Quarterly Journal of Political Science, 2023, 18: 105–130

Democracy and its Vulnerabilities: Dynamics of Democratic Backsliding

Zhaotian Luo[1] and Adam Przeworski[2*]

[1] *University of Chicago, Chicago, IL, USA; luozhaotian@uchicago.edu*
[2] *New York University, New York, NY, USA; adam.przeworski@nyu.edu*

ABSTRACT

The puzzle entailed in erosion of democracy by backsliding — a process in which the incumbent government takes every opportunity to reduce citizens' ability to remove it by democratic means — is how a catastrophic situation can be gradually brought about by steps against which people who would be adversely affected do not react in time. We investigate conditions which render democracy impregnable to backsliding and conditions which make it vulnerable. Democracy is sustainable, free from the threat of backsliding, when opposing politicians are neither very attractive nor very unattractive to citizens. To sustain it, citizens must allow more appealing incumbents to gain some security in office. Backsliding occurs either when citizens knowingly consent to erosion of democracy because they find the incumbent highly appealing or when citizens unconditionally oppose the incumbent, so that the incumbent can remain in office only by backsliding.

Keywords: Democracy; backsliding; dynamics; sustainability

*For comments we thank John Ferejohn, Joanne Fox-Przeworski, Diego Gambetta, Roberto Gargarella, Robert Gulotty, Steven Holmes, Monika Nalepa, Andrew Little, Beatriz Magaloni, Bernard Manin, Jack Paine, Pasquale Pasquino, Arturas Rozenas, Kenneth Shotts, Susan Stokes, Milan Svolik, Andrea Vindigni, and two anonymous reviewers.

Online Appendix available from:
http://dx.doi.org/10.1561/100.00021112_app
MS submitted on 25 June 2021; final version received 18 January 2022
ISSN 1554-0626; DOI 10.1561/100.00021112
© 2023 Z. Luo and A. Przeworski

Most, if not all, democracies were established as a reaction against "despotic," "tyrannical," or "autocratic" rule. Their institutions were designed to prevent incumbents from holding onto office independently of popular will and from adopting measures that would curtail individual freedoms. The resulting institutional arrangements have varied but the goal everywhere was to establish a system in which each part of the government would want to and have the means to prevent usurpation of power by any other part. The father of constitutionalism, Montesquieu (1748, p. 326), insisted that "For the abuse of power to be impossible, it is necessary that by the disposition of things, the power stop the power." Or, in an often-cited Madison's passage (Madison, 1788), "the great security against a gradual concentration of several powers in the same department consists in giving to those who administer each department the necessary constitutional means and personal motives to resist encroachments of the others. . . . Ambition must be made to counteract ambition." The effect of the separation of powers would be "limited" or "moderate" government.[1]

Not everyone was confident that institutional checks would be sufficient to maintain the balance of powers.[2] But if these internal controls were to fail, if governments were to commit flagrantly unconstitutional acts, people would rise in a revolution aimed at restoring the constitutional status quo. Montesquieu (1748, p. 19) thought that if any power succeeded to violate fundamental laws, "everything would unite against it"; there would be a revolution, "which would not change the form of government or its constitution: for revolutions shaped by liberty are but a confirmation of liberty." In this tradition, Weingast (1997, 2015) argued that if a government were to conspicuously violate the constitution, cross a "bright line," citizens would coordinate against it and, anticipating this reaction, the government would not commit such violations. Fearon (2011) thought that the same would occur if a government were not to hold an election or commit flagrant fraud. Hence, the combination of internal and external controls would make democratic institutions impregnable to the "encroaching spirit of power" (Madison, 1788), the desire of politicians for enduring and unlimited power.

[1] For models in which separation of powers, treated as unitary actors, generates moderate equilibriums see Persson *et al.* (2001) and Dragu *et al.* (2014).

[2] As Palmer (1959, p. 262) observed,

> The real problem (and it was a real problem) was to prevent the powers thus constituted from usurping more authority than they have been granted. According to one school, the several constituted powers of government, by watching and balancing and checking one another, were to prevent such usurpation. According to another school, which regarded the first school as undemocratic or mistrustful of the people, the people itself must maintain a constant vigilance and restraint upon the powers of government.

This is the view of democracy we inherited and this is the view we are now forced to question. By now we have seen Turkey under the government of AKP, Venezuela under Chavez and Maduro, Hungary under the second government of Fidesz, Poland under the second government of PiS, India under Narendra Modi, as well as the United States under Donald Trump. All these are, albeit to different extent, instances of democratic "backsliding" (or "deconsolidation," "erosion," "retrogression"): "a process of incremental (but ultimately still substantial) decay in the three basic predicates of democracy — competitive elections, liberal rights to speech and association, and the rule of law" (Ginsburg and Huq, 2018, p. 17). As this process advances, the opposition becomes unable to win elections or assume office if it wins, established institutions lose the capacity to control the executive, while manifestations of popular protest are repressed by force.

The puzzle entailed in destruction of democracy by backsliding is how a catastrophic situation can be gradually brought about by steps against which people who would be adversely affected do not react in time. As Ginsburg and Huq (2018, p. 91) pose it, "The key to understanding democratic erosion is to see how discrete measures, which either in isolation or in the abstract might be justified as consistent with democratic norms, can nevertheless be deployed as mechanisms to unravel liberal constitutional democracy."

We think as follows. A government wins an election.[3] This government is characterized by its attractiveness to citizens. Its appeal may be based on some policy outcomes, such as income growth, but it may be also be rooted in ideological affinities, such as Islamization in Turkey, "Bolivarianism" in Venezuela, "preserving the purity of the nation" in Hungary, "defending Christianity" in Poland, Hinduism in India, or xenophobia in the United States. The government decides whether to take steps to protect its tenure in office from any opposition, steps such as changing electoral formulae, redistricting, changing voting qualifications, harassing the partisan opposition, imposing restrictions on NGOs, reducing judicial independence, imposing partisan control over the state apparatuses, or controlling the media.[4] If the government takes such steps, citizens face a trade-off between enjoying the outcomes the government generates and preserving their ability to remove the incumbent in

[3]The definition of "victory" is not as simple as it may appear. Electoral laws play an important role: in Turkey, the AKP won 34.3 percent of votes to obtain 66.0 percent of seats when it first assumed power in 2002, in Hungary Fidesz won 53 percent of votes and 68 of seats in 2010, in Poland PiS got 37.5 percent of votes and 51.0 of seats, in the 2016 election in the United States Donald Trump won with 46.09 percent of popular vote against 48.18 for his opponent. The ascension of Chavez to office in Venezuela was convoluted: the traditional parties actually won the legislative election of 1998, Chavez won the presidential election with 56.4 percent, the referendum for a new constitution was passed by 71.8 percent, then Chavez won a new election with 59.8 percent while his party obtained 44.4 percent of votes and 55.7 of seats in the legislative election of 2000.

[4]On the role of media, see Li *et al.* (2020).

some future when they would prefer a challenger. Yet whether the incumbent can be removed from office depends on how far this process has already advanced, on how much advantage the incumbent has already mustered.

The obvious question is what makes democracy vulnerable to backsliding. What are the different ways in which backsliding occurs? What may induce governments to take actions that limit the ability of the opposition to remove it? Once a government begins to backslide, can it be ever stopped short of realization of complete domination? Would the potential opposition be able to remove a backsliding government? Are there some conditions under which democracy is impregnable to backsliding?

As noted by Grillo and Prato (2019), the burgeoning literature on backsliding differs in two dimensions: the motivation of the incumbent, which can be either to increase policy discretion or to extend tenure in office, and the identity of the restraining agent, which can be either "horizontal" (other institutions, parties, elites) or "vertical" (citizens, voters) (Table 1).[5] The distinction between seeking policy discretion and extending tenure in office is not perfectly sharp in that the incumbent may need policy discretion in order to advance his electoral advantage: Viktor Orbán, for example, centralized the revenue Hungarian cities receive from parking violations, in order to reduce the budget of the Budapest municipal government, controlled by the opposition. As distinct from Howell and Wolton (2018) or Howell *et al.* (2019), we model the process by which the government accumulates electoral advantage, not the process by which the incumbent weakens horizontal checks on his policy discretion. This distinction is important because the erosion of institutional constraints is a lasting legacy, while partisan advantage of an incumbent is not inherited by a partisan opponent.

