbent leader must decide on one of two actions $a_t \in \{0, 1\}$. The action $a_t = 1$ represents the leader abusing her position or cheating. In contrast, $a_t = 0$ represents the leader abiding by or respecting the (unwritten) rules. The leader's time-t payoff from taking either of these actions is determined by two elements: (i) the type, $h \in \mathbb{R}$, representing the leader's level of honesty (e.g., the leader's sense of fiduciary duty) or ability to cheat; and (ii) the institutional strength $N_t \in \mathbb{R}$. Specifically, we assume

$$u(a_t, N_t, h) := b - a_t(N_t + h).^{7}$$

[5] Almond (1956) argues the role of "political culture" on the functioning of government (see also Almond and Verba, 1963; Diamond, 1999).

[6] Besley and Persson (2024) study a different model in which organizational culture is formulated as the distribution of "types" in an organization which affect project choices.

[7] We implicitly assume that violating a norm is bad for society. One could easily modify the model to a situation in which respecting a norm does not necessarily bring a positive value: for example, whether the supreme court adheres to precedents or not and whether a president issues an executive order or not.

The first term $b \geq 0$ is the benefit of being in power (e.g., non-pecuniary benefits from holding office and pecuniary benefits such as wages and office perks).[8] Thus, if the leader respects the rules (i.e., $a_t = 0$), then the payoff is b. If the leader abuses her position (i.e., $a_t = 1$), then the payoff is $b - (N_t + h)$. The stronger the institution, the higher the cost of abusing (i.e., the deviation from respecting the position). Likewise, the higher the honesty type, the higher the cost of abusing.[9]

A natural extension of the model would be to consider the case in which both actions and norms are a vector (see Online Appendix B.1). For ease of exposition, we constrain our analysis to the scalar case.

Importantly, when deciding which action to take, the leader also takes into consideration how her action, together with the institutional strength, affects her probability of remaining in power. We denote the replacement probability at time t by the function $\lambda(a_t, N_t)$. We assume: (i) $0 \leq \lambda(a_t, N_t) \leq 1$; (ii) $\lambda_0(N_t) := \lambda(0, N_t)$ is non-increasing and continuous; and (iii) $\lambda_1(N_t) := \lambda(1, N_t)$ is non-decreasing and continuous. The first assumption is needed since λ is a probability. Assumptions (ii) and (iii) imply that $\lambda_1(N_t) - \lambda_0(N_t)$ is non-decreasing in N_t. This is meant to capture the idea that the stronger the institution, the more likely it is that abusing power will lead to losing the position. As we do not impose any assumption on the relative magnitudes of λ_1 and λ_0, our setup allows for the existence of $\tilde{N}$ such that for $N_t < \tilde{N}$ abusing power enhances the probability of remaining in office $\lambda_1(N_t) - \lambda_0(N_t) < 0$ while for $N_t > \tilde{N}$ abusing power lowers the probability of remaining in office $\lambda_1(N_t) - \lambda_0(N_t) > 0$. The interpretation of this is that when the institutions are weak, abusing power allows the politician to engage in activities that might help her get re-elected such as: patronage and clientelism, bread and circuses, or directly meddling with the elections, while facing little risk of a scandal. See Section 5 for discussions on how λ could capture political competition, independence of media or judicial independence. The model also allows for the extreme possibility that when the institutions are sufficiently weakened a leader can guarantee remaining in power by abusing her position, i.e., $\lambda_1(N_t) = 0$.

We move on to specifying the evolution of institutional strength N_t. There is no consensus as to how to formalize norms, their evolution or more importantly their impact on institutional strength. Our approach is to use a specification that allows us to capture both the role of formal rules $\overline{N} \in \mathbb{R}$ and norms as governed by the history of actions by past

[8] Our main results extend to the case in which the benefit from being in power depends on institutional strength N_t or the current action a_t. See Online Appendix B.3.

[9] Thus, an interesting case is when $N_t + h$ takes a negative value. Also, our main results extend to the case in which the payoff from abusing is separable and decreasing in N_t and h.

leaders in a tractable yet flexible way. Specifically, we assume $N_1 = \overline{N}$ and

$$N_{t+1} = (1-\delta)N_t + \delta\overline{N} + (1-2a_t)\gamma. \tag{1}$$

If the leader respects the rules, $a_t = 0$, then the institutions are strengthened by $\gamma \geq 0$. If the leader abuses her position, $a_t = 1$, then the institutions are weakened by γ. Thus, γ measures the short-run sensitivity of institutional strength to behavior.[10]

The parameter $\delta \in (0, 1]$ is akin to a rate of depreciation in capital accumulation models. Lower δ implies a longer lasting impact of current actions on future norms and institutional strength. The first two terms, $(1-\delta)N_t + \delta\overline{N}$, have the effect of mean reversion to $\overline{N}$. This highlights the sense in which the formal written rules have a more persistent role. In fact, with the absence of the effect of the leaders' actions, the institutional strength converges to $\overline{N}$ in the long run. Despite this force, however, we will demonstrate that the institutions may be absorbed into different regimes when norms are affected by the leaders' actions.

Lemma 1. *Assume $\delta < 1$ and $\gamma > 0$.*

1. *$N_t \in \left(\overline{N} - \frac{\gamma}{\delta}, \overline{N} + \frac{\gamma}{\delta}\right)$ for all $t \in \mathbb{N}$.*
2. *If $a_t = 0$, then $N_{t+1} > N_t$.*
3. *If $a_t = 1$, then $N_{t+1} < N_t$.*

Lemma 1 implies that $N_{t+1} < N_t$ if and only if $a_t = 1$.[11] For ease of notation, we denote by $N_L := \overline{N} - \frac{\gamma}{\delta}$ and $N_H := \overline{N} + \frac{\gamma}{\delta}$.

For a leader of type h, the discounted value from following the leader's strategy $a = (a_t, a_{t+1}, \dots)$ at time t given the institutional strength N_t is:

$$V(N_t, h \mid a) := \sum_{s=t}^{\infty} \beta^{s-t} \Pi_s u(a_s, N_s, h),$$

where $\beta \in (0, 1)$ is the leader's discount factor and Π_s denotes the probability that the leader

[10] To capture the possibility that institutions are more easily undermined than strengthened, we can define the law of motion for institutional strength as $N_{t+1} = (1-\delta)N_t + \delta\overline{N} + (1-a_t)\gamma_R - a_t\gamma_A$, A and R denote abuse and respect, respectively, and $\gamma_A > \gamma_R$ formalizes the bias toward institutional decay. Additionally, we can explore a variant where norm updating is contingent on the leader's survival (λ), such that changes occur only if the incumbent remains in power. These modifications do not qualitatively alter the main findings.

[11] The main results remain robust to alternative specifications of institutional dynamics, provided that the sequence $(N_t)_{t\in\mathbb{N}}$ satisfies the final two properties of Lemma 1. The first property—requiring N_t to be bounded—is necessary only to characterize institutional strength dynamics in a manner consistent with Theorem 2. The main results go through with any initial condition $N_1 \in \left(\overline{N} - \frac{\gamma}{\delta}, \overline{N} + \frac{\gamma}{\delta}\right)$.

is still in power in a given future period s. It can be defined recursively as:

$$\Pi_s := \begin{cases} 1 & \text{if } s = t \\ (1 - \lambda(a_{s-1}, N_{s-1}))\Pi_{s-1} & \text{if } s > t \end{cases}.$$

Lastly, if the leader gets replaced at the end of time t, then a new leader is drawn from a distribution with full support $H_t = \left[\underline{h}_t, \overline{h}_t\right] \subseteq \mathbb{R}$. Although the evolution of H_t plays no role in determining the leader's decision at time t, it can have implications for the long-run properties of the institution. We will first consider the case $H_t = H$ for all t and postpone to Section 6.3 the case of an endogenous evolution of H_t.

3 Main Analysis

We divide our main analysis into two subsections. Section 3.1 studies the optimal sequence of actions for a given leader with a given honesty type. In principle the leader could choose an arbitrary sequence of actions but, importantly, we are able to show that it is optimal for the leader not to switch from one action to another. This allows us to derive an explicit closed-form characterization for the cutoff type for given institutional strength N which we denote $\tilde{h}(N)$ and also the leader's value function. With that important property established, Section 3.2 studies the dynamics.

3.1 Characterization of a Leader's Decision

Consider a leader with honesty $h \in H$ when the institutional strength is $N \in (N_L, N_H)$. The leader's problem can be stated recursively by the following Bellman equation:

$$\begin{aligned} V(N, h) = \max_{a \in \{0,1\}} \; & b - a(h + N) + \beta \left(1 - \lambda(a, N)\right) V(N', h) \\ & \text{subject to } N' = (1 - \delta)N + \delta\overline{N} + (1 - 2a)\gamma. \end{aligned}$$

Note that the value function V that satisfies the above Bellman equation exists uniquely. To see this, the right-hand side of the Bellman equation is well-defined, as the maximum is taken over the binary actions. Then, existence and uniqueness follow from the fact that the operation that defines the right-hand side of the Bellman equation is a contraction mapping, as the usual Blackwell conditions are satisfied.

To characterize the leaders' optimal actions, consider the effects of choosing abuse versus choosing respect. Firstly, the flow payoff changes. If the leader abuses at time t, then she

gets an extra $-(N_t + h)$ flow payoff at t. Secondly, there are two additional effects on the continuation payoffs. First, by abusing, the probability of staying in power in the next period changes from $1 - \lambda_0(N_t)$ to $1 - \lambda_1(N_t)$. Second, conditional on remaining in power, the continuation value changes from $V((1-\delta)N_t + \delta\overline{N} + \gamma, h)$ to $V((1-\delta)N_t + \delta\overline{N} - \gamma, h)$. Given these various effects, it is hard to solve for the optimal policy directly.

Instead, we rely on the property that $N_{t+1} > N_t$ if and only if $a_t = 0$ (Lemma 1) to make progress. This property implies that if there exists a non-increasing threshold function $\tilde{h}$ such that the policy function a^* satisfies the following: the leader of type h takes $a^*(N, h) = 1$ if $h < \tilde{h}(N)$ and $a^*(N, h) = 0$ if $h > \tilde{h}(N)$, then the optimal action sequence is constant over time. To see this, suppose that it is optimal for the leader to abuse today, i.e., $h < \tilde{h}(N_t)$. Then, since $N_{t+1} < N_t$ and thus $h < \tilde{h}(N_t) \leq \tilde{h}(N_{t+1})$, it is optimal for the leader to abuse tomorrow as well.

Next, we guess and verify that the threshold function $\tilde{h}$ is non-increasing. Given the conjecture, if $h < \tilde{h}(N)$, then the leader abuses the position forever, as $h < \tilde{h}(N) \leq \tilde{h}(N_t^1)$, where N_t^1 denotes the decreasing path when $a = (1, 1, \dots)$ with $N_1^1 = N$. Thus,

$$V(N, h \mid (1, 1, \dots)) = \sum_{t=1}^{\infty} \beta^{t-1} \left(\prod_{s=1}^{t-1} (1 - \lambda_1(N_s^1)) \right) \left(b - \left(N_t^1 + h \right) \right).$$

On the other hand, if $h > \tilde{h}(N)$, then the leader respects forever, as $h > \tilde{h}(N) \geq \tilde{h}(N_t^0)$, where N_t^0 denotes the increasing path when $a = (0, 0, \dots)$ with $N_1^0 = N$. Thus,

$$V(N, h \mid (0, 0, \dots)) = \sum_{t=1}^{\infty} \beta^{t-1} \left(\prod_{s=1}^{t-1} (1 - \lambda_0(N_s^0)) \right) b.$$

Then, the threshold function can be computed by solving for

$$V(N, \tilde{h}(N) \mid (0, 0, \dots)) = V(N, \tilde{h}(N) \mid (1, 1, \dots)). \tag{2}$$

This is because, if the leader's type is $\tilde{h}(N)$ when the institutional strength is N, then she is indifferent between abusing forever and respecting forever. On the one hand, since the replacement probability λ_0 is non-increasing in N and the flow payoff is constant when $a = (0, 0, \dots)$, the left-hand side of Expression (2) is non-decreasing in N and does not depend on h. On the other hand, since the replacement probability λ_1 is non-decreasing in N and the flow payoff is non-increasing in N and h when $a = (1, 1, \dots)$, the right-hand side of Expression (2) is non-increasing in N and h. Therefore, it must be the case that $\tilde{h}$ is non-increasing in N.