While our model fits into tenure–voter cell, as distinct from static models (Graham and Svolik, 2020; Grillo and Prato, 2019; Nalepa *et al.*, 2020), we

Table 1: Literature on democratic backsliding.

	Restrainer	
Motivation	Horizontal	Vertical
Discretion	Howell and Wolton (2018)	Graham and Svolik (2020)
	Howell *et al.* (2019)	Grillo and Prato (2019)
Tenure	Helmke *et al.* (2021)	Nalepa *et al.* (2020)
		This paper

[5]Miller (2021) has a setup in which both the opposition elites and citizens play roles in restraining the incumbent: the opposition elites can mobilize citizens and provide information while citizens can vote against the incumbent.

treat the value of democracy as evolving endogenously. For us, the value of democracy is the ability to remove any incumbent from office by democratic procedures when citizens believe that a different government would be better for them, and this value evolves endogenously as a consequence of actions taken by the incumbent rather than being fixed once and for all.

In what follows, we first preview the assumptions of the model and summarize the central results in nontechnical terms. The model is presented next. A conclusion closes the paper. Most proofs are in the Online Appendix.

Democracy and its Vulnerabilities

Democracy is an institutional arrangement in which citizens are able to replace incumbents through elections whenever they believe that a different government would be better for them. Ideally, the outcomes of democratic procedures by which citizens select their rulers would depend only on citizens' comparisons of the current and the prospective governments: which would make them better off, whether materially or ideologically? People want to be governed by politicians they find appealing, so that even if they do not care about institutions per se, they value being able to decide who would govern them. In turn, because citizens would try to remove the incumbent government from office when they find a competitor more attractive, incumbents are insecure in office, which may encourage some of them to take steps to protect their tenure from the voice of the people. We refer to these incumbents as "*authoritarian-minded leaders*." Yet even under democracy, citizens may have to allow incumbents to gain some security in office. This reward must be sufficient for the incumbent government not want to push its domination too far, while still being tolerable for citizens.

Democracy is vulnerable when the incumbent government is either very appealing or very unattractive to citizens, relatively to their expectations about potential challengers. When people are highly satisfied with the current government, they see it as unlikely that a competitor would be better, so they retain the incumbent, which in turn makes the leader free to take steps that increase his chances to remain in office. When the incumbent government is relatively unappealing, citizens are afraid that it would hold onto office, and to preempt this possibility they want to remove the incumbent even if they see the current challenger as less attractive. In turn, because the incumbent knows that citizens want to remove it no matter what, it does everything possible to solidify its hold on office. Hence, in both situations the incumbent engages in *backsliding*, that is, takes every opportunity to advance its partisan interest, all the way to the level at which it risks being removed by nondemocratic means, such as coups or popular uprisings.

To flesh out intuitively the logic of this analysis, consider the following model, presented formally below. The authoritarian-minded leader ("leader, he") generates some fixed level of satisfaction to a representative citizen ("citizen, she").[6] This level of satisfaction may be due to some policies of the incumbent, to manipulation of beliefs, or just to ideological congruence of values. At each time the leader faces a challenger whom the citizen finds more or less attractive, better or worse, than the incumbent. The probability that any future challenger would be better is given. The leader's "*advantage*" is the probability that he remains in office when the citizen wants to remove him: this definition encompasses situations in which the representative citizen is not decisive as well as those in which the incumbent loses an election and manages to stay in office. This probability, in turn, depends on the actions of the leader: in each period the leader faces an opportunity to take a step of varying magnitude to increase his advantage, a step which he may or may not take. Citizens observe whether some acts increase incumbent advantage and by how much only when their effects materialize. Imagine that a government extends voting rights to citizens residing abroad, or adopts legislation to require additional documentation at the polling place, or relaxes the rules regulating private political financing, all while offering democratic arguments in their favor: "We want to extend rights to all citizens," "We want to prevent fraud," "We are protecting the freedom of expression." Only ex post we learn that Erdogan won an election by the vote of Turks in Berlin, that Republicans won because poor people who did not have the required documents were prevented from voting, or that the Indian Bharatiya Janata Party (BJP) enjoyed a massive financial advantage in the election.[7] In turn, at each time, the citizen decides whether to retain the incumbent, having observed the quality[8] of the challenger but not knowing how large is the opportunity of the leader to increase his advantage if he is not removed.

These assumptions are sufficient to identify the conditions under which democracy is sustainable and when it is vulnerable to two situations which induce the leader to engage in backsliding, either with the support of the citizen or against her opposition. Figure 1 illustrates and compares the average survival rate (1− hazard rate) of the leader in all the three cases.

[6]To the extent to which the leader may take actions that affect the level of satisfaction of the citizen, perhaps the best way to think is that the flow payoff represents the highest level of appeal the leader can achieve.

[7]This delay in observing the effect corresponds to the notion of "stealth" as in Varol (2015).

[8]We use "quality" interchangeably with "attractiveness" and "appeal" to indicate the extent to which the citizen likes a particular candidate. These terms are not intended as an objective measure of a candidate's ability; they just describe the citizen's subjective preference.

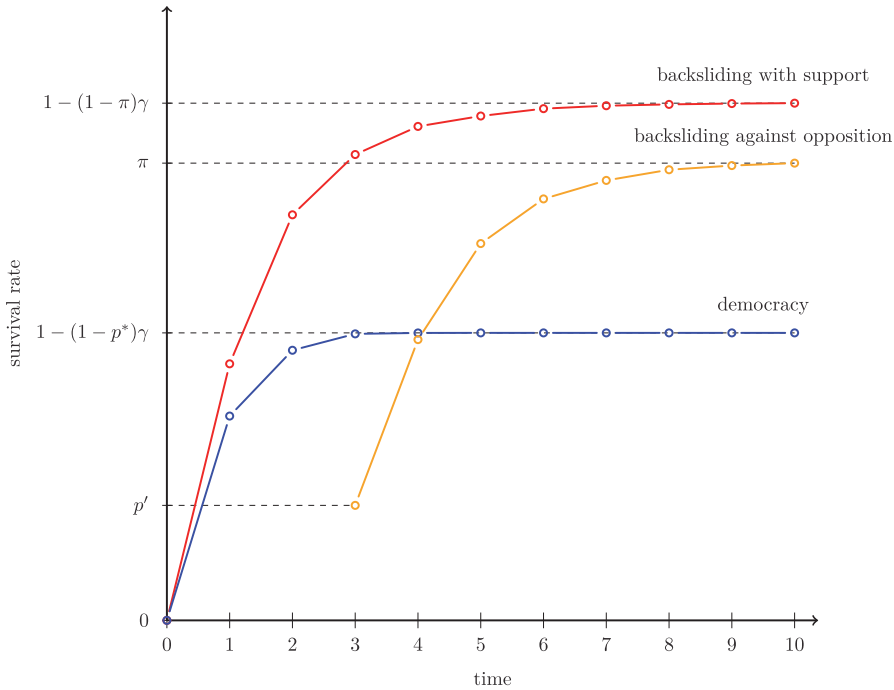

Figure 1: Survival rates.

Average survival rate (1 − hazardrate) from 1000 runs of the model presented below.
Red: backsliding with support; yellow: backsliding against opposition starting at p'; blue: democracy with the threshold of p^*.
Parameters: $\pi = 0.9$, $F(q) = \frac{q}{\pi}$, $\gamma = 0.5$, $p' = 0.6$, $p^* = 0.5$.

Democracy is sustainable, free from the danger of backsliding, only if there exists a threshold level of advantage that is simultaneously satisfactory to the incumbent, so that the leader increases his advantage up to this level but stops short of complete domination, and acceptable to the citizen, so that the citizen supports the incumbent against worse challengers as long as the threshold level is not passed and credibly threatens to oppose unconditionally if it is crossed. If such a threshold level exists, the citizen would keep some ability to remove the leader when a better challenger arises and as a result, the leader's long-term survival rate would be capped at a relatively fair level shown by the blue curve in Figure 1. Such a threshold level does exist, so that democracy is sustainable, when incumbents are neither very attractive or very unattractive to the citizen in comparison to potential challengers. But citizens face a trade-off between the attractiveness of the competing candidates and

the ability to replace the incumbent. To sustain democracy, citizens must be willing to tolerate a higher level of advantage for more attractive politicians. Given the attractiveness of the incumbent and the average quality of the potential challengers, the citizen is better off when the leader has a lower advantage. But, perhaps to the dismay of democratic purists, because citizens are better off when the quality of all politicians — the incumbent and the challengers — is higher, they are best off in a sustainable democracy in which they allow the incumbent a high level of security in office.