This allows us to obtain closed-form solutions for the value function in both cases and verify that indeed the implied optimal policy threshold function is non-increasing in N as conjectured.[12] Formally:

Theorem 1. *The leader's optimal action is constant over time. For any given $N \in (N_L, N_H)$, there exists $\tilde{h}(N) \in \mathbb{R}$ such that if $h < \tilde{h}(N)$ the leader abuses her position and if $h > \tilde{h}(N)$ the leader respects the rules. The threshold $\tilde{h}(N)$ is non-increasing in N and is given by:*

$$\tilde{h}(N) = \left(1 - \frac{\sum_{t=1}^{\infty} \beta^{t-1} \left(\prod_{s=1}^{t-1}(1-\lambda_0(N_s^0))\right)}{\sum_{t=1}^{\infty} \beta^{t-1} \left(\prod_{s=1}^{t-1}(1-\lambda_1(N_s^1))\right)}\right) b - N_L$$
$$- \frac{\sum_{t=1}^{\infty} (\beta(1-\delta))^{t-1} \left(\prod_{s=1}^{t-1}(1-\lambda_1(N_s^1))\right)}{\sum_{t=1}^{\infty} \beta^{t-1} \left(\prod_{s=1}^{t-1}(1-\lambda_1(N_s^1))\right)} (N - N_L), \tag{3}$$

where $(N_t^0)_{t=1}^{\infty}$ denotes the increasing path when $a = (0, 0, \dots)$, $N_{t+1}^0 = (1-\delta)N_t^0 + \delta\overline{N} + \gamma$, with $N_1^0 = N$; and $(N_t^1)_{t=1}^{\infty}$ the decreasing path when $a = (1, 1, \dots)$, $N_{t+1}^1 = (1-\delta)N_t^1 + \delta\overline{N} - \gamma$, with $N_1^1 = N$.

When the replacement probability λ does not depend on institutional strength, denoting by λ_0 and λ_1, respectively, the replacement probabilities after respect and abuse, the threshold function $\tilde{h}$ given by Expression (3) reduces to a simple affine equation:

$$\tilde{h}(N) = \frac{\beta(\lambda_0 - \lambda_1)}{1 - \beta(1-\lambda_0)} b - N_L - \frac{1 - \beta(1-\lambda_1)}{1 - \beta(1-\delta)(1-\lambda_1)}(N - N_L). \tag{4}$$

3.2 Dynamics of Institutional Strength

We now study the dynamics of institutional strength. We assume that the set from which leader types are drawn, H, is a compact interval $H = [\underline{h}, \overline{h}]$ with $-\infty < \underline{h} < \overline{h} < \infty$ and that, when a leader is replaced, the next leader's type is drawn (independently of histories) from a distribution F_H with full support H. At time $t = 1$, the institution starts with $N_1 = \overline{N}$. A leader with type h_1 is drawn according to the distribution F_H. Then, the leader makes her decision $a_1 = a^*(\overline{N}, h_1)$, which leads to the institutional strength $N_2 = \overline{N} + (1 - 2a_1)\gamma$ at the beginning of the next period. In period $t \geq 2$, with probability $1 - \lambda(a_{t-1}, N_{t-1})$, the incumbent stays in power: $h_t = h_{t-1}$. Otherwise, with probability $\lambda(a_{t-1}, N_{t-1})$, a new leader with type $h_t \in H$ is drawn. In either case, the leader at time t takes $a^*(N_t, h_t)$, which determines the institutional strength N_{t+1} at the beginning of the next period. For ease of presentation, this subsection assumes $\lambda(\cdot, \cdot) \in (0, 1)$.[13]

[12] Note that the stationarity of the environment implies that $\tilde{h}$ is time-independent. As we show in Section 6.2, term limits would make $\tilde{h}$ time-dependent, and thus it might be optimal for a leader to switch actions.

[13] Theorem 2 can be easily modified when we allow for $\lambda(\cdot, \cdot) \in [0, 1]$.

To highlight the importance of the endogenous norms, we first discuss the case in which norms are constant, i.e., $N_t = \overline{N}$. This is the case when $(\delta, \gamma) = (1, 0)$. In this case, there are three possibilities: (i) if $\tilde{h}(\overline{N}) > \overline{h}$ then all types want to abuse power and that is the only outcome observed; (ii) if $\tilde{h}(\overline{N}) < \underline{h}$ then no type abuses power and rules are always respected; and (iii) $\overline{h} > \tilde{h}(\overline{N}) > \underline{h}$ then there is a subset of types that would abuse power and a subset that wouldn't. As a result, we will observe transitions from abuse to respect and vice versa as the type of a leader changes.

Importantly, with constant norms it is not possible for two countries to have very different long-run outcomes if they start with the same initial condition. In contrast, when norms evolve endogenously this arises as a possibility. To see this, consider a situation as in (iii) above with $\overline{h} > \tilde{h}(\overline{N}) > \underline{h}$. Now suppose in one country the initial sequence of elected leaders has $h_1 > \tilde{h}(\overline{N})$ and thus no abuse takes place. This implies that the institution gets stronger and as a result, the cutoff type decreases $\tilde{h}(\overline{N}) > \tilde{h}(N_1) > \tilde{h}(N_2) \ldots$. If the string of good leaders is sustained sufficiently long, then we might reach a point in which $\tilde{h}(N_t) < \underline{h}$ and, at this point, the institution is so strong that even if the worst possible leader is elected she will still respect the rules. As a result, the institution will just keep getting stronger, $N_t \to N_H$, and the rules will always be respected from then on. Yet, for the same initial condition, the opposite might also be possible. A draw of bad leaders early on, who choose to abuse the institution, can lead the institution to weaken to a point at which $\tilde{h}(N_t) > \overline{h}$. From that point on, not even the best possible leader would respect the rules. Thus, rules are never again respected and the institutional strength just keeps on drifting down: $N_t \to N_L$. Thus, we can have two very different absorbing steady states.

Our main result formalizes this discussion. For ease of notation, denote by $\tilde{h}(N_L) := \lim_{N \downarrow N_L} \tilde{h}(N)$ and $\tilde{h}(N_H) := \lim_{N \uparrow N_H} \tilde{h}(N)$.

Theorem 2. *The following four cases characterize the long-run dynamics.*

1. *If (i) $\underline{h} < \tilde{h}(N_H)$ and (ii) $\overline{h} < \tilde{h}(N_L)$, then $N_t \downarrow N_L$ almost surely along any path.*
2. *If (i) $\underline{h} > \tilde{h}(N_H)$ and (ii) $\overline{h} > \tilde{h}(N_L)$, then $N_t \uparrow N_H$ almost surely along any path.*
3. *If (i) $\underline{h} < \tilde{h}(N_H)$ and (ii) $\overline{h} > \tilde{h}(N_L)$, then there exists a full-support limit distribution on $N_\infty \in (N_L, N_H)$.*
4. *If (i) $\underline{h} > \tilde{h}(N_H)$ and (ii) $\overline{h} < \tilde{h}(N_L)$, then almost surely along any path, either $N_t \downarrow N_L$ or $N_t \uparrow N_H$. There exists a limit distribution on $N_\infty \in \{N_L, N_H\}$.*

In Case 1, depicted in the top left panel of Figure 1, almost surely along any path, the institutional strength converges to the lowest level. Put differently, the leaders' actions

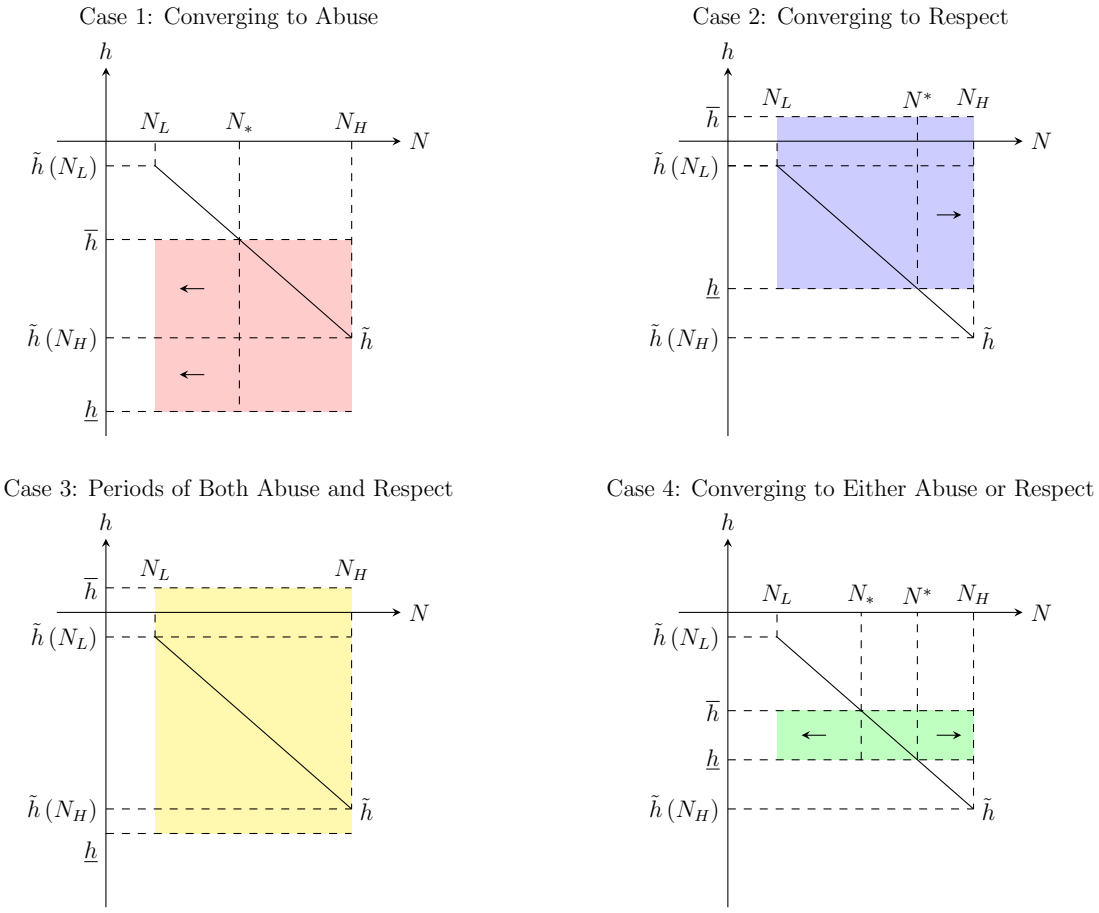

Figure 1: Theorem 2. The top left panel depicts Case 1; the top right panel Case 2; the bottom left panel Case 3; and the bottom right panel Case 4. For each panel, the colored rectangle depicts the set of honesty types $H = [\underline{h}, \overline{h}]$ for institutional strength $N \in (N_L, N_H)$.

satisfy $a_t = 1$ for all but finitely many times. For this to be the case we need two conditions to hold: (i) for any institutional strength, there are some types who want to abuse the position; and (ii) once the institution is sufficiently weak (i.e., below N_* in the panel), even the most honest type wants to abuse. The first condition implies that any path almost surely reaches sufficiently low institutional strength and the second that once that point is reached it is absorbing.

In contrast, in Case 2, depicted in the top right panel of Figure 1, almost surely along any path, the institutional strength converges to the highest level. The leaders' actions satisfy $a_t = 0$ for all but finitely many times. The conditions for this are the exact opposite to Case 1: (i) there must always be a type willing to respect the rules; and (ii) once the institution is sufficiently strong (i.e., above N^* in the panel), no type wants to abuse.

For a non-degenerate limiting distribution to exist, Case 3, depicted in the bottom left panel of Figure 1, it must be the case that (i) however strong the institution is, some types

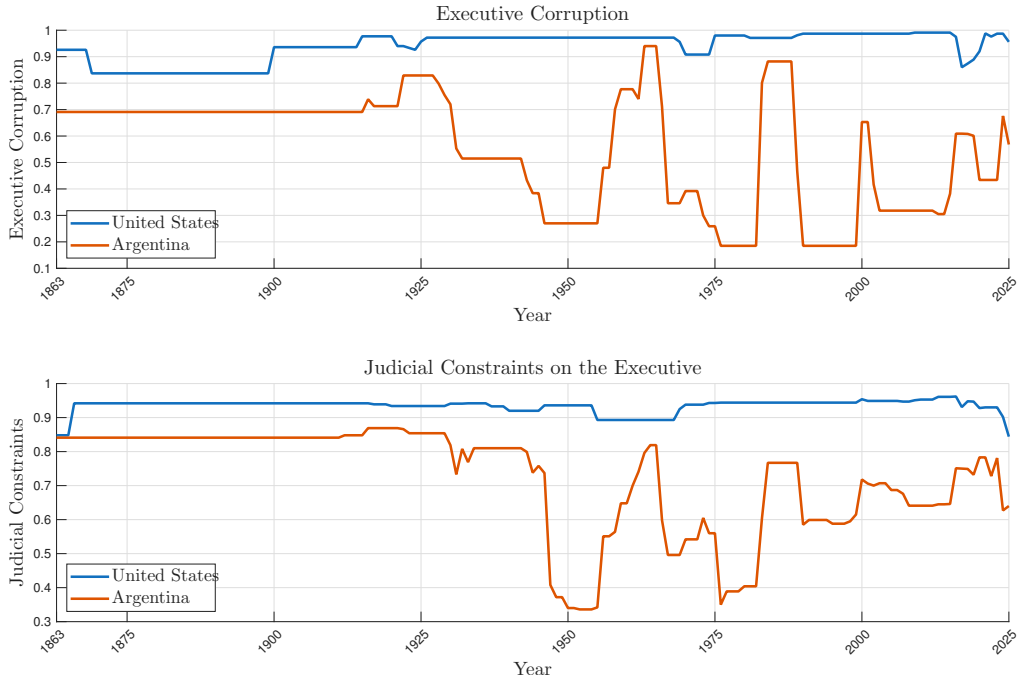

Figure 2: Path Dependence: Executive Corruption Index and Judicial Constraints on the Executive Index from V-Dem (Coppedge et al., 2026).

are willing to abuse; and that (ii) however weak the institution is, some types are willing to respect the rules. In this case, the leaders take $a_t = 0$ and $a_t = 1$ infinitely often. In their study of democratic reversals, Kapstein and Converse (2008) document that, among those democracies that were reversed, some such as Pakistan, Peru, and Thailand, experienced second and even third democratization episodes.

Case 4, depicted in the bottom right panel of Figure 1, is perhaps the most interesting and highlights the importance of how leaders can shape the institutions. For this case to arise, (i) once the institution is sufficiently strong (i.e., above N^* in the panel), no type wants to abuse, and (ii) once the institution is sufficiently weak (i.e., below N_* in the panel), even the most honest type wants to abuse. In this case, almost surely along any path, the institutional strength either converges to N_L or converges to N_H. Put differently, almost surely along any path, either $a_t = 1$ for all but finitely many times or $a_t = 0$ for all but finitely many times. This happens because a string of very honest leaders, who respect the rules, can raise through their actions the norms and thus the institutional strength to a sufficiently high point such that when this level is reached, a future leader, however low her type is, never abuses the position. Conversely, if a sequence of bad leaders abuse the position, then the institution might become so weak that even if a better leader will be elected she will still be tempted to abuse the position.

Figure 2 illustrates how, while both started with similar scores on both measures and similar institutions, Argentina and the United States have followed very different paths in terms of executive corruption and judicial constraints on the executive.[14] Until recently, the US has slightly improved over time.[15] Instead Argentina has followed a more volatile path and seems to have converged to a lower steady state. The two measures naturally move together since the executive has an incentive to undermine the courts as a way to be able to get away with their illegal actions. They also have incentives to remain in power to make use of the immunity it grants. For example, the former President Cristina Fernández de Kirchner was found guilty of corruption and sentenced to 6 years in prison. Her political immunity as vice President kept her out from jail for several years, while her government tried very hard to manipulate the courts to be able to steer her legal proceedings. President Trump himself claimed that "He who saves his country does not violate any law."[16]

Theorem 2 speaks to the persistent effect that early leaders can have on institutions or the culture of organizations. Thus, young organizations must devote extra care in the selection of their leaders. In the political context, Keefer (2007) and Kapstein and Converse (2008) document that young democracies are especially at risk of reversal, and suggest that the absence of checks and balances such as political competition is among the most powerful predictors of democratic failure. In the context of our model, a young democracy whose institutional strength is between N_* and N^* is indeed at risk of reversal and the democratic failure may occur as the institutional strength deteriorates. The absence of checks and balances corresponds to a lower λ_1, under which leaders tend to abuse more and norms are more likely to deteriorate.

To highlight path dependence further, it is important to reiterate that Case 4 only obtains when norms endogenously respond to the leaders' actions. If we let $(\delta, \gamma) = (1, 0)$, then we get $N_t = \overline{N}$ for all t. In this case, Cases 1-3 are still possible but not Case 4. In fact, if we were in Case 4 before and we made $(\delta, \gamma) = (1, 0)$ then we would find ourselves in Case 3.

Corollary 1. *If $(\delta, \gamma) = (1, 0)$, then an optimal action sequence cannot converge to two*

[14] In the figure, Executive Corruption Index measures "how routinely do members of the executive, or their agents grant favors in exchange for bribes, kickbacks, or other material inducements, and how often do they steal, embezzle, or misappropriate public funds or other state resources for personal or family use." To be consistent with our usage of institutional strength, we plot $1 - x_{i,t}$ where $x_{i,t}$ is country i's Executive Corruption Index at time t. Thus, it ranges from 0 (high corruption) to 1 (low corruption). Judicial Constraints on the Executive Index, which is a continuous measure from 0 (low) to 1 (high), measures the extent to which the executive respects the constitution and comply with court rulings and the judiciary is able to act in an independent fashion. The starting point in the figure is 1863, ten years after the ratification of the Constitution of the Argentine Nation, which was modeled after the Constitution of the United States.

[15] In 2025, Judicial Constraints on the Executive Index dropped from 0.902 to 0.884.

[16] https://www.reuters.com/world/us/trump-if-it-saves-country-its-not-illegal-2025-02-16/ (Date of Access: March 20, 2026).

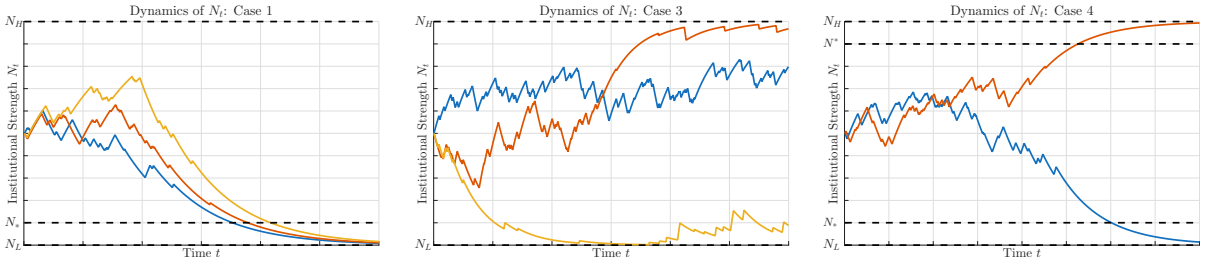

Figure 3: Institutional Dynamics. The left panel depicts Case 1: along any path, eventually the institutional strength converges to the lowest level. The central panel depicts Case 3: a stationary distribution on N_∞ exists. The right panel depicts Case 4: with the same initial conditions, the institutional strength converges to either the lowest or highest level.

constant actions, i.e., Case 4 is not possible.

Together, Theorem 2 and Corollary 1 provide the first formal characterization of how endogenous norm evolution generates path-dependent divergence in long-run institutional outcomes. Unlike the existing literature on democratic backsliding—where past leadership does not shift the incentives or survival prospects of successors (Luo and Przeworski, 2023), or where path dependence and divergent steady states are absent (Howell et al., 2023)—Case 4 shows that institutions governed by identical formal rules can converge to fundamentally different absorbing states, with the early realization of leader types as the sole determinant of which prevails. Crucially, as Corollary 1 establishes, this divergence is impossible when norms are held fixed: it is precisely the feedback between leadership behavior and endogenous institutional strength that generates the multiplicity of long-run outcomes.

Figure 3 depicts the dynamics of institutional strength.[17] The left panel depicts Case 1: along each path, eventually the institutional strength converges to the lowest level. The central panel depicts Case 3, in which the institutional strength remains ergodic. Note, however, that for a significant amount of time, the institutional strength is close to one of the extremes. This is because once it reaches such a level, the mass of types that switch the action relative to their predecessor, even though it is positive, is relatively small. In addition, $\lambda_1(N_L)$, the replacement probability after abuse at the limit N_L, is small.[18]

The right panel depicts Case 4: starting from the initial formal rules, an institution can converge to two very different steady states. Distinguishing Case 3 from Case 4 might not be empirically very easy since, in practice, the type space could have long but thin tails. Yet it is important that in both cases the key is that we will have path dependence and persistence close to the extremes. Figure 4 depicts the simulated long-run distribution of institutional

[17]For our numerical simulations, we discretize $H = \{h^1, \ldots, h^n\}$ with $h^1 = \underline{h}$ and $h^n = \overline{h}$ and assume that F_H is a uniform distribution.

[18]Our discussion in Section 4.1 is also closely related to this point.

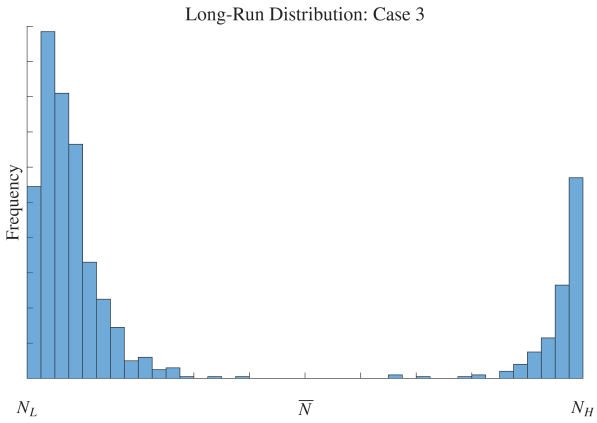

Figure 4: Simulated Long-Run Institutional Strength Distribution for Case 3.

strength for Case 3. Beyond the fact that most of the density is in the extremes, it is worth noting that this distribution is likely to be skewed towards the left. This is the case because when institutions are very weak regimes can last longer. Instead, when institutions are strong, turnover is higher.

Mainwaring and Bizzarro (2019) document 91 transitions (in 79 countries) to democracy from 1974 to 2012. Consistently with our theory, their result suggests extreme long-run outcomes of either full breakdown or restoration. Namely, out of these 91 cases of "third-wave" democratization, 62 experienced either breakdowns or stagnation at a low level; 27 either achieved major democratic advances or attained high levels of democracy from their first year of democracy to 2017; and in 2 (Ecuador and Poland), levels of democracy eroded substantially while the regime remained a democracy per their classification.

4 Democratic Backsliding

As discussed in the law and political science literature, many autocracies are the result of a slow erosion of institutions rather than a rapid wholesale shift (e.g., Huq and Ginsburg, 2018). As illustrated in Figure 5, recent examples include Cháves in Venezuela, Erdogan in Turkey, Orbán in Hungary, and Putin in Russia.[19] High-income countries and older democracies may also experience a slow erosion of institutions, even if they do not transition into an autocracy.