Democracy is vulnerable in two situations.

Whenever the citizen finds the leader highly appealing, the unique equilibrium is one in which the leader backslides with the support of the citizen. The citizen retains the leader whenever she finds him more attractive than the challenger and the leader takes every opportunity to undermine democracy. Because the leader is highly attractive, the citizen believes that any challenger is unlikely to be better, so that she values relatively little her future ability to remove the incumbent. In turn, because the citizen wants to retain an attractive incumbent, the leader feels *free* to use all opportunities to increase his advantage.

In contrast, a situation may emerge in which the citizen wants to remove the incumbent regardless of the attractiveness of the challenger and the leader undermines democracy as long as he remains in office. Backsliding against opposition arises when the citizen finds the leader relatively unappealing and the leader has already managed to achieve considerable advantage, above the threshold which sustains democracy. Because the leader is unattractive, the citizen fears that he would further increase his advantage, so that she wants to remove him even when she finds the challenger less attractive. Because the citizen unconditionally opposes him, the leader finds it *necessary* to use all opportunities to accumulate advantage.

In both cases, the ability of the citizen to remove the incumbent declines and the citizen becomes worse off as backsliding proceeds, yet the citizen does not react in time. When the leader is very attractive, the citizen values retaining the leader so highly that she wants him to remain in office even at the cost of being unable to remove him in the unlikely possibility that a better challenger would appear. When, however, the incumbent is less attractive, the citizen would want to unconditionally remove the incumbent if he would cross some tolerable level of advantage but she cannot react in time because she does not immediately observe the consequences of incumbents' actions. Hence, the answer to the puzzle differs across situations.

With support or against opposition, once the leader begins backsliding, he takes every opportunity to increase his advantage, going all the way to the level at which citizens lose all hope of being able to remove him by democratic

means and conflicts spill outside institutional boundaries.[9] Because under backsliding with support the citizen wants to retain the leader as long as the challenger is not better while under backsliding against opposition she wants to remove him unconditionally, the probability that the leader would complete more steps toward complete domination is higher in the first situation. Yet in both situations, unless the leader is removed early into the process, removing him by democratic means becomes difficult. This is shown by the red and yellow curves in Figure 1: conditional on surviving all past periods, the leader's probability to stay in power for yet another period is larger the longer he has remained in power.

Model

Model Setup

An *authoritarian-minded leader* ("he," and "leader" for short) and a representative *citizen* ("she") interact in infinitely many periods, $t = 0, 1, 2, \ldots$. The leader provides the citizen with the flow payoff of $x \in (0, 1)$ each period he holds office. At each t, the citizen chooses the office holder between the *incumbent*, the leader at $t = 0$ and whoever holds office at $t - 1$ for any $t \geq 1$, and a randomly drawn *challenger* with two possible types $y_t = 0, 1$.[10] A *high-type* challenger brings the citizen the flow payoff of $y_t = 1 > x$ each period in office while a *low-type* challenger brings $y_t = 0 < x$. A high-type challenger is drawn at each t with a fixed probability $Pr(y_t = 1) = \gamma \in (0, 1)$. The leader loses office forever once being removed.[11]

The leader enters office under perfect democracy, in which the citizen has full control over who would be in office. The leader, however, may undermine democracy, securing his position in office against potential opposition by the citizen. Formally, at any t when the leader is the incumbent, he remains in office with probability p_t. This probability measures the leader's ability to hold on power against the citizen's will and therefore is referred to as his *advantage*.

[9] Evo Morales in Bolivia is a good example. He was term-limited, called for a referendum to abolish the term limits, lost it, appealed to the constitutional courts which he previously appointed, received a favorable ruling, run for office again, and declared his victory in spite of widespread allegations of fraud. The result was a popular rebellion, in which he was abandoned by the police and the armed forces.

[10] Allowing the challenger's quality to be continuously distributed in $[0, 1]$ would require the citizen to form an expectation about their quality, greatly complicating the algebra. But in the end the equilibrium would still depend on the comparison between the appeal of the incumbent and the expected attractiveness of challengers and at each time the decision of the citizen would still depend on a particular draw from the distribution. Hence, qualitative results would remain in the same.

[11] Note that we use the terms "incumbent" for short to refer to the government in office. Moreover, the "leader" is the head of the incumbent government even when it is not the same person: for example, Maduro became the "leader" upon the death of Chavez.

The leader begins with no advantage at all, $p_0 = 0$.[12] At each t when the leader is the incumbent, he gets an opportunity to take a *step* that increases his advantage from p_t to some q_t, drawn from a distribution F conditional on $q_t > p_t$. Letting $b_t = 0, 1$ denote the action of the incumbent, where $b_t = 1$ indicates taking the step,

$$p_{t+1} = (1 - b_t)p_t + b_t q_t. \tag{1}$$

The distribution F has full support on $[0, \pi]$, where $\pi < 1$ represents the largest possible advantage the leader can ever achieve.

At the beginning of any period when the incumbent is the leader, p_t is observed both by the leader and the citizen. The leader observes q_t and decides whether to take the step to increase his advantage, b_t. The citizen observes y_t and chooses between the leader and the challenger.[13] For convenience, assume that incumbents other than the leader have no advantage.

The leader cares only about staying in power while the citizen cares only about the flow payoffs office holders produce. Both are forward looking and discount future payoffs at the rate of $\delta \in (0, 1)$. Let $\ell_t = 0, 1$ denote whether the leader holds office at t and $u_t = 0, x, 1$ denote the flow payoff the office holder at t provides to the citizen. At any t, the leader seeks to maximize the average discounted expected payoff ("expected payoff" for short)

$$(1 - \delta)E\left(\sum_{\tau=t}^{\infty} \delta^{\tau-t}\ell_\tau | p_t, q_t\right), \tag{2}$$

while the citizen seeks to maximize

$$(1 - \delta)E\left(\sum_{\tau=t}^{\infty} \delta^{\tau-t}u_\tau | p_t, y_t\right). \tag{3}$$

Benchmark

To build up intuition, consider first a benchmark in which the leader has a fixed level of advantage p. The leader is not strategic and only the citizen's decision matters.

The best case scenario for the citizen is to replace the incumbent with a high-type challenger who brings the highest possible flow payoff of 1 every period. When such a challenger is available, the citizen clearly prefers the leader to be removed.

[12] The model produces the same results for any p_0 that is not too large. In fact, a larger p_0 makes democracy more robust against backsliding.

[13] b_t and q_t are not directly observable to the citizen in period t. But because the citizen observe p_{t+1}, she can infer b_t and q_t in period $t + 1$. In other words, the citizen observes the incumbent's action to increase advantage and its consequence with a delay.

The citizen's preference is less clear with a low-type challenger. In this case, the citizen faces an *intertemporal trade-off*. By keeping the leader in office, the citizen gets a higher flow payoff $x > 0$ in the current period. If a high-type challenger is drawn in the next period, however, the citizen would fail to place him into office with probability p, due to the leader's advantage. Through replacing the leader with a low-type challenger, which succeeds only with probability $1 - p$, the citizen gets a lower flow payoff $0 < x$ in the current period, but in the future she would be perfectly able to place a high-type challenger into office whenever such a challenger is available. Her continuation value, denoted as $w(\gamma)$, is identical to her expected payoff in any period in which the current incumbent is the low-type challenger from an earlier period. In such a period, with probability γ, a high-type challenger competes for office, in which case the citizen would choose the challenger and get 1 in the current and every future period. With probability $1 - \gamma$, a low-type challenger competes for office, in which case the citizen gets 0 in the current period and $w(\gamma)$ in the future. Therefore,

$$\begin{aligned} w(\gamma) &= \gamma + (1-\gamma)\delta w(\gamma) \\ &= \frac{\gamma}{1-\delta(1-\gamma)}. \end{aligned} \tag{4}$$

Proposition 1 (benchmark)**.** *Suppose the leader has a fixed advantage p. The citizen prefers to replace the leader with a low-type challenger if and only if $p > \frac{x}{\delta w(\gamma)}$. Moreover, the citizen's expected payoff in any period with the leader being the incumbent is*

$$\overline{U}(p,\gamma,x) = \begin{cases} \frac{(1-\delta)(1-(1-p)\gamma)}{1-\delta(1-(1-p)\gamma)}x + 1 - \frac{(1-\delta)(1-(1-p)\gamma)}{1-\delta(1-(1-p)\gamma)}, & p \leq \frac{x}{\delta w(\gamma)} \\ \frac{(1-\delta)p}{1-\delta p}x + \left(1 - \frac{(1-\delta)p}{1-\delta p}\right) w(\gamma), & p > \frac{x}{\delta w(\gamma)} \end{cases}, \tag{5}$$

which is strictly decreasing in p.

When the leader has a sufficiently large advantage, the citizen prefers to replace him even with a challenger who is less appealing. The threshold level of advantage equals to the ratio of what the citizen loses in the current period by removing the leader, x, and the present value of what she expects to gain in the future, $\delta w(\gamma)$.

Therefore, even if citizens do not value democracy because of some ideals, such as political equality or liberty, with which they associate it, caring only about the flow payoffs governments provide, they value the ability to choose a better office holder whenever one would be available, if not now, in the future. This value provides citizens with an incentive to remove an incumbent who deprives them of this opportunity. When democracy is threatened by an overly large advantage, this incentive is strong enough to induce citizens to remove such an incumbent as soon as possible, even if doing so means having a less appealing politician in office.

Equilibrium Concept

Now consider the full version of the model, which endogenizes the leader's advantage. The leader begins with no advantage and in each period he chooses whether to take a step that increases his advantage.

The equilibrium concept is *Markov perfect equilibrium* ("equilibrium" for short) with the state variables being the incumbent's identity and, if the incumbent is the leader, his advantage. The game is trivial in any period when the incumbent is not the leader: the citizen keeps the incumbent in office if the challenger is worse and removes the incumbent if the challenger is better. In any such period, the citizen gets the payoff of 1 if the incumbent is a high-type challenger from some earlier period and the expected payoff of $\gamma + (1-\gamma)\delta w(\gamma) = w(\gamma)$ if the incumbent is a low-type challenger from an earlier period. At any t with the leader being the incumbent, the leader's strategy maps his current level of advantage, p_t, and the level at the next period if he takes the step, q_t, into a decision $\beta(q_t, p_t) = 0, 1$ on whether to take the step. The citizen always chooses the challenger when the challenger is of the high type, $y_t = 1$. Hence, each strategy of the citizen maps the current level of the leader's advantage, p_t, into a decision $\kappa(p_t) = 0, 1$ as to whether to keep the leader in office when the challenger has the low type, $y_t = 0$.

Let $L_\kappa(p)$ denote the leader's expected payoff in any period when he is the incumbent and has the advantage of p, given the citizen's strategy κ. The leader expects to be retained in such period with probability

$$(1-\gamma)\kappa(p) + ((1-\gamma)(1-\kappa(p)) + \gamma)\, p = p + (1-p)(1-\gamma)\kappa(p).$$

Specifically, with probability $(1-\gamma)\kappa(p)$, the citizen chooses the leader, in which case the leader stays for sure. With probability $(1-\gamma)(1-\kappa(p))+\gamma$, the citizen chooses the challenger, in which case the leader stays with probability p. Given that the leader survives in office, he gets a flow payoff of 1 in the current period and his continuation value in the future is $L_\kappa(q)$ if the leader takes the step to increase his advantage to q and $L_\kappa(p)$ if he gives up taking the step. The leader would take the step if and only if

$$L_\kappa(q) > L_\kappa(p). \tag{6}$$

In turn,

$$\begin{aligned} L_\kappa(p) \;=\; & (p + (1-p)(1-\gamma)\kappa(p)) \\ & \times \left(1 - \delta + \delta \int_p^\pi \max\left\{L_\kappa(q), L_\kappa(p)\right\} dF_p(q)\right), \end{aligned} \tag{7}$$

where $F_p(q) = F(q|q > p)$.

Let $U_\beta(p)$ denote the citizen's expected payoff in any period when the leader is the incumbent with the advantage of p, given the incumbent's strategy β.

The citizen faces three possibilities. First, a high-type challenger replaces the leader, in which case the citizen gets the best possible payoff of 1. Second, a low-type challenger replaces the leader, in which case the citizen gets 0 currently and the expected payoff of $w(\gamma)$ in the future, so that her total expected payoff is $\delta w(\gamma)$. Finally, if the leader stays in power, the citizen gets x currently. Given the leader's strategy β, if $\beta(q,p)=1$, the leader's advantage in the next period would be q, in which case the citizen's continuation value would be $U_\beta(q)$; if $\beta(q,p)=0$, the leader's advantage in the next period would be p, in which case the citizen's continuation value would be $U_\beta(p)$. In expectation, the citizen gets the continuation value of

$$E_\beta(U_\beta|p)=\int_p^\pi \left(\beta(q,p)U_\beta(q)+(1-\beta(q,p))\,U_\beta(p)\right)dF_p(q)$$

when the leader stays in office and her expected payoff in total is $(1-\delta)x+\delta E_\beta(U_\beta|p)$. The citizen prefers to retain the leader rather than replacing him with a low-type challenger if and only if

$$(1-\delta)x+\delta E_\beta(U_\beta|p)\geq \delta w(\gamma). \tag{8}$$

Due to the leader's advantage, with probability p he stays in office regardless of the citizen's choice. With probability $(1-p)\gamma$ the citizen's choice matters and she would place a high-type challenger into office. With probability $(1-p)(1-\gamma)$, the citizen's choice matters and she would choose either the leader or a low-type challenger, whoever brings a better expected payoff. Therefore,

$$\begin{aligned} U_\beta(p) \;=\; & p\left((1-\delta)x+\delta E_\beta(U_\beta|p)\right)+(1-p)\left(\gamma+(1-\gamma)\right.\\ & \left.\times\,\max\left\{(1-\delta)x+\delta E_\beta(U_\beta|p),\delta w(\gamma)\right\}\right). \end{aligned} \tag{9}$$

Definition 1. *A strategy profile (β^*,κ^*) constitutes an equilibrium if*

1. *given κ^*, for all p and $q>p$,*

$$\beta^*(q,p)=\begin{cases}1, & L_{\kappa^*}(q)>L_{\kappa^*}(p)\\ 0, & L_{\kappa^*}(q)\leq L_{\kappa^*}(p)\end{cases}; \tag{10}$$

2. *given β^*, for all p,*

$$\kappa^*(p)=\begin{cases}1, & (1-\delta)x+\delta E_{\beta^*}(U_{\beta^*}|p)\geq\delta w(\gamma)\\ 0, & (1-\delta)x+\delta E_{\beta^*}(U_{\beta^*}|p)<\delta w(\gamma)\end{cases}. \tag{11}$$

In what follows, we characterize the vulnerability of democracy to subverting steps by an authoritarian-minded leader and its sustainability against such steps. We analyze two kinds of the leader's strategies.

Definition 2. *1. (Stopping strategy.) A strategy of "stopping at p" is β such that $\beta(q|p') = 0$ for all $p' \leq p$ and $q > p$;*

2. (Backsliding strategy.) A strategy of "backsliding at p" is β such that $\beta(q|p') = 1$ for all $p' \geq p$ and $q > p$.

In the benchmark, we assumed that the leader has a fixed level of advantage p. Given Definition 2, this is equivalent to the leader having the advantage of p and pursuing a strategy of stopping at p. Specifically, under such a strategy, when the leader's current advantage is below p, he would refrain from taking any step so large that increases his advantage over it. Once the leader's advantage reaches p, he stops taking any subsequent steps.