For this, it is useful to consider the "abuse" action as including ones such as replacing key figures that might play an important role in limiting the leader's power. Three relevant

[19]Electoral Democracy Index plotted in the figure is an aggregate measure of free and fair elections as well as freedom of expressions and associations, which ranges continuously from 0 (low) to 1 (high).

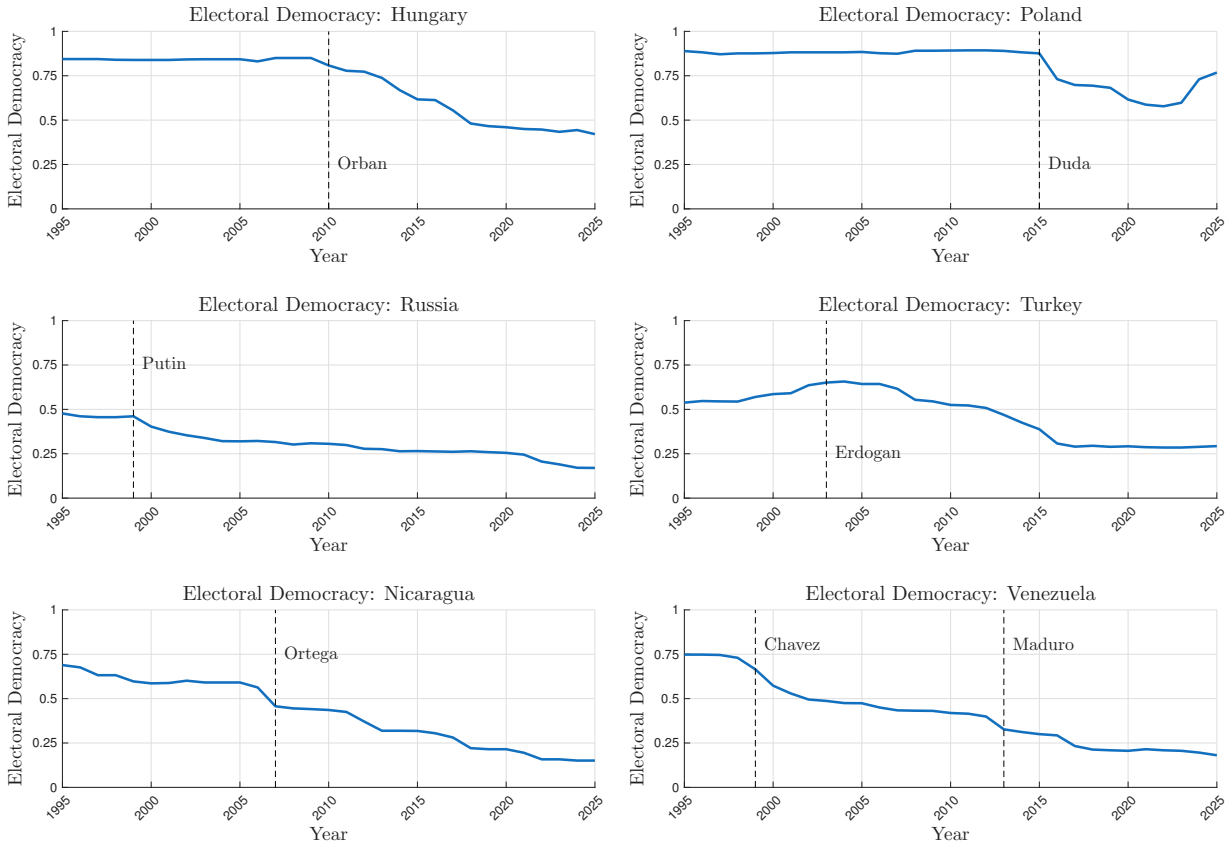

Figure 5: Democratic Backsliding. Each panel depicts Electoral Democracy Index from V-Dem (Coppedge et al., 2026).

examples are: (i) changing the composition of the courts, for example by expanding the supreme court; (ii) changing the people in charge of running/supervising elections from honest brokers to puppets; and (iii) manipulating public debate through media. President Trump's attempts to overturn the 2020 election were to a large extent derailed by Department of Justice leaders that were unwilling to do his bidding. Indeed, President Trump has strongly endorsed many candidates in the 2022 election largely on the basis that they denied the outcome of the 2020 election. Under Prime Minister Orbán in Hungary, "elections rules have been modified 20 times, paralysing opposition parties; and Fidesz has heaped pressure on the independent judiciary" (Szelényi, 2022).

Our model can capture democratic backsliding if the replacement probability $\lambda_1(N)$ decreases as the leader undermines the institutional safeguards piece by piece. Such piecemeal subversion of norms is less visible and attracts less resistance than a wholesale shift such as a coup. As in Venezuela, Turkey, Hungary, and Russia, many leaders who have subverted democratic norms have retained the support of a majority or a ruling coalition through several election cycles. In the limit, we could reach the situation in which the leader consolidates

herself as an autocrat and is never replaced, i.e., $\lambda_1(N) = 0$, once the institutions are sufficiently eroded. In this sense, while Hungary and Poland have suffered a significant erosion of their democratic institutions their leaders are not quite as cemented in power as those of Russia, Nicaragua, and Venezuela. Indeed, following the election in Poland in 2023, the new government led by Prime Minister Donald Tusk "started (re-)democratization immediately trying to undo state capture and restore freedom of expression" (Nord et al., 2026).

Finally, it is worth noting that even if the economy has experienced a long history of respectable behavior leading to strong institutions, democratic backsliding can occur in our model. This speaks to the concerns of what the return of President Trump to power would entail for American institutions. As Kagan (2023) put it, "There is a clear path to dictatorship in the United States, and it is getting shorter every day." As reported by Nord et al. (2026), these fears seem to be well justified: "[t]he speed with which American democracy is currently dismantled is unprecedented in modern history."

4.1 Restoration of Democratic Practices

Once an economy has fallen into a despotic regime rather than $\lambda_1(N) = 0$, we might think that there is still a very small probability of replacing the current leader. In this case, it is natural to think that the set of possible replacement types would also differ.

For ease of presentation, consider $H_t = \left\{h_{t-1}, h^{\mathrm{h}}\right\}$, where h_{t-1} can represent the despot replaced by a family member (as in North Korea) or a political rival that would continue with the current practices (as when one war lord deposes another). Instead, h^{h} represents a hero type that is willing to potentially risk her life to depose the current leader.[20] There are several historic figures that we might associate with such a type. For ease of presentation, assume h^{h} is such that this type would not abuse the position once in power. This would give institutions a chance to recover and reestablish the necessary checks and balances for a proper functioning of democracy.

This, of course, is not easy and can help explain the difficulty in restoring democratic practices in former autocratic regimes. Since institutions are very weak, the temptation for a new leader to abuse is very high. This is particularly hard when such a heroic figure is absent or replaced too soon, such as evidenced recently in Egypt, Libya, and Yemen. This might also help explain why regime changes from the outside tend to fail. Diamond (2021) presents a list of 20 countries where mass public protests or an unexpected defeat of an authoritarian incumbent might have resulted in a transition to democracy for the period of

[20]This is similar in spirit to the "prominent" agents in the model by Acemoglu and Jackson (2015) that can restore cooperative behavior.

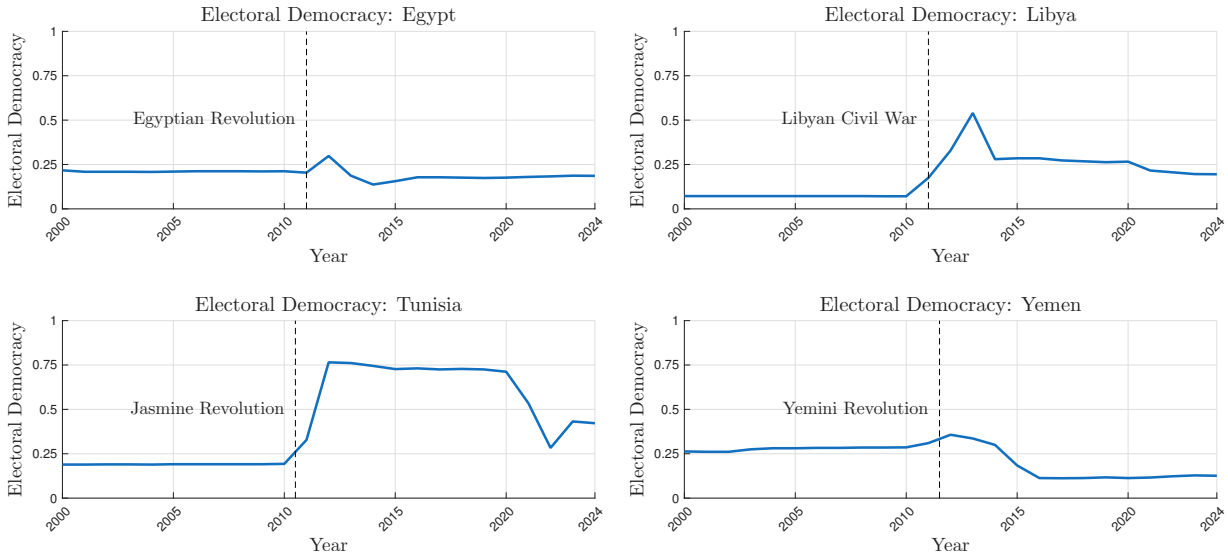

Figure 6: Restoration of Democratic Practices. Each panel depicts Electoral Democracy Index from V-Dem (Coppedge et al., 2026).

2009-2020. At the time the paper was written, only 2 out of 20 countries in the list (one of which is Tunisia to be discussed below) had resulted in democratic transitions.[21]

Figure 6 illustrates several examples of these failed attempts at restoring democratic practices from the Arab Spring. A particularly interesting case to highlight is Tunisia which until recently it seemed to have managed to succeed. Former Prime Minister and President Essebsi played an important role on that initial success. Unfortunately, upon his death in 2019, Kais Saied, who won the first presidential election in which a presidential debate was held, dismissed the parliament and carried out a self-coup in 2021. He has ruled by decree and passed a new constitution since then.

A related problem is that faced by young nations at the end of colonial rule. As African countries gained independence in the second half of the 20th century, several struggled in terms of consolidating strong democratic practices. As highlighted in our model, early leadership played a very important role. In this respect, Chad with Tombalbaye, Libya with Gaddafi, and Zimbabwe with Mugabe represent some of the worst performers. In contrast, Khama played a very positive role in Botswana. The left panel of Figure 7 compares Electoral Democracy Index for Botswana and Zimbabwe. The right panel of Figure 7 compares Electoral Democracy Index for Senegal and Chad, both of which gained independence in 1960 from France. Senegal presents a more gradual path towards democratic consolidation. Although its first president Leopold Senghor has a mixed record, he did open up political competition towards the end of his mandate and had a peaceful transition to his handpicked

[21] Under weak institution, the replacement probability λ_0 after respect may be high.

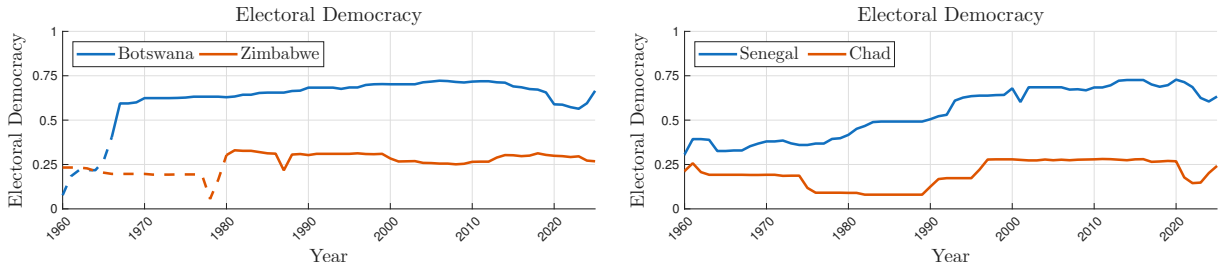

Figure 7: Post-Colonial Rule. Each panel depicts Electoral Democracy Index from V-Dem (Coppedge et al., 2026). Dashed line indicates colonial periods.

successor Abdou Diouf. He continued opening up political competition and had the first peaceful transition to an opposition party.

5 Endogenizing Accountability and Leader Replacement

In our main analysis, the reduced-form replacement probability λ allowed us to characterize both the leader's decision and institutional dynamics. Although we did not commit to a particular micro-foundation, the generality of λ in fact makes it possible to capture a wide range of environments. In this section we develop one such micro-foundation—a noisy-signal voter model—that simultaneously (i) pins down λ_0 and λ_1 as explicit functions of institutional strength N_t, and (ii) satisfies the assumptions on λ required for Theorems 1 and 2. We then use this model as a common lens through which to interpret the roles of media independence, political patronage, and political competition.

5.1 The Noisy-Signal Voter Model

There is a unit mass of citizens. The incumbent is replaced at the end of the period if the fraction voting *against* her exceeds an exogenous threshold $\tau \in (0, 1)$, which can be interpreted as reflecting the constitutional election rules and possible effects from unfair elections. Citizens are endogenously partitioned each period into two groups: (i) *core voters*, with mass $\mu(a_t, N_t)$, who always support the incumbent; and (ii) *retrospective voters*, with mass $1 - \mu(a_t, N_t)$, who stochastically participate, update their position on a noisy signal of the leader's behavior and punish perceived abuse.

Extensive margin: capture function. The core-voter share is

$$\mu(a_t, N_t) := m_0 + c(N_t) \cdot a_t, \tag{5}$$

where $m_0 \in (0,1)$ is the baseline share of loyal supporters (party-line voters, ideological supporters, etc) and $c : (N_L, N_H) \to (0, 1 - m_0)$ is the *capture function*, non-increasing in N_t. When the incumbent abuses ($a_t = 1$), she can convert $c(N_t)$ additional citizens into core supporters through patronage, vote-buying, or electoral manipulation: the weaker the institution, the more citizens she can capture. When she respects ($a_t = 0$), no such conversion occurs and the retrospective voter mass is $1 - m_0$.

Intensive margin: noisy signals. Each retrospective voter i observes a *private* binary signal $s_i \in \{0, 1\}$ drawn independently conditional on the incumbent's action:

$$P(s_i = 1 \mid a_t) := \begin{cases} \eta(N_t) & \text{if } a_t = 1, \\ 1 - \eta(N_t) & \text{if } a_t = 0, \end{cases}$$

where $\eta : (N_L, N_H) \to (\frac{1}{2}, 1)$ is non-decreasing in N_t. The parameter $\eta(N_t)$ is the signal precision: a stronger institution (more independent media, more robust political opposition) makes the signal more informative about whether abuse occurred. Each retrospective voter votes *against* the incumbent if $s_i = 1$.

Stochastic participation and replacement probabilities. In each period t, retrospective voters participate with probability ξ_t, where $\xi_t \sim \text{Uniform}[0,1]$ i.i.d., independent of signals and the leader's type h_t.[22] By the law of large numbers, conditional on ξ_t, the fraction of all citizens voting against the incumbent equals $\xi_t \cdot D_{a_t}(N_t)$, where the *effective opposition mass* is

$$D_0(N_t) := (1 - m_0)(1 - \eta(N_t)) \text{ and} \tag{6}$$

$$D_1(N_t) := (1 - m_0 - c(N_t))\eta(N_t). \tag{7}$$

[22]Since the action depends on type, there is still a correlation between voting and incumbent type. For explicit reputation models, see, for instance, Myerson (2006) and Kartik et al. (2025).

The incumbent is replaced whenever this fraction exceeds τ, i.e., whenever $\xi_t > \tau / D_{a_t}(N_t)$. Since ξ_t is uniform, the replacement probabilities are

$$\lambda_0(N_t) = \max\left(1 - \frac{\tau}{(1-m_0)(1-\eta(N_t))}, 0\right) \text{ and} \tag{8}$$

$$\lambda_1(N_t) = \max\left(1 - \frac{\tau}{(1-m_0-c(N_t))\eta(N_t)}, 0\right). \tag{9}$$

Note that $\lambda(a_t, N_t) < 1$ in both cases, since the denominator $D_{a_t}(N_t)$ is strictly positive. As noted below in Remark 1, denominator of $\lambda_1(N_t)$ tends to zero, so $\lambda_1(N_t) \to 0$, capturing the autocratic consolidation in which an abusive leader faces a vanishing probability of removal.

Two margins of accountability. Expressions (8) and (9) make two accountability margins explicit. The first is extensive margin (capture function c): the incumbent who abuses converts a mass $c(N_t)$ of retrospective voters into core supporters, shrinking the pool that can punish her. Weaker institutions allow larger conversions. The second is intensive margin (signal precision η): among the remaining retrospective voters, the probability that each one correctly identifies abuse is $\eta(N_t)$. A more independent media or stronger opposition raises $\eta(N_t)$, making it harder for the incumbent to escape accountability.

We now confirm that the micro-founded λ_0 and λ_1 satisfy the assumptions imposed in Section 2.

Proposition 1. *In the noisy-signal voter model, λ satisfies the following properties.*

1. *$\lambda_0(N_t)$ is non-increasing in N_t.*
2. *$\lambda_1(N_t)$ is non-decreasing in N_t.*

Consequently, $\lambda_1(N_t) - \lambda_0(N_t)$ is non-decreasing in N_t.

Remark 1 (Autocratic consolidation)**.**

1. As $N_t \to N_L$, if the capture function satisfies $c(N_t) \to 1 - m_0$, then $\lambda_1(N_L) \to 0$: a leader who abuses under maximally weak institutions faces vanishing replacement risk, consistent with the autocratic consolidation limit in Section 4 where we allowed $\lambda_1(N_t) = 0$.

2. The explicit condition for abuse to *lower* replacement risk relative to respect, $\lambda_1(N_t) < \lambda_0(N_t)$, reduces under (8)–(9) to

$$(1 - m_0 - c(N_t))\eta(N_t) < (1 - m_0)(1 - \eta(N_t)),$$

which would hold when N_t is sufficiently small (so that $c(N_t)$ is close to $1 - m_0$ and $\eta(N_t)$ is close to $\frac{1}{2}$). This would also pin down the threshold $\tilde{N}$ in Section 2.

The rest of this section uses the noisy-signal voter model as a common lens to interpret media independence, political patronage, and political competition. In each case we show how the relevant institution maps into one or both margins and how its erosion through abuse generates the dynamic feedback central to democratic backsliding.

5.2 Media

The role of media in democratic governance has long been recognized. In a letter to James Currie, Thomas Jefferson eloquently wrote: "Our liberty depends on the freedom of the press, and that cannot be limited without being lost." The First Amendment to the United States Constitution protects freedom of the press. The press has been recognized as the fourth estate or the fourth power. A strong and independent media is important to hold leaders accountable for their actions.

Media as signal precision. Within the noisy-signal voter model, media independence operates primarily through the *intensive margin*: it determines how accurately retrospective voters can infer whether the leader has abused her position. Formally, let the signal precision η be a non-decreasing function of institutional strength N_t, where part of what N_t captures is the independence of the press. When media is free and investigative, $\eta(N_t)$ is high: bad behavior is more likely to be reported, and voters correctly identify abuse with higher probability. When media is compromised, $\eta(N_t) \to \frac{1}{2}$: the signal is nearly uninformative and voters cannot distinguish abuse from legitimate governance. This affects $\lambda_1(N_t)$ directly through $D_1(N_t)$: a less independent press lowers $\eta(N_t)$, reduces $D_1(N_t)$, and thereby lowers the probability that an abusive leader is removed.

Good deeds are over-reported under a captured press. A captive press not only fails to expose abuse—it can distort signals in ways that benefit the incumbent regardless of her action. A fully Bayesian voter who observes the degree of media capture would discount the signal accordingly, attenuating this effect. Two considerations partially restore it. First, voters may not be fully Bayesian about capture: because institutional erosion is gradual and opaque, citizens continue to respond to signals at face value. Second, a co-opted press can be viewed as an information designer who strategically chooses signal precision—suppressing the punishment signal when the leader abuses while preserving noise when she respects. Either way, $\lambda_1(N_t) - \lambda_0(N_t)$ widens as N_t falls, in line with Proposition 1.

Dynamic feedback. Beyond the within-period effect on accountability, the abuse action may itself undermine media independence, creating an additional dynamic channel. If the

abuse action at period t includes, say, granting privileged access to pro-incumbent outlets or enacting legislation that captures broadcast licensing, then the resulting decrease in N_{t+1} (via the law of motion (1) for institutional strength) directly lowers $\eta(N_{t+1})$. This tightens the intensive-margin accountability in future periods: $D_1(N_{t+1})$ falls and $\lambda_1(N_{t+1})$ declines. The endogenous norm evolution therefore generates a self-reinforcing dynamic: early abuse weakens media, which reduces accountability, which lowers the cost of further abuse. The study of backsliding cases by Haggard and Kaufman (2021a,b) concludes that "the media and courts are typically early targets in the cases that end up regressing to autocratic rule. Delegitimizing or shutting down the media creates an alternative reality that strengthens autocratic discretion." Our model provides a formal mechanism for this sequence.

After the 2016 failed coup attempt, Turkey closed around 150 media organizations, including major newspapers, and jailed around 160 journalists.[23] Nicaraguan journalist Carlos F. Chamorro describes the demolition of the rule of law in Nicaragua as proceeding precisely through the "criminalisation of both freedom of the press and freedom of expression."[24] In both cases, early media capture corresponds in our model to a reduction in $\eta(N_t)$ that lowered $\lambda_1(N_t)$, making subsequent abuse cheaper. This persistent intertemporal benifit helps rationalize the empirical pattern that media intervention is typically an *early* move in backsliding episodes.

In the literature on media and political accountability, Besley and Prat (2006), Fearon (2011), and Guriev and Treisman (2020) all show that greater media scrutiny disciplines politicians.[25] Our model nests this insight: $\eta(N_t)$ is the intensive-margin counterpart to their media-scrutiny parameter, and the feedback through the law of motion for institutional strength generates an additional dynamic that is absent in their static or partial-equilibrium analyses.

5.3 Political Patronage

Politicians can distort state resources to reward wealthy voters and interest groups for electoral support. When the incumbent leader is less constrained to distort state resources, she is more likely to be re-elected (Wantchekon, 2013).

[23] https://www.bbc.com/news/world-europe-36910556 and https://www.reuters.com/article/us-turkey-security-newspaper/turkish-court-orders-release-of-journalists-during-their-trial-idUSKCN1GL2OR (Date of Access: June 12th, 2023).

[24] https://reutersinstitute.politics.ox.ac.uk/news/2023-reuters-memorial-lecture-how-report-under-dictatorship-lessons-nicaragua-and-beyond (Date of Access: June 12th, 2023).

[25] For surveys on theoretical and empirical work on media capture, see Prat and Strömberg (2013) and Strömberg (2015). The model of Besley and Prat (2006) can also be applied to a setting in which the agent who engages in media capture is not a government, such as a corporation.

Patronage as voter capture. Within the noisy-signal voter model, political patronage operates primarily through the *extensive margin*: it increases the core-voter mass $\mu(a_t, N_t)$. Formally, the capture function c in Expression (5) encapsulates this force. When the institutions are weak, state resources are more readily diverted for electoral purposes: $c(N_t)$ is large. When institutions are strong, legal constraints and public scrutiny raise the cost of patronage: $c(N_t)$ is small. Under abuse ($a_t = 1$), the fraction of citizens who can be converted into reliable supporters is $c(N_t)$, shrinking the mass of retrospective voters to $1 - m_0 - c(N_t)$ and thereby reducing $D_1(N_t)$. This lowers $\lambda_1(N_t)$, making it easier for the patron incumbent to survive. Under respect ($a_t = 0$), no conversion occurs, so $\lambda_0(N_t)$ is unaffected by c.