In contrast, under a strategy of backsliding at p, when the leader's current advantage is above p he would take every step that increases it further. As is clear from the definition, a strategy of backsliding at p is also a strategy of backsliding at any $p' > p$. This implies that once the leader starts backsliding, he never stops unless and until he is removed from office. Suppose the leader starts backsliding at t, pursuing a strategy of backsliding at p_t. The leader would take the step to increase advantage at t, so that if he survives to $t+1$, his advantage increases to $p_{t+1} = q_t > p_t$, for which the leader would continue backsliding. Iteratively, at any time since t, as long as the leader continues to be the incumbent, he would not stop backsliding.

Vulnerabilities of Democracy

Consider any period when the incumbent is the leader with the advantage of p. If the leader employs a strategy of stopping at p, the citizen's expected payoff, according to Proposition 1, is $\overline{U}(p, \gamma, x)$.

Lemma 1. *For any β and p,*

$$\overline{U}(\pi, \gamma, x) \leq U_\beta(p) \leq \overline{U}(p, \gamma, x). \tag{12}$$

Lemma 1 shows that, given the incumbent is the leader with the advantage of p, the citizen's expected payoff is bounded. It cannot be better than if the leader pursues a strategy of stopping at p, refraining from any step that further increases his advantage, and cannot be worse than if the leader already has the highest possible level of advantage π.

When choosing between the leader and a low-type challenger, the citizen compares the expected payoffs $(1-\delta)x + \delta E_\beta(U_\beta|p)$ and $\delta w(\gamma)$. In this comparison, only the citizen's continuation value of retaining the leader, $E_\beta(U_\beta|p)$, depends on the leader's strategy β and his current advantage p. On the one hand, according to Lemma 1, this continuation value cannot be worse than if the leader already has the highest possible level of advantage π, as

$$E_\beta(U_\beta|p) \geq E_\beta\left(\overline{U}(\pi, \gamma, x)|p\right) = \overline{U}(\pi, \gamma, x).$$

Therefore, if

$$(1-\delta)x + \delta\overline{U}(\pi,\gamma,x) \geq \delta w(\gamma), \tag{13}$$

condition (8) holds for any β and p. This implies that if the citizen prefers the leader rather than a low-type challenger to hold office when the leader already has the advantage of π, she would choose the leader over a low-type challenger when the leader has any strategy and any level of advantage. As shown in Proposition 1, Equation (13) holds if and only if $\pi \leq \frac{x}{\delta w(\gamma)}$, or equivalently, $x \geq \pi\delta w(\gamma)$. On the other hand, $E_\beta(U_\beta|p)$ cannot be better than if the leader has a strategy of stopping at p, refraining from increasing his advantage anymore, as

$$E_\beta(U_\beta|p) \leq E_\beta\left(\overline{U}(\cdot,\gamma,x)|p\right) \leq \overline{U}(p,\gamma,x),$$

where the second inequality is due to $\overline{U}$ being strictly decreasing. Therefore, if

$$(1-\delta)x + \delta\overline{U}(p,\gamma,x) < \delta w(\gamma), \tag{14}$$

condition (8) fails for any β and any $p' \geq p$. This implies that if the citizen prefers to replace the leader with a low-type challenger when the leader stops at p, she would choose a low-type challenger over the leader when the leader has any strategy and any level of advantage greater than p. As shown in Proposition 1, Equation (14) holds if and only if $p > \frac{x}{\delta w(\gamma)}$.

Proposition 2. [14]

1. (*Backsliding with support.*) *If* $x \geq \pi\delta w(\gamma)$, *there exists a unique equilibrium* (β^*, κ^*), *in which* β^* *is a strategy of backsliding at* $p = 0$ *and* $\kappa^*(p) = 1$ *for all* p.

2. (*Backsliding against opposition.*) *If* $x < \pi\delta w(\gamma)$, *for any equilibrium* (β^*, κ^*), β^* *is a strategy of backsliding at any* $p > \frac{x}{\delta w(\gamma)}$ *and* $\kappa^*(p) = 0$ *for all* $p > \frac{x}{\delta w(\gamma)}$.

Proposition 2 indicates two ways backsliding could happen in equilibrium. First, when the citizen derives a sufficiently large flow payoff from the leader, so that $x \geq \pi\delta w(\gamma)$, backsliding occurs with the citizen's support. Namely, the leader takes every possible step to accumulate advantage ever since he enters office, while knowing this, the citizen acts as if myopically, choosing at each period whoever provides a larger flow payoff. The citizen would try

[14] Note that this proposition holds even if we modify our timing assumption to allow the citizen observing the leader's step to undermine democracy in the current period. Essentially, the citizen is fully aware that the leader is taking every step to increase his advantage in both the case of backsliding with support and backsliding against opposition.

to remove the leader when a high-type challenger competes for office. If the citizen is not so lucky to have a high-type challenger right away, however, she would wait and the longer she waits, the less able she would be to remove the leader and place a high-type challenger into office when such a challenger is available.

Second, when the leader cannot provide a large enough flow payoff to the citizen to backslide with support, $x < \pi\delta w(\gamma)$, yet has gained an excessive level of advantage, $p > \frac{x}{\delta w(\gamma)}$, backsliding occurs against the citizen's opposition. Namely, once the leader's advantage exceeds $\frac{x}{\delta w(\gamma)}$, he takes every subsequent step to further increase advantage, while the citizen unconditionally opposes the leader, trying to remove him whoever the challenger is. Facing the citizen's unconditional opposition, the leader may still survive in office through his advantage. The longer the leader holds on to power, the more advantage he accumulates, the more likely he would survive yet another period.

Although the leader acts similarly in backsliding with support and in backsliding against opposition, he does so for different reasons. In the case of backsliding with support, the citizen's lenience makes taking subversive steps a *costless* investment in extra office security in case a high-type challenger competes against the leader. In the case of backsliding against opposition, the citizen's antagonism makes the leader's advantage his only resort to stay in power and, therefore, taking subversive steps is a *necessary* measure for survival.

According to Proposition 2, backsliding is inevitable when $x \geq \pi\delta w(\delta)$. In this case, there is a unique equilibrium in which the leader starts backsliding with support immediately at $t = 0$. When $x < \pi\delta w(\delta)$, however, backsliding happens only when the leader accumulates an excessively high level of advantage. Backsliding may be avoided in this case if there exists an equilibrium path on which the leader would never acquire that much advantage.

Sustainable Democracy: Definition

In this section, we define and characterize sustainable democracy, robust against authoritarian-minded leaders' attempts of backsliding. Formally,

Definition 3. *An equilibrium* (β^*, κ^*) *sustains democracy if* β^* *is a strategy of stopping at some* $p^* < \pi$. *Democracy is sustainable if an equilibrium that sustains democracy exists.*

In an equilibrium that sustains democracy, the leader would refrain from any step that is large enough to increase his advantage above some critical level p^*. Because the leader enters office with no advantage at all, $p_0 = 0$, his advantage would never exceed p^*. Therefore, every sustainable democracy is

associated with an endogenous *upper bound* on the leader's advantage. Once the leader's advantage reaches the upper bound, it would become stationary and the interaction between the leader and the citizen would be identical to the benchmark case discussed above. The smaller the upper bound, the higher is the ability of citizens to choose their preferred candidates for office. This upper bound on the leader's advantage, therefore, indicates how *undemocratic* is a particular sustainable democracy. A *perfect* sustainable democracy is one with the upper bound of $p^* = 0$.

Proposition 3. *If* (β^*, κ^*) *is an equilibrium that sustains democracy in which* β^* *is a strategy of stopping at* $p^* < \pi$*, then for any* $p \leq p^*$*,* $\kappa^*(p) = 1$ *and* $\beta^*(q, p) = 1$ *for all* p, q *such that* $p < q \leq p^*$*.*

Proposition 3 characterizes the leader's and the citizen's equilibrium behavior in a sustainable democracy. The leader enters office with no advantage and whenever his advantage is below its upper bound, the citizen chooses the leader over a low type challenger, while the leader takes every step that increases his advantage a bit but is small enough not to make it above the upper bound. This leads to an equilibrium path on which the citizen always chooses whoever provides a better flow payoff and the leader's advantage increases as long as he continues to hold office up to its upper bound.