Dynamic feedback. As with media, there is a dynamic feedback between patronage and norm evolution. If the abuse action at time t includes state-resource diversion that increases $c(N_{t+1})$ (because weaker institutions reduce legal scrutiny of procurement or public-spending audits), then the extensive margin widens over time. The leader therefore has an intertemporal incentive to erode the institution early in order to expand the clientelist network available to her in later periods. This is consistent with the observation, stressed by Acemoglu et al. (2004), that kleptocratic regimes survive precisely because a well-organized elite coalition captures enough rents to sustain loyalty. In the language of the model, m_0 is endogenously augmented by $c(N_t)$, and the erosion of N_t via abuse expands the set of buyable voters over time.

Symbiotic elite support. If we view elites as the relevant "voters" in the model, kleptocracy corresponds to a situation where $c(N_t)$ is large even for moderate values of N_t, reflecting the existence of a patron–client network that benefits directly from the incumbent's abuse. This symbiotic relationship is an extreme form of patronage in which the capture function c is essentially flat and large, sustaining high λ_1-suppression across a wide range of institutional strengths. Theorem 2 then implies that the economy is likely to fall into Case 1 (convergence to the low-norm steady state) regardless of initial conditions, which is consistent with the persistence of kleptocratic regimes documented by Acemoglu et al. (2004).[26]

5.4 Political Competition

Competition among political parties is vital to democracy, as it creates a system of checks and balances. The institutional arrangements that protect political opponents, ensure fair

[26] Acemoglu and Robinson (2008) and Bidner et al. (2015) study models in which institutions can be entrenched by elites. For a survey on patronage distribution/clientelism in autocratic regimes, see Brancati (2014).

elections, and guarantee freedom of association are important elements of a well-functioning democracy.

Political competition operates on both margins. Through the lens of the noisy-signal voter model, political competition affects *both* margins simultaneously.

First, the *intensive margin*: a strong opposition party acts as an information intermediary, gathering evidence of abuse, aggregating it, and communicating it to voters. This raises the signal precision $\eta(N_t)$: retrospective voters are more likely to correctly identify an abusive action because the opposition amplifies the relevant information. Formally, $\eta(N_t)$ captures not only media independence but also the capacity of the opposition to monitor and publicize the incumbent's behavior.

Second, the *extensive margin*: in the presence of competitive alternatives, voters are harder to capture because credible exit options are available. A citizen who can readily vote for a competitive opposition party is less susceptible to clientelist inducements. This reduces the capture function c: institutional arrangements that protect political competition lower the mass of voters the incumbent can convert into core supporters through patronage.

Both effects together imply that suppressing political competition—by eliminating opposition parties, manipulating electoral rules, or creating an uneven playing field—simultaneously lowers $\eta(N_t)$ and raises $c(N_t)$. By Expressions (6)–(7), this lowers $D_1(N_t)$ and raises $D_0(N_t)$, widening the range of N_t over which $\lambda_1(N_t) < \lambda_0(N_t)$, i.e., the range over which abusing power actually *lowers* replacement risk.

Dynamic feedback. The strategic logic of suppressing competition is particularly transparent in the model. An incumbent who abuses at time t causes $N_{t+1} < N_t$. If the abuse action includes handicapping the opposition—through legal harassment, funding restrictions, or outright banning—then both $\eta(N_{t+1})$ and $1 - m_0 - c(N_{t+1})$ decrease, compounding the direct norm-erosion effect. Russia provides prominent examples of this logic: Mikhail Khodorkovsky's assets were seized and he was imprisoned after founding Open Russia, a civil-society organization; Alexei Navalny was poisoned, subsequently imprisoned and finally killed. In the model, both actions correspond to moves that reduce $D_1(N_t)$ in future periods by eliminating the opposition's capacity to raise $\eta(N_t)$, thereby lowering $\lambda_1(N_t)$ and making the incumbent more secure. If we allow participation ξ_t to be driven by the charisma or competence of the opposition leader, eliminating the most electable ones also secures the leaders position. Stalin's purge of all the original Bolsheviks is rationalized through this lens. This can be captured in the model by a change in the distribution of ξ_t.

Electoral fraud. A particularly direct form of competition suppression is outright electoral manipulation. The threshold τ in our model naturally captures this: a higher τ means that a larger fraction of citizens must vote against the incumbent to trigger her removal, reflecting gerrymandering, biased electoral administration, or voter suppression that raises the effective hurdle for losing office. As mentioned before, under Prime Minister Orbán in Hungary, electoral rules have been modified twenty times (Szelényi, 2022)—each modification corresponding in the model to an upward shift in τ that directly lowers $\lambda(a_t, N_t)$ for any given N_t, independently of the signal precision or capture function.

6 Discussions

This section provides comparative statics results and additional discussions such as the role of term limits and endogenous leader types.

6.1 Comparative Statics

Our closed-form characterization of the threshold function $\tilde{h}$ given by Equation (3) allows us to perform comparative statics.[27]

A stronger set of formal rules (i.e., a higher $\overline{N}$) implies (i) that leaders would be more likely to respect the rules; and (ii) a higher probability of getting absorbed into the steady state of respect.[28] Although this clearly suggests we would want to start with strong formal rules, this is not easy. Determining an effective set of formal rules from observed outcomes is hard since it requires to condition on the sequence of leader types. For example, while the US constitutional framework is usually regarded as being strong, there have been many examples of countries adopting very similar frameworks yet experiencing very different outcomes.

An increase in δ, which governs the reversion to $\overline{N}$, may be construed as conferring less flexibility to the interpretation of the constitution and thus allowing less room for the role of informal rules. When the leaders abuse the position, the institutional strength decreases more slowly from the initial level. Thus, leaders are less able to influence their future replacement probability and their flow payoff. When the replacement probability λ_0 after respect stays the same, leaders are less likely to abuse the position. Thus, conferring less flexibility to the interpretation of the constitution may deter democratic backsliding.

[27] Sharper results can be obtained when the replacement probability λ_0 after respect does not depend on the institutional strength, i.e., when $\tilde{h}$ is given by Expression (4).

[28] In our model, there is no countervailing force favoring a weaker set of formal rules. An interior optimum level of formal rules might arise when flexibility is valuable and a stronger set of rules may hinder it and backfire. See Gratton and Lee (2024) and Invernizzi and Ting (2024).

When institutions are more malleable, which corresponds to a higher γ, leaders are able to decrease the replacement probability and increase the flow payoff in the future. When the replacement probability λ_0 after respect stays the same, leaders have more incentives to abuse their position.

As discussed in Section 5, one can interpret λ_1 as the scrutiny of media, political competition, or the independence of the supreme court.[29] As oversight increases, the likelihood of abuse decreases.

Next, we consider the effect of the benefit b of being in office. It can be seen from Equation (3) that the effect of b on the leader's behavior depends on the replacement probability λ. For simplicity, we focus on the case in which $\lambda_1(\cdot) > \lambda_0(\cdot)$: the abuse action is more likely to lead to losing the position for any institutional strength. In this case, since the coefficient on b in Equation (3) is negative, an increase in b leads to a decrease in the threshold $\tilde{h}(N)$ for given institutional strength, meaning that the leader is more likely to respect the institution. In other words, the leader is more likely to respect the institution for the "re-election" motives. This comparative statics result is consistent with the empirical findings of Ferraz and Finan (2010) and Gagliarducci and Nannicini (2013) that a salary increase for politicians improves political performance, although they also point out another effect that a salary increase leads to a better selection of politicians.[30] However, if the institution is sufficiently weak, then it is possible that $\lambda_1(N) < \lambda_0(N)$, and thus an increase in b may incentivize a leader to abuse the position.

Next, we consider the effect of the discount factor β.[31] Let us suppose first that the leader's action has no effect on the replacement probability, i.e., $\lambda_1 = \lambda_0$. It is important to note that even in such a case, the leader's problem is not static because the future payoff from abusing is affected by its actions today. In particular, suppose that the leader is currently indifferent between abusing forever or always respecting. If we increase β, that would increase the benefit of abusing because the benefits of abusing are increasing over time due to the weakening of the institution while the payoff from respect is constant over time. If, in addition, $\lambda_1 \neq \lambda_0$, then there is a further consideration arising from the change in the replacement probability. If abusing lowers the replacement probability, $\lambda_1(N) < \lambda_0$, then this effect reinforces the leaders' incentive to abuse as we increase β. Instead, if respect lowers the replacement probability, $\lambda_1(N) > \lambda_0$, then there is a countervailing force. This

[29] In the corporate setting, one can interpret λ_1 as the independence of the corporate board or the strength of the minority shareholder rights.

[30] In our model, if H_t is endogenized and if a higher b leads to a "higher" H_t (in a set-theoretical sense or putting a larger mass on higher $h \in H_t$), then the leader is more likely to respect the institution as well.

[31] While we can interpret β literally as the discount factor, β may also be affected by the prevalence of political assassinations.

effect can dominate when b is sufficiently large.

Thus, when the replacement probability is constant over time, an increase in β leads to a higher threshold, i.e., leaders are more likely to abuse their position. However, in general, the sign of the comparative statics with respect to β depends on a particular functional form of λ.

Finally, we could also consider the effect of distributions on H (for simplicity, assume $H = [\underline{h}, \overline{h}]$). When a distribution first-order stochastically dominates another, under the former distribution, the leaders are likely to be of a higher type and thus the long-run outcome is more likely to be an absorption into the high steady state. When it comes to second-order stochastic dominance, however, the effect of variance is not straightforward anymore, as for a weak institution, higher variance might give a higher chance for the reversal, i.e., the absorption into the high steady state.

6.2 Term Limits

For analytical convenience, we focused on a stationary model with no explicit term limits. Yet, since it is interesting to study them from a policy perspective, we consider here the role of term limits: there exists a time T such that a leader will be replaced for sure if she has served for T periods. First, constant actions may not necessarily be optimal. In particular, a term limit may encourage a leader to switch her action from respect to abuse toward the end of the term.[32] This arises, for instance, when the benefit b is sufficiently high, the leader's type h is low, and the replacement probabilities satisfy $\lambda_1(N) - \lambda_0(N) > 0$. Consider the two-period model. In this case, abusing in the first period is costly because of the loss of b in the second period. In the second period, the effect of λ is irrelevant, and thus the leader would take a myopically best action, provided that $h + N_2 < 0$.

Second, extending the term may have opposing effects.

Remark 2. Consider an extension from two to three periods. On the one hand, not to lose the benefit b of being in office, a leader may go from $(a_1, a_2) = (0, 1)$ to $(a_1, a_2, a_3) = (0, 0, 1)$. On the other hand, the leader may have an incentive to undermine the institution earlier since now she can reap the benefits from abusing longer. Thus, the leader may go from $(a_1, a_2) = (0, 1)$ to $(a_1, a_2, a_3) = (1, 1, 1)$.

Thus, while term limits create a natural end-of-term effect which increases the incentive for the leader to behave myopically, there is also a countervailing consideration that calls for shorter term limits. Institutional erosion might not be profitable in the short run and

[32] This is consistent with, for instance, Ferraz and Finan (2011), who empirically study the effect of a term limit on corruption through an anti-corruption audit program by the Brazilian government.

a leader might only want to engage in it if it has enough time to reap its benefits. Thus, shortening the term limit can be a disincentive for abuse. The less the leader can erode the institutions, the longer the optimal term limits.

Furthermore, the effect that a leader may abuse her position at the end of her term may affect the dynamics characterization. For Case 2, while the institutional strength still converges towards the upper bound, for some parameters the leaders would abuse at the last period of their term, introducing a momentary reduction in the institutional strength. For Case 4, term limits generate a regression-towards-the-mean effect with respect to types and institutional strength, and the speed of convergence may become slower. Furthermore, it increases the likelihood of the convergence to the steady state of abuse. While these effects are interesting, they do not qualitatively change the conclusion of Theorem 2, and highlight the advantage of using the stationary model for our main analysis.

6.3 Endogenous Leader Types

In the main analysis, we have assumed that the type distribution on H is constant over time and, in particular, independent of the history and the institutional strength. It is natural to think that this might not be the case. For example, when the institutions are strong, the internal process of selecting a leader in a political party would favor higher types. In the opposite direction, when the institutions deteriorate significantly, those types more willing to cheat or use patronage to buy support are more likely to enter the political process or succeed at early stages and thus be more relevant, moving the distribution of types down. Thus, the weaker the institutions, the lower the probability that a potential new leader is of a higher type. In a related vein, this framework formalizes the intuition that the strategy "we go high when they go low" may only be electorally viable within the context of sufficiently strong institutions.

We denote the support of the distribution at time t by H_t. It is important to note that the endogeneity of H_t will not change the optimal response of the current leader. This implies that the only effects will be on the long-run properties of the institution. The endogeneity of H_t will give more "inertia" to the system: if institutions deteriorated from time t to $t+1$, then with an endogenous H_t, it would be more likely to continue deteriorating (and vice-versa for an improvement).[33]

If the change does not affect the support of the distributions, then Theorem 2 will continue to hold as stated. The only difference is that convergence will be faster for the cases with

[33]When λ is the replacement probability at an election as in Section 5, this may justify the assumption that, as institutions weaken, the replacement probability λ_1 decreases, as the candidates are more concentrated at lower honesty types.

absorbing regions and for Case 3 with a long-run stationary distribution, we will observe more mass on the extremes of the long-run distribution. If the support moves, then, in addition, previous parametrizations that lead to having a long-run distribution (Case 3) will instead now fall into Case 4 in which the economy gets absorbed into either the steady state with strong institutions or the steady state with weak institutions. Thus, for a given legal framework, the early realization of its leader's types will have more important long-term consequences. Historians debate to what extent individuals play an outsize role in shaping outcomes relative to broad forces. In our model, both play a role. Yet, the possibility of Case 4 suggests that the relative importance is time dependent, where individuals play particularly important roles early on.

6.4 Additional Extensions

In the Online Appendix we explore three additional extensions to the model. First, we show how the model can be extended to study richer action sets. Allowing for an extensive margin in the abuse decision could allow us to capture the idea of the slippery slope where first the transgressions are small but as they get away with it they become more egregious. We could also allow for a vector of norms and how abuse can happen in some dimensions first and in others later. For example, it might be worth undermining the oversight of the media or the courts first and only later engage in vote buying or other corrupt activities.

Second, we consider the case where formal rules can also evolve over time. With this extension we want to capture two phenomena. On one hand, the possibility that norms can be replaced by formal rules as a reaction to prior abuse. For example, after the norm on term limits was broken by President Roosevelt in 1940, the Twenty-Second Amendment to the United States Constitution was introduced in 1951 to limit a President to two terms.[34] Conversely, autocrats can change the formal rules to allow for indefinite tenure. For example, in Venezuela, Chávez managed to abolish term limits in 2009. It is worth noting that Chávez illegally used the resources of the State to accomplish his goals.

Third, we show that the main results are robust to the possibility that the benefit from being in power can also be endogenously evolving over time.

[34]There used to be no formal term limits, as Alexander Hamilton had even written in Federalist No. 69: "That magistrate is to be elected for four years; and is to be re-eligible as often as the people of the United States." Despite this, after George Washington and Thomas Jefferson served for just two terms, this became effectively the norm.

7 Conclusion

This paper provides a parsimonious model of the evolution of institutional strength and the behavior of a leader that they induce. The leader's action has a persistent effect on the behaviors of the future leaders. As demonstrated in Theorem 1, this can lead to different long-run behaviors even for institutions with the same initial level of formal rules. The evolution of norms plays a crucial role in path dependence of institutional strength. Especially, the early history of leaders may play a crucial role in determining which outcome prevails. Thus, the paper suggests the importance of conditioning on the history of past leaders in evaluating the quality of governance. This may explain why a regime change from the outside tends to fail. Theorem 1 also sheds light on the long-run effect of the selection process of institutional leaders on the dynamics of institutional strength. Our model can capture democratic backsliding, whereby institutional strength is gradually eroded. To the best of our knowledge, our paper is the first paper that formally elucidates the role of the evolution of norms on democratic backsliding.

We believe that our simple model admits many other interesting extensions for future research. The previous section has sketched some of them. For others beyond our baseline model, for instance, one may consider multiple countries in which the action of a leader in one country may affect the incentives of the leaders of the other countries. It corresponds to cross-diffusion of anti-democracies: Rydgren (2005) studies the emergence of the party family of extreme right-wing populist parties in Western Europe, beginning with the electoral breakthrough in 1984 of the French *Front National* led by Jean-Marie Le Pen.

As discussed by Haggard and Kaufman (2021a,b), democratic backsliding processes are usually accompanied by increasing polarization. As they highlight, polarization fuels democratic backsliding by pushing citizens into hostile "us versus them" binaries, making them more willing to tolerate—or even support—authoritarian measures against those they see as enemies rather than political opponents. Once a society is deeply divided, aspiring autocrats can exploit that distrust to delegitimize opposition, erode institutional checks, and justify "strong measures" that would otherwise be unthinkable. Our model is not rich enough to capture all of these elements, which we believe present an interesting avenue for future study.

Finally, the mechanisms developed in this paper extend naturally to the corporate world, where the analog to democratic backsliding is board capturing: a CEO who gradually shapes board composition until it no longer provides meaningful oversight. As Schein (2017) emphasizes the outsized role of early leadership in setting organizational culture, our model formalizes how an executive's choices to respect or exploit governance structures have persistent effects on the norms that constrain successors. These patterns are consistent with

Bertrand and Schoar (2003), who document that individual managers' characteristics persistently affect corporate behavior and performance, and with Graham et al. (2020), who show that the sudden death of an entrenched CEO generates a roughly 3% higher abnormal return—a market signal of the rents that entrenchment creates.[35]

The corporate setting is also a promising venue for empirical investigation of our model's predictions. Unlike political institutions, the corporate world offers a panel structure with frequent CEO transitions and, in some cases, plausibly exogenous turnover events, allowing CEO fixed effects to isolate the contribution of norm evolution to governance outcomes. Together, the corporate governance literature and our theoretical framework offer complementary perspectives on how institutional norms shape long-run organizational outcomes.

A Proofs

A.1 Proof for Section 2

Proof of Lemma 1. 1. First, $(N_t)_t$ is bounded below from $(N_t^0)_t$ given by $N_1^0 = \overline{N}$ and $N_{t+1}^0 = (1-\delta)N_t^0 + \delta\overline{N} + \gamma$. Solving this recursive equation, we obtain:

$$N_t^0 = N_H + (1-\delta)^{t-1}\left(N_1^0 - N_H\right) = N_H - \frac{\gamma}{\delta}(1-\delta)^{t-1},$$

where, as in the main text, $N_H := \overline{N} + \frac{\gamma}{\delta}$. Hence, $N_t \leq N_t^0 < N_H$, where the second inequality follows because $\delta \in (0,1)$ and $\gamma > 0$.

Second, $(N_t)_t$ is bounded above from $(N_t^1)_t$ given by $N_1^1 = \overline{N}$ and $N_{t+1}^1 = (1-\delta)N_t^1 + \delta\overline{N} - \gamma$. Solving this recursive equation, we obtain:

$$N_t^1 = N_L + (1-\delta)^{t-1}\left(N_1^1 - N_L\right) = N_L + \frac{\gamma}{\delta}(1-\delta)^{t-1},$$

where, as in the main text, $N_L := \overline{N} - \frac{\gamma}{\delta}$. Hence, $N_L < N_t^1 \leq N_t$, where the first inequality follows because $\delta \in (0,1)$ and $\gamma > 0$.

2. Suppose $a_t = 0$. Then, Expression (1) reduces to

$$N_{t+1} = (1-\delta)N_t + \delta\overline{N} + \gamma = (1-\delta)N_t + \delta N_H = N_t + \delta\left(N_H - N_t\right) > N_t,$$

where the last inequality follows because $N_t < N_H$ by Part 1 of this remark.

[35] On a related point, Syverson (2004) and Hsieh and Klenow (2009) report persistent performance differences among seemingly similar enterprises.

3. Suppose $a_t = 1$. Then, Expression (1) reduces to

$$N_{t+1} = (1-\delta)N_t + \delta\overline{N} - \gamma = (1-\delta)N_t + \delta N_L = N_t - \delta\left(N_t - N_L\right) < N_t,$$

where the last inequality follows because $N_t > N_L$ by Part 1 of this remark.

□

A.2 Proofs for Section 3.2

Proof of Theorem 2. 1. In each period, h falls into $\left[\underline{h}, \tilde{h}\left(N_H\right)\right)$ with positive probability and the institutional strength decreases. Also, there is a threshold institutional strength N_* below which N_t deterministically converges to N_L. Hence, $N_t \to N_L$ almost surely along any path.

2. In each period, h falls into $\left(\tilde{h}\left(N_L\right), \overline{h}\right]$ with positive probability and the institutional strength increases. Also, there is a threshold institutional strength N^* above which N_t deterministically converges to N_H. Thus, $N_t \to N_H$ almost surely along any path.

3. For each t and for any $N_t \in (N_L, N_H)$, we have $N_{t+1} = (1-\delta)N_t + \delta\overline{N} + \gamma$ with strictly positive probability and $N_{t+1} = (1-\delta)N_t + \delta\overline{N} - \gamma$ with strictly positive probability. Thus, a limit distribution exists and has full support.

4. There is N_* such that if $N_t \leq N_*$ for some t then N_t deterministically converges to N_L. Likewise, there is N^* such that if $N_t \geq N^*$ for some t then N_t deterministically converges to N_H. In each period t, if $N_t \in (N_*, N^*)$, then with positive probability, either N_t decreases over time and is below N_* in some finite time or N_t increases over time and is above N^* in some finite time. Thus, the measure of paths $(N_t)_t$ such that $N_t \in (N_*, N^*)$ for infinitely many t is zero. This establishes the statement.

□

Proof of Corollary 1. Assume $(\delta, \gamma) = (1, 0)$. First, if $\tilde{h}(\overline{N}) \geq \overline{h}$, then, almost surely along any path, the optimal action sequence is always to abuse, i.e., Case 1 obtains. Note that if $\tilde{h}(\overline{N}) > \overline{h}$, then the optimal action sequence is deterministically always to abuse. Second, if $\tilde{h}(\overline{N}) \leq \underline{h}$, then, almost surely along any path, the optimal action sequence is always to abide by the rules, i.e., Case 2 obtains. Note that if $\tilde{h}(\overline{N}) < \underline{h}$, then the optimal action sequence is deterministically always to abide by the rules. Third, if $\tilde{h}(\overline{N}) \in (\underline{h}, \overline{h})$, then there exists a limit distribution on the set of action sequences, i.e., Case 3 obtains. The proof is complete, as these cases are exhaustive. □

A.3 Proof for Section 5

Proof of Proposition 1. 1. Recalling that λ_0 is given by Expression (8), it is enough to show that $(1 - m_0)(1 - \eta(N_t))$ is non-increasing in N_t. However, this assertion follows because m_0 is a constant and η is non-decreasing in N_t.

2. Recalling that λ_1 is given by Expression (9), it is enough to show that $(1 - m_0 - c(N_t))\eta(N_t)$ is non-decreasing in N_t. Since c is non-increasing in N_t, $(1 - m_0) - c$ is non-decreasing in N_t. Also, η is non-decreasing in N_t. Since $(1 - m_0 - c(N_t)) > 0$ and $\eta(N_t) > 0$, it follows that $(1 - m_0 - c(N_t))\eta(N_t)$ is non-decreasing in N_t.

□

References

Acemoglu, D., G. Egorov, and K. Sonin (2021): "Institutional Change and Institutional Persistence," in *Handbook of Historical Economics*, Elsevier, 365–389.

Acemoglu, D. and M. O. Jackson (2015): "History, Expectations, and Leadership in the Evolution of Social Norms," *Review of Economic Studies*, 82, 423–456.

Acemoglu, D., S. Johnson, and J. A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369–1401.

Acemoglu, D., S. Johnson, J. A. Robinson, and P. Yared (2008): "Income and Democracy," *American Economic Review*, 98, 808–842.

Acemoglu, D. and J. A. Robinson (2008): "Persistence of Power, Elites, and Institutions," *American Economic Review*, 98, 267–293.