Let $L^*(p, \gamma|p^*)$ be the leader's and $U^*(p, \gamma, x|p^*)$ be the citizen's expected payoff on this equilibrium path in any period when the leader is the incumbent and has the advantage of $p \leq p^*$. Because the citizen chooses whoever provides her with a better flow payoff, the leader stays in office in such a period with probability $p + (1-p)(1-\gamma) = 1 - (1-p)\gamma$. Given that the leader stays, he gets the flow payoff of 1 and the citizen gets the flow payoff of x in the current period and there are two possible cases for their continuation values. If the leader's step to increase advantage is small, so that $q \leq p^*$, he would take the step and his advantage would become q at the next period, yielding the leader the continuation value of $L^*(q, \gamma|p^*)$ and the citizen the continuation value of $U^*(q, \gamma, x|p^*)$. If the step is large, so that $q > p^*$, the leader would give up and his advantage would remain at p in the next period, yielding the leader the continuation value of $L^*(p, \gamma|p^*)$ and the citizen that of $U^*(p, \gamma, x|p^*)$. If the leader is removed, which is possible only with a high-type challenger, he gets 0 while the citizen gets 1 in the current and all future periods. Therefore,

$$\begin{aligned} L^*(p, \gamma|p^*) \;=\; & (1 - (1-p)\gamma) \\ & \times \left(1 - \delta + \delta \left(\begin{array}{l} \int_p^{p^*} L^*(q, \gamma|p^*) dF_p(q) \\ + (1 - F_p(p^*))\, L^*(p, \gamma|p^*) \end{array} \right)\right) \end{aligned} \tag{15}$$

and

$$U^*(p,\gamma,x|p^*) = (1-(1-p)\gamma) \times \left((1-\delta)x + \delta\left(\begin{array}{l}\int_p^{p^*} U^*(q,\gamma,x|p^*)dF_p(q) \\ +(1-F_p(p^*))\,U^*(p,\gamma,x|p^*)\end{array}\right)\right) + (1-p)\gamma. \quad (16)$$

Lemma 2. *$L^*(p,\gamma|p^*)$ is strictly increasing in p, p^*, strictly decreasing in γ, and*

$$U^*(p,\gamma,x|p^*) = L^*(p,\gamma|p^*)x + 1 - L^*(p,\gamma|p^*). \quad (17)$$

The above lemma characterizes the expected payoff the leader and the citizen gets on the equilibrium path of a sustainable democracy with the upper bound of p^* on the leader's advantage. Note that $L^*(p,\gamma|\pi)$ and $U^*(p,\gamma,x|\pi)$ are their expected payoffs under backsliding with support — when the citizen chooses the leader over a low-type challenger and the leader never stops until he is removed or gets the maximal level of advantage π.

The leader becomes better off each time he takes a small step to increase his advantage and manages to stay in office. The leader's position in office is least secure when he first enters office with no advantage at all, $p = 0$, in which case he gets the continuation value of $L^*(0,\gamma|p^*)$.

The citizen's expected payoff is always a weighted average between the flow payoff the leader provides, x, and that a high-type challenger provides, 1. The weight on x is exactly the leader's expected payoff. Intuitively, the leader's expected payoff can be understood as the proportion of periods when the leader holds office, with future periods being discounted at the rate of δ. In each of these periods, the citizen gets the flow payoff of x. Because the leader can be replaced only by a high-type challenger, in each of the other periods when the leader is out of office, the citizen gets the flow payoff of 1. Each time the leader takes a small step and manages to stay in power, he becomes better off and, because $x < 1$, the citizen becomes worse off. The worst possible case for the citizen on the equilibrium path, therefore, is when the leader's advantage reaches the upper bound p^*, where he stops taking subsequent steps.

Sustainable Democracy: Conditions

To keep democracy with the upper bound of p^* on the leader's advantage sustainable, the leader has to stay on the equilibrium path described in Proposition 3 and refrain from getting any advantage above the upper bound.

The strongest incentive the citizen can possibly offer is to punish the leader's deviation by unconditional removal. According to Proposition 2, If the citizen has such a strategy off the equilibrium path, the leader would have to backslide against the citizen's unconditional opposition whenever he gains any advantage $p > p^*$.

Let $\underline{L}(p)$ be the leader's and $\underline{U}(p, \gamma, x)$ be the citizen's expected payoff off under backsling against opposition. Due to the citizen's unconditional opposition, the leader stays in power in such a period with probability p, solely through his advantage. Given that the leader stays, he gets the flow payoff of 1 and the citizen gets x in the current period. In the next period, the leader's advantage would grow to $q > p$ and would continue increasing, so that the leader would get the continuation value of $\underline{L}(q)$ while the citizen would receive the continuation value of $\underline{U}(q, \gamma, x)$. Given that the leader is removed, he would get 0 in the current and all subsequent periods. For the citizen, with probability γ, the leader is replaced by a high-type challenger, in which case she gets 1 in the current and all future periods. With probability $1 - \gamma$, the leader is replaced by a low-type challenger, in which case the citizen gets 0 in the current period and the continuation value of $w(\gamma)$ in the future. The citizen's expected payoff when the leader is removed, in turn, is $\gamma + (1-\gamma)\delta w(\gamma) = w(\gamma)$. Therefore,

$$\underline{L}(p) = p\left(1 - \delta + \delta \int_p^{\pi} \underline{L}(q) dF_p(q)\right) \tag{18}$$

and

$$\underline{U}(p, \gamma, x) = p\left((1-\delta)x + \delta \int_p^{\pi} \underline{U}(q, \gamma, x) dF_p(q)\right) + (1-p)w(\gamma). \tag{19}$$

Lemma 3. *$\underline{L}(p)$ is strictly increasing in p and*

$$\underline{U}(p, \gamma, x) = \underline{L}(p)x + (1 - \underline{L}(p))\, w(\gamma). \tag{20}$$

Given the citizen's threat of unconditional removal, if the leader deviates, he would backslide against opposition and would be better off each time he successfully remains in power through his advantage. Therefore, assuming the citizen would unconditionally oppose the leader off the equilibrium path, the best case scenario for the leader when he deviates is to increase his advantage right away to the highest possible level π, in which case he gets the continuation value of

$$\underline{L}(\pi) = \pi\left(1 - \delta + \delta \underline{L}(\pi)\right) = \frac{(1-\delta)\pi}{1 - \delta\pi}.$$

In turn, given the citizen's threat of unconditional removal, the leader would never deviate when his worst possible continuation value on the equilibrium path is better than his best possible continuation value off the equilibrium path, that is, when

$$L^*(0,\gamma|p^*) \geq \frac{(1-\delta)\pi}{1-\delta\pi}. \tag{21}$$

The citizen's expected payoff when the leader backslides against opposition is always a weighted average between the flow payoff the leader provides, x, and the expected payoff a randomly drawn challenger provides, $\gamma + (1-\gamma)\delta w(\gamma) = w(\gamma)$. The intuition is similar to the case on the equilibrium path. The difference is that under backsliding against opposition, the leader may be removed regardless of the challenger's type. As a result, in each period the leader is out of office, the citizen's expected payoff is $w(\gamma)$. Note that backsliding against opposition is possible only when $x < \pi\delta w(\gamma) < w(\gamma)$. Hence, as the leader gets better off each time he manages to stay in power and continues backsliding, the citizen becomes worse off. Therefore, assuming the leader would backslide against opposition off the equilibrium path, the best case scenario for the citizen if the leader deviates is when the deviation is infinitesimal to $p^* + \epsilon$ for $\epsilon > 0$ as close as possible to 0, in which case the citizen's continuation value of retaining the leader is

$$\begin{aligned}(1-\delta)x + \delta\int_{p^*}^{\pi} \underline{U}(q,\gamma,x)dF_{p^*}(q) \;=\; &\left(1-\left(1-\int_{p^*}^{\pi}\underline{L}(q)dF_{p^*}(q)\right)\delta\right)x \\ &+\left(1-\int_{p^*}^{\pi}\underline{L}(q)dF_{p^*}(q)\right)\delta w(\gamma).\end{aligned}$$

In turn, the citizen always prefers a low-type challenger to the leader and thus would be willing to unconditionally remove the leader off the equilibrium path, in other words, her threat of unconditional removal is incentive compatible, if and only if

$$\begin{aligned}&\left(1-\left(1-\int_{p^*}^{\pi}\underline{L}(q)dF_{p^*}(q)\right)\delta\right)x \\ &\quad+\left(1-\int_{p^*}^{\pi}\underline{L}(q)dF_{p^*}(q)\right)\delta w(\gamma) \leq \delta w(\gamma).\end{aligned} \tag{22}$$

Proposition 4. *An equilibrium (β^*, κ^*) that sustains democracy in which β^* is a strategy of stopping at $p^* < \pi$ exists if and only if*

1. (*Incentive compatibility for the leader.*) $\gamma \leq g(p^*)$, *where* $g(p^*)$ *is strictly increasing in* p^*, $g(0) = 1 - \pi$, *and* $g(p^*) < \frac{1-\pi}{1-p^*}$ *for all* $p^* > 0$;[15]

2. (*Incentive compatibility for the citizen.*) $p^*\delta w(\gamma) \leq x \leq h(p^*)\delta w(\gamma)$, *where* $h(p^*)$ *is strictly increasing in* p^*, $h(\pi) = \pi$, *and* $h(p^*) > p^*$ *for all* $p^* < \pi$.

Democracy is sustainable if and only if $\gamma < g(\pi)$ *and* $g^{-1}(\gamma)\delta w(\gamma) \leq x < \pi\delta w(\gamma)$.

As illustrated by the dashed lines in Figure 2, two conditions have to be met to sustain democracy with the upper bound of p^* on the leader's advantage. First, to make it incentive compatible for the leader, the probability that a high-type challenger rises to compete for office must be sufficiently low, so that

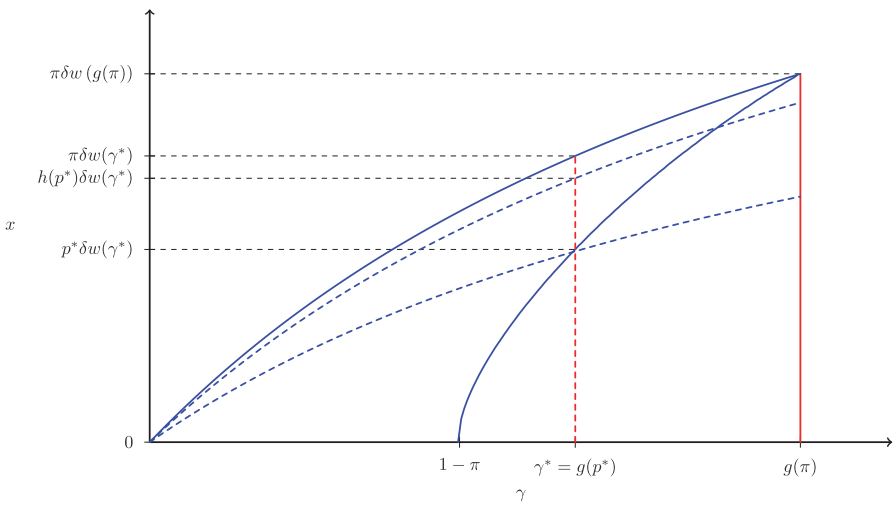

Figure 2: Sustainability of democracy.

Blue solid: $x = \pi\delta w(\gamma)$ and $x = g^{-1}(\gamma)\delta w(\gamma)$; red solid: $\gamma = g(\pi)$.
Region encircled by solid lines: democracy is sustainable.
Blue dashed: $x = h(p^*)\delta w(\gamma)$ and $x = p^*\delta w(\gamma)$; red dashed: $g(p^*)$.
Region encircled by dashed lines: democracy with the upper bound of p^* is sustainable.
Parameters: $\pi = 0.8$, $F(q) = \frac{q}{\pi}$, $\delta = 0.8$, $p^* = 0.6$, $\gamma^* = g(p^*) = 0.138$.

[15]Note that as long as $p_0 < \frac{x}{\delta w(\gamma)}$, it is relevant only for the incentive compatibility for the leader. With a generic p_0, the leader's incentive compatibility condition (21) becomes $L^*(p_0, \gamma|p^*) \geq \frac{(1-\delta)\pi}{1-\delta\pi}$. Because $L^*(p_0, \gamma|p^*)$ is strictly increasing in p_0, this condition is easier to hold with a larger p_0. Intuitively, a larger p_0 makes the leader's position in office more secure and therefore leaves him a weaker incentive to deviate to gain excessive advantage at the risk of triggering the citizen's unconditional opposition.

$\gamma \leq g(p^*)$. Intuitively, the citizen cannot commit to retain the leader when a high-type challenger competes for office. The leader faces office insecurity on the equilibrium path due to the citizen's incentive to place a high-type challenger into office. The more likely such a challenger would rise, the more severe is the office insecurity the leader suffers on the equilibrium path, and as a result, the more tempted the leader would be to deviate by gaining a level of advantage above the upper bound.

Second, to make democracy with the upper bound of p^* incentive compatible for the citizen, the flow payoff the leader provides to the citizen must be intermediate, so that $p^*\delta w(\gamma) \leq x \leq h(p^*)\delta w(\gamma)$. If the leader provides an overly low flow payoff $x < p^*\delta w(\gamma)$, the citizen would not allow him to increase his advantage close to the upper bound on the equilibrium path. When the leader does gain any advantage close enough to the upper bound, the citizen would unconditionally oppose him, triggering backsliding against opposition. If the leader instead provides an overly high flow payoff $x > h(p^*)\delta w(\gamma)$, the citizen would not be willing replace the leader with a low-type challenger off the equilibrium path when he increases his advantage above the upper bound. In other words, the citizen's threat of unconditional opposition when the leader gains any advantage above the upper bound would not be credible from the perspective of the leader. Knowing this, the leader is both willing and able to push his advantage above the upper bound.

In general, democracy is sustainable, that is, there exists at least one equilibrium that sustains democracy, when the probability that a high-type challenger arises is not too high, so that $\gamma < g(\pi)$ and the leader provides to the citizen with an intermediate flow payoff, so that $g^{-1}(\gamma)\delta w(\gamma) \leq x < \pi\delta w(\gamma)$. These conditions are illustrated by the region encircled by the solid lines in Figure 2. When the citizen likes the leader too much in comparison to an average challenger, so that $x \geq \pi\delta w(\gamma)$, democracy is vulnerable because the leader would backslide with support. When the citizen likes an average challenger too much in comparison to the leader, so that $g^{-1}(\gamma)\delta w(\gamma) > x$, democracy is vulnerable because the leader would be tempted to take large steps for excessive levels of advantage, which would lead to backsliding against opposition.

Sustainable Democracy: Comparison

Every sustainable democracy is associated with an upper bound on the leader's advantage. This upper bound, p^*, can be thought of as indicating how undemocratic is a particular sustainable democracy: a lower upper bound means that citizens have more control over whom they can place in office.

Because the leader enters office with no advantage at all, $p_0 = 0$, the citizen's ex-ante expected payoff in a sustainable democracy with the upper

bound of p^* on the leader's advantage is

$$U^*(0, \gamma, x|p^*) = L^*(0, \gamma|p^*)x + 1 - L^*(0, \gamma|p^*) \tag{23}$$

According to Lemma 2, $L^*(0, \gamma|p^*)$ is strictly increasing in p^*, so that $U^*(0, \gamma, x|p^*)$ is strictly decreasing in p^*. Hence, keeping x and γ fixed and assuming sustainability, the citizen is always better off in a more democratic regime that has a more stringent constraint on the leader's advantage and the best regime is perfect democracy in which $p^* = 0$.

Obviously, the citizen's ex-ante expected payoff is strictly increasing in x, which measures how attractive the leader is to the citizen. Moreover, due to Lemma 2, $L^*(0, \gamma|p^*)$ is strictly decreasing in γ, so that the citizen's ex-ante expected payoff $U^*(0, \gamma, x|p^*)$ is also strictly increasing in γ, which is her expectation about the potential challengers. Sustainability imposes upper limits on both x and γ.

First, to make a democracy with the upper bound of p^* sustainable, x cannot exceed the upper limit of $h(p^*)\delta w(\gamma)$, which is strictly increasing in p^*. In other words, to keep democracy sustainable when the incumbent is an authoritarian-minded leader, this leader can be more attractive to citizens in a less democratic regime than he could be in a more democratic regime.

Second, to sustain a democracy with the upper bound of p^*, γ cannot exceed the upper limit of $g(p^*)$, which is strictly increasing in p^*. This implies that while keeping democracy sustainable against an authoritarian-minded leader, citizens can have a better expectation about potential challengers in a less democratic regime than they could in a more democratic regime.