Acemoglu, D., T. Verdier, and J. A. Robinson (2004): "Kleptocracy and Divide-and-Rule: A Model of Personal Rule," *Journal of the European Economic Association*, 2, 162–192.

Ahmed, A. (2022): "A Theory of Constitutional Norms," *Michigan Law Review*, 120, 1361–1418.

Almond, G. and S. Verba (1963): *The Civic Culture: Political Attitudes and Democracy in Five Nations*, Princeton University Press.

Almond, G. A. (1956): "Comparative Political Systems," *Journal of Politics*, 18, 391–409.

Alston, L. J. and A. A. Gallo (2010): "Electoral Fraud, the Rise of Peron and Demise of Checks and Balances in Argentina," *Explorations in Economic History*, 47, 179–197.

Andvig, J. C. and K. O. Moene (1990): "How Corruption May Corrupt," *Journal of Economic Behavior and Organization*, 13, 63–76.

Ashforth, B. E. and V. Anand (2003): "The Normalization of Corruption in Organizations," *Research in Organizational Behavior*, 25, 1–52.

Azari, J. R. and J. K. Smith (2012): "Unwritten Rules: Informal Institutions in Established Democracies," *Perspectives on Politics*, 10, 37–55.

Bertrand, M. and A. Schoar (2003): "Managing with Style: The Effect of Managers on Firm Policies," *Quarterly Journal of Economics*, 118, 1169–1208.

Besley, T. and T. Persson (2019): "Democratic Values and Institutions," *American Economic Review: Insights*, 1, 59–76.

——— (2024): "Organizational Dynamics: Culture, Design,and Performance," *Journal of Law, Economics, and Organizations*, 40, 394–415.

Besley, T. and A. Prat (2006): "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, 96, 720–736.

Bidner, C. and P. Francois (2013): "The Emergence of Political Accountability," *Quarterly Journal of Economics*, 128, 1397–1448.

Bidner, C., P. Francois, and F. Trebbi (2015): "A Theory of Minimalist Democracy," Working paper.

Biggerstaff, L., D. C. Cicero, and A. Puckett (2015): "Suspect CEOs, Unethical Culture, and Corporate Misbehavior," *Journal of Financial Economics*, 117, 98–121.

Brancati, D. (2014): "Democratic Authoritarianism: Origins and Effects," *Annual Review of Political Science*, 17, 313–326.

Bryce, J. (1888 [1995]): *The American Commonwealth*, vol. I, Liberty Fund.

Coppedge, M., J. Gerring, C. H. Knutsen, S. I. Lindberg, J. Teorell, D. Altman, F. Angiolillo, M. Bernhard, A. Cornell, M. S. Fish, L. Fox, L. Gastaldi, H. Gjerløw, A. Glynn, A. Good God, A. Hicken, K. Kinzelbach, K. L. Marquardt, K. McMann, V. Mechkova, A. Neundorf, P. Paxton,

D. Pemstein, J. Pernes, J. von Römer, B. Seim, R. Sigman, S.-E. Skaaning, J. Staton, A. Sundström, M. Tannenberg, E. Tzelgov, Y.-t. Wang, T. Wig, and D. Ziblatt (2026): "V-Dem Codebook v16," Varieties of Democracy (V-Dem) Project.

Currie, T., P. Turchin, J. Bednar, P. J. Richerson, G. Schwesinger, S. Steinmo, R. Wacziarg, and J. Wallis (2016): "Evolution of Institutionsand Organizations," in *Complexity and Evolution: Toward a New Synthesis for Economics*, ed. by D. S. Wilson S. and A. Kirman, MIT Press, 201–236.

Dessein, W. and A. Prat (2022): "Organizational Capital, Corporate Leadership, and Firm Dynamics," *Journal of Political Economy*, 130, 1477–1536.

Diamond, L. (2021): "Democratic Regression in Comparative Perspective: Scope, Methods, and Causes," *Democratization*, 28, 22–42.

Diamond, L. J. (1999): *Developing Democracy: Towards Consolidation*, Johns Hopkins University Press.

Dixit, A., G. M. Grossman, and F. Gul (2000): "The Dynamics of Political Compromise," *Journal of Political Economy*, 108, 531–568.

Fearon, J. D. (2011): "Self-Enforcing Democracy," *Quarterly Journal of Economics*, 126, 1661–1708.

Ferraz, C. and F. Finan (2010): "Motivating Politicians: The Impacts of Monetary Incentives on Quality and Performance," Working paper.

——— (2011): "Electoral Accountability and Corruption: Evidence from the Audits of Local Governments," *American Economic Review*, 101, 1274–1311.

Foran, C. (2016): "An Erosion of Democratic Norms in America," *Atlantic*, November 22.

Gagliarducci, S. and T. Nannicini (2013): "Do Better Paid Politicians Perform Better? Disentangling Incentives From Selection," *Journal of the European Economic Association*, 11, 369–398.

Glaeser, E. L., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2004): "Do Institutions Cause Growth?" *Journal of Economic Growth*, 9, 271–303.

Graham, J. R., H. Kim, and M. Leary (2020): "CEO-board Dynamics," *Journal of Financial Economics*, 137, 612–636.

Gratton, G. and B. E. Lee (2024): "Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies," *Review of Economic Studies*, 91, 340–371.

Grillo, E., Z. Luo, M. Nalepa, and C. Prato (2024): "Theories of Democratic Backsliding," *Annual Review of Political Science*, 27, 381–400.

Grillo, E. and C. Prato (2023): "Reference Points and Democratic Backsliding," *American Journal of Political Science*, 67, 71–88.

Guiso, L., P. Sapienza, and L. Zingales (2015): "The Value of Corporate Culture," *Journal of Financial Economics*, 117, 60–76.

Guiso, L., L. Zingales, and P. Sapienza (2016): "Long-term Persistence," *Journal of the European Economic Association*, 14, 1401–1436.

Guriev, S. and D. Treisman (2020): "A Theory of Informational Autocracy," *Journal of Public Economics*, 186, 104158.

Haggard, S. and R. Kaufman (2021a): "The Anatomy of Backsliding: Why is Democracy Consuming Itself?" *Political Violence At A Glance*, March 3.

——— (2021b): *Backsliding: Democratic Regress in the Contemporary World*, Cambridge University Press.

Helmke, G., M. Krorger, and J. Paine (2022): "Democracy by Deterrence: Norms, Constitutions, and Electoral Tilting," *American Journal of Political Science*, 66, 434–450.

Howell, W. G., K. A. Shepsle, and S. Wolton (2023): "Executive Absolutism: The Dynamics of Authority Acquisition in a System of Separated Powers," *Qarterly Journal of Political Science*, 18, 243–275.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *Quartrely Journal of Economics*, 124, 1403–1448.

Huq, A. and T. Ginsburg (2018): "How to Lose a Constitutional Democracy," *U.C.L.A. Law Journal*, 65, 78–169.

Invernizzi, G. M. and M. M. Ting (2024): "Institutions and Political Restraint," *American Journal of Political Science*, 68, 58–71.

Jones, B. F. and B. A. Olken (2005): "Do Leaders Matter? National Leadership and Growth Since World War II," *Quarterly Journal of Economics*, 120, 835–864.

Kagan, R. (2023): "A Trump Dictatorship is Increasingly Inevitable. We should Stop Pretending," *Washington Post*, November 30.

Kamarck, E. (2021): "Did Trump Damage American Democracy?" *Brookings Institution*, July 9.

Kapstein, E. B. and N. Converse (2008): *The Fate of Young Democracies*, Cambridge University Press.

Kartik, N., E. Lipnowski, and H. Pei (2025): "Replacement and Reputation," Working paper.

Keefer, P. (2007): "Clientelism, Credibility, and the Policy Choices of Young Democracies," *American Journal of Political Science*, 51, 804–821.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. Vishny (1999): "The Quality of Government," *Journal of Law, Economics, and Organization*, 15, 222–279.

Levitsky, S. and L. Way (2015): "The Myth of Democratic Recession," *Journal of Democracy*, 26, 45–58.

Levitsky, S. and D. Ziblatt (2018): *How Democracies Die*, Crown Publishing.

Linz, J. J. (1978): *The Breakdown of Democratic Regimes: Crisis, Breakdown and Reequilibration*, Johns Hopkins University Press.

——— (1990): "Transitions to Democracy," *Washington Quarterly*, 13, 143–164.

Luo, Z. and A. Przeworski (2023): "Democracy and its Vulnerabilities: Dynamics of Democratic Backsliding," *Quarterly Journal of Political Science*, 18, 105–130.

Lust, E. and D. Waldner (2015): "Unwelcome Change: Understanding, Evaluating, and Extending Theories of Democratic Backsliding," U.S. Agency for International Development.

Mainwaring, S. and F. Bizzarro (2019): "The Fates of Third-Wave Democracies," *Journal of Democracy*, 30, 99–113.

Myerson, R. B. (2006): "Federalism and Incentives for Success of Democracy," *Quarterly Journal of Political Science*, 1, 3–23.

——— (2011): "Toward a Theory of Leadership and State Building," *Proceedings of the National Academy of Sciences*, 108, 21297–21301.

Nord, M., D. Altman, T. Fernandes, A. Good God, and S. I. Lindberg (2026): "Democracy Report 2026: Unraveling The Democratic Era?" University of Gothenburg: V-Dem Institute.

North, D. C. (1990): *Institutions, Institutional Change, and Economic Performance*, Cambridge University Press.

O'Donnell, G. and P. C. Schmitter (1986): *Transitions from Authoritarian Rule: Tentative Conclusions about Uncertain Democracies*, Johns Hopkins University Press.

O'Donnell, G. A. (1996): "Illusions about Consolidation," *Journal of Democracy*, 7, 34–51.

Paldman, M. (2002): "The Cross-Country Pattern of Corruption: Economics, Culture and the Seesaw Dynamics," *European Journal of Political Economy*, 18, 215–240.

Persson, T. and G. Tabellini (2009): "Democratic Capital: The Nexus of Political and Economic Change," *American Economic Journal: Macroeconomics*, 1, 88–126.

Pfiffner, J. P. (2021): "Donald Trump and the Norms of the Presidency," *Presidential Studies Quarterly*, 51, 96–124.

Pierson, P. (2000): "Increasing Returns, Path Dependence, and the Study of Politics," *American Political Science Review*, 94, 251–267.

Prat, A. and D. Strömberg (2013): "The Political Economy of Mass Media," in *Advances in Economics and Econometrics: Tenth World Congress*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, 135–187.

Putnam, R. (1993): *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press.

Renan, D. (2018): "Presidential Norms and Article II," *Harvard Law Review*, 131, 2187–2282.

Rydgren, J. (2005): "Is Extreme Right-Wing Populism Contagious? Explaining the Emergence of a New Party Family," *European Journal of Political Research*, 44, 413–437.

Schein, E. H. (2017): *Organizational Culture and Leadership*, John Wiley & Sons, 5th ed.

Shleifer, A. and R. W. Vishny (1993): "Corruption," *Quarterly Journal of Economics*, 108, 599–617.

Strömberg, D. (2015): "Media and Politics," *Annual Review of Economics*, 7, 173–205.

Svolik, M. W. (2013): "Learning to Love Democracy: Electoral Accountability and the Success of Democracy," *American Journal of Political Science*, 57, 685–702.

Syverson, C. (2004): "Product Substitutability and Productivity Dispersion," *Review of Economics and Statistics*, 86, 534–550.

Szelényi, Z. (2022): "Viktor Orbán's Machiavellian Genius: Will His Illiberal Experiment Return to Haunt Him?" *UnHerd*.

Tanzi, V. (1998): "Corruption Around the World: Causes, Consequences, Scope, and Cures," *IMF Staff Papers*, 45, 559–594.

Wantchekon, L. (2013): "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin," *World Politics*, 55, 399–422.

Yuchtman, N. (2024): "Fight Fire with Fire: the Erosion of Israel's Democracy and the Popular Response," in *The Transition to Illiberal Democracy Economic Drivers and Consequences*, ed. by A. Razin, CEPR Press, 35–42.