Therefore, the citizen faces a trade-off between more democratic regime and more attractive politicians. A more democratic regime sustainable against authoritarian-minded leaders requires that citizens find these leaders less attractive and, at the same time, that they are more pessimistic about potential challengers. In the extreme, perfect democracy can be sustained only if the citizen receives an extremely low flow payoff from the leader, $x \leq h(0)\delta w(\gamma)$, and faces an extremely low probability of a high-type challenger running for office, $\gamma \leq g(0) = 1 - \pi$. Allowing the leader to gain some advantage makes democracy imperfect, but benefits the citizen through loosening these constraints.

The best ex-ante expected payoff the citizen can get in a sustainable democracy with the upper bound of p^* is

$$V^*(p^*) := U^*\left(0, g(p^*), h(p^*)\delta w\left(g(p^*)\right)|p^*\right),$$

when she receives the highest possible flow payoff from the leader, $x = h(p^*)\delta w(\gamma)$, and has the highest possible probability to see a high-type challenger competing for office, $\gamma = g(p^*)$, while keeping such a democracy sustainable.

Proposition 5. *For any* $p^* < \pi$,

$$V^*(p^*) = \frac{(1-\delta)\pi}{1-\delta\pi} h(p^*)\delta w\left(g(p^*)\right) + 1 - \frac{(1-\delta)\pi}{1-\delta\pi} \tag{24}$$

is strictly increasing in p^*.

Proposition 5 indicates that in the citizen's trade-off between more democratic regime and more attractive politicians, the citizen always favors the latter. Indeed, under the constraints of sustainability, perfect democracy is the worst possible regime from the citizen's perspective. It is simply too costly to sustain.

The upper bound on the leader's advantage has three competing effects on the citizen's ex-ante expected payoff. On the one hand, it has a direct *negative* effect. Keeping x and γ fixed, $U^*(0, \gamma, x|p^*)$ is strictly decreasing in p^* because $L^*(0, \gamma|p^*)$, as the weight on x, is strictly increasing in p^*. Intuitively, when a high-type challenger is available, the citizen would benefit from replacing the leader with such a challenger and she is less able to remove the incumbent in a less democratic regime. On the other hand, the upper bound on the leader's advantage has two indirect *positive* effects through the constraints of sustainability. A larger p^* makes the incentive compatibility constraint of the citizen weaker, allowing the citizen to get a better flow payoff from the leader, that is, a larger x. It also makes the incentive compatibility constraint of the leader weaker, allowing the citizen a higher probability to have a high-type challenger running for office, that is, a larger γ. Most importantly, the direct negative effect and the second indirect positive effect through γ neutralize each other. The upper limit of γ is a result of binding the leader's incentive compatibility constraint (21), so that

$$L^*\left(0, g(p^*)|p^*\right) = \frac{(1-\delta)\pi}{1-\delta\pi},$$

which is constant in p^*. Consequently, the only effect of p^* that remains is the positive indirect effect through x. Therefore, a higher upper bound on the leader's advantage has a net positive effect on the citizen's ex-ante expected payoff.

Conclusion

Defending democracy imposes a difficult challenge on individual citizens. To act against a government that is undermining democracy, people must weigh their satisfaction with the incumbent against the effect of backsliding on their future ability to replace it by a better one. Even if individuals have consistent time preferences (Akerlof, 1991) and even if they fully anticipate that the

government will not stop short of seeking complete domination, they may still prefer the incumbent over potential opponents, giving up future possibilities to remove him. In turn, when the effects of antidemocratic measures are not visible immediately, it may be too late to remove the incumbent once these effects materialize. To prevent backsliding, people must both value their ability to choose governments and understand that while each of the backsliding measures may be perfectly legal, their effect is to protect the incumbent from being defeated in the future.

The fact is that most backsliding governments find ways to remain in power. President Trump did lose the election and in spite of some attempts failed to hold onto office. Yet while several other backsliding governments suffered temporary reversals, they were able to recover and continue: With 40.9 percent of votes, the AKP failed to win a majority of seats in the election of June 7, 2015 but it called for a new election and won 49.5 percent of the vote five months later. Three years later, in June 2018, Erdogan won the presidential election with 52.6 percent. In Poland, PiS won an absolute majority of parliamentary seats in October 2019 and maintained the presidency in July 2020. In Hungary, Fidesz and its allies won a reelection in April 2018 with 44.9 percent of the vote. In Venezuela, Chavez won a re-election in 2006 with 62.8 percent of the vote and again in 2102 with 55.1 percent. He enjoyed majority support in the polls and the opposition became majoritarian only after his death (Venezuelabarometro). In Brazil, Bolsonaro continues to enjoy popular support in spite of his disastrous handling of the Covid pandemic. The implication must be that either many people do not care about preserving their ability to remove backsliding incumbents or that they do not see the consequences of supporting them.[16]

The optimism that citizens would effectively threaten to punish governments that commit transgressions against democracy and thus prevent them from taking this path is sadly tenuous. This view is based on the assumption that when a government commits some acts that flagrantly threaten liberty, violate constitutional norms, or undermine democracy, people will unify against it. Yet people may not react to such violations even when they observe them or they may be unable to assess their consequences. And if citizens do not stop the government from taking some series of steps, it may be too late to prevent it from doing whatever it wants.

References

Akerlof, G. A. 1991. "Procrastination and Obedience". *American Economic Review*. 81: 1–19.

[16]For a recent empirical study on how voters react to backsliding, see Svolik (2021).

Dragu, T., X. Fan, and J. Kuklinski. 2014. "Designing Checks and Balances". *Quarterly Journal of Political Science*. 9: 1–42.
Fearon, J. 2011. "Self-enforcing Democracy". *Quarterly Journal of Economics*. 126: 1661–708.
Ginsburg, T. and A. Huq. 2018. *How to Save a Constitutional Democracy*. Chicago: University of Chicago Press.
Graham, M. and M. W. Svolik. 2020. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States". *American Political Science Review*. 114(2): 392–409.
Grillo, E. and C. Prato. 2019. "Opportunistic Authoritarians, Reference-dependent Preferences, and Democratic Backsliding". *Working Paper*.
Helmke, G., M. Kroeger, and J. Paine. 2021. "Democracy by Deterrence: Norms, Constitutions, and Electoral Tilting". *Working Paper*.
Howell, W. G., K. A. Shepsle, and S. Wolton. 2019. "Executive Absolutism: A Model". *Working Paper*.
Howell, W. G. and S. Wolton. 2018. "The Politician's Province". *Quarterly Journal of Political Science*. 13(2): 119–46.
Li, A., D. Raiha, and K. W. Shotts. 2020. "Propaganda, Alternative Media, and Accountability in Fragile Democracies". *Working Paper*.
Madison, J. 1788. "The Federalist Papers by Alexander Hamilton, James Madison, and John Jay". In: ed. G. Wills. 1982nd ed. New York: Bantam Books.
Miller, M. K. 2021. "A Republic If You Can Keep It: Breakdown and Erosion in Modern Democracies". *Journal of Politics*. 83(1): 198–213.
Montesquieu, B. 1748. *De l'Esprit de Lois*. 1995th ed. Paris: Gallimard.
Nalepa, M., G. Vanberg, and C. Chioopris. 2020. "A Wolf in Sheep's Clothing: Citizen Uncertainty and Democratic Backsliding". *Working Paper*.
Palmer, R. R. 1959. *The Age of the Democratic Revolution*. Vol. 1. Princeton: Princeton University Press.
Persson, T., G. Roland, and G. Tabelini. 2001. "Separation of Pwer and Accountability: Towards a Formal Approach to Comparative Politics". *Working Paper*.
Svolik, M. W. 2021. "Voting Against Autocracy". *Working Paper*.
Varol, O. 2015. "Stealth Authoritarianism". *Iowa Law Review*. 100: 1673–742.
Weingast, B. R. 1997. "Political Foundations of Democracy and the Rule of Law". *American Political Science Review*. 91: 245–63.
Weingast, B. R. 2015. "Capitalism, Democracy, and Countermajoritarian Institutions". *Supreme Court Economic Review*. 23: 255–77.