Reference Points and Democratic Backsliding*

Edoardo Grillo† Carlo Prato‡

May 19, 2021

Running Title: Reference Points and Democratic Backsliding

Word Count: 9,834

Keywords: Democratic Backsliding, Reference Points, Autocratic Backsliding, Context-Dependent Voting

*We thank Avi Acharya, Sheri Berman, Peter Buisseret, Jon Eguia, Teresa Esteban-Casanelles, Tim Frye, Diego Gambetta, John Huber, Giovanna Invernizzi, Kimuli Kasara, Krzysztof Krakowski, Zhaotian Luo, John Marshall, Anne Meng, Davide Morisi, Monika Nalepa, Jacopo Perego, Chiara Superti, Michael Ting and Stephane Wolton, as well as seminar and conference participants at UC San Diego, Virtual Formal Theory Workshop, and OVERS, for their helpful comments and conversations. Antonio Camara, Stefano Cravero, and Colleen Wood provided outstanding research assistance. The paper was previously circulated under the title "Opportunistic Authoritarians, Reference-Dependent Preferences, and Democratic Backsliding."

†Collegio Carlo Alberto, Piazza Arbarello 8, Torino, 10122, Italy *Email*: edoardo.grillo@carloalberto.org

‡Department of Political Science, Columbia University, 420 W. 118th St, NY 10027 *Email*: cp2928@columbia.edu

Abstract

We propose a theory of democratic backsliding where citizens' retrospective assessment of an incumbent politician depends on expectations that are endogenous to the incumbent's behavior. We show that democratic backsliding can occur even when most citizens *and most politicians* intrinsically value democracy. By challenging norms of democracy, an incumbent can lower citizens' expectations; by not doubling down on this challenge, he can then beat this lowered standard. As a result, gradual backsliding can actually enhance an incumbent's popular support not despite of, but *because of* citizens' opposition to backsliding. This mechanism can only arise when citizens are uncertain enough about incumbents' preferences (e.g., owing to programmatically weak parties). Mass polarization, instead, can reduce the occurrence of backsliding while simultaneously increasing its severity.

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: `https://doi.org/10.7910/DVN/UVFOWU`.

1. Introduction

In the summer of 2019, after withdrawing his party from the cabinet where he was serving as Deputy Prime Minister, Matteo Salvini asked voters to grant him "full powers, to carry out what we promised in full, without holdups or stumbling blocks." During his tenure, Salvini opened investigations against the judges who struck down his executive order denying asylum seekers access to public services and threatened to remove police protection from a journalist who criticized him. He also defied the constitutional authority of Italy's President Sergio Mattarella over the appointment of Paolo Savona—the author of a plan detailing Italy's exit from the Eurozone—as finance minister. In the end, the verdicts stood, the security details remained in place, and a less controversial figure was appointed to the finance ministry.

These attempts to weaken judicial independence, silence the media, and set off a constitutional clash between the executive and the head of state were not popular among voters (Mattarella's approval rating, for instance, remained stable throughout the confrontation and well above Salvini's). And yet, they brought substantial gains in the polls: support for Salvini's party almost doubled in little over a year, and during his clash with Mattarella, Salvini's own approval rating grew by nine percentage points.[1]

These patterns are hardly exceptional. From Boris Johnson's prorogation of Parliament in the United Kingdom to the forced retirement of judges in Poland, from Viktor Orban's weakening of the Supreme Court in Hungary to the repetition of Istanbul's mayoral election in Turkey, scholars and observers are increasingly concerned about democratic backsliding, the loosening of constraints of accountability on the actions of democratically elected leaders (Waldner and Lust, 2018; Levitsky and Ziblatt, 2018; Przeworski, 2019). And, since observational and experimental evidence shows that voters, all else equal, dislike challenges to democratic norms (Graham and Svolik, 2020), we should expect these challenges to reduce the popularity of an incumbent, not improve it.

[1]Sources: Istituto Ixé "Political Environment" surveys conducted on May 13, 2018 and June 19, 2018 and Istituto Piepoli "Trust in Leader" surveys conducted on May 16, 2018 and June 21, 2018, `https://bit.ly/32WYYMF`, accessed November 17, 2020.

This paper shows that democratic backsliding can occur even when most citizens and most politicians intrinsically value democracy. We propose a theory of context-dependent retrospection where citizens evaluate incumbents according to a standard that can be manipulated. This leads to *opportunistic authoritarians*—incumbents who attack democratic institutions to enhance their popularity. Our results suggest that the programmatic weakening of political parties is crucial for the emergence of opportunistic authoritarians, while the effect of checks and balances and mass polarization on backsliding is more subtle than previously theorized.

Our theory is built on the premise that (i) citizens and politicians share a primitive aversion to violations of democratic norms, but (ii) some of them (a minority of both groups) are willing to accept them in order to achieve radical policy change, and (iii) politicians also value popular support. Consistent with the idea of backsliding as a gradual process, we assume that the incumbent first chooses whether to challenge democracy and then how much to double down (i.e., the severity of the challenge).

Our key innovation is that a citizen's assessment of the incumbent is not based solely on an absolute standard—his performance in office—but also on context-dependent factors, captured by a *reference point*. The reference point corresponds to citizens' expectation about the material payoff the incumbent will yield them. If the payoff citizens actually experience is above this expectation, their support for the incumbent increases; if the payoff falls below this expectation, their support decreases.

Context-dependent preferences have a long history in social and behavioral sciences. A large body of evidence documents their importance for electoral choices (Quattrone and Tversky, 1988), attitudes towards legislators (Kimball and Patterson, 1997), the executive (Waterman, Jenkins-Smith and Silva, 1999), and democratic institutions (Corazzini et al., 2014).

In our model, citizens form their reference points before the incumbent's choice of doubling down but after his choice of challenging democratic norms.[2] Hence, citizens' reference points respond to incumbent behavior. If they believe that the incumbent is likely to double down on dismantling democratic norms, then their reference point is low. If the incumbent ends

[2]The timing is crucial, but also quite natural. See the discussion in Section 3.

up not doubling down, his performance exceeds the reference point. This produces a sense of relief: citizens think that "it could have been worse", and their retrospective assessment of the incumbent improves. As a result, an incumbent can challenge democratic norms and enjoy substantial support not despite citizens' aversion to democratic backsliding, but *precisely because of* it.

The increase in support associated with a challenge followed by a partial retreat can make democratic backsliding politically appealing. Because of reference-dependence, public opinion ends up rewarding challenges to democratic norms instead of encouraging incumbents to respect these norms. The psychological mechanism we identify, however, also encourages some incumbents to hold off on their initial challenge, since doubling down generates disappointment (performance falling below citizens' reference points) and depresses support. While responsiveness to public opinion can rein in the impulses of autocratic incumbents, this paper shows that it can also encourage gradual backsliding.

In our model, citizens' reference points are not arbitrary. Rather, they are derived from correct conjectures about incumbents' future (equilibrium) behavior, in line with a rational-expectation approach (Kőszegi and Rabin, 2007). This approach produces sharp, testable predictions that distinguish our model from alternative accounts of backsliding.

First, we show that citizens' uncertainty about the incumbent's ideology increases the likelihood of democratic backsliding. We relate this result to the documented link between the rise of populist authoritarians and the disintermediation of representation from political parties (Mair, 2002; Rosenblum, 2010) in favor of direct communication via social media. Challenging democracy is a more viable strategy when citizens' expectations about leaders' behavior are not anchored to parties' programmatic identities or the fact-based reporting of traditional media outlets. Going back to our initial example, Salvini's tenure as the leader of the *Lega Nord* coincided with a large shift in the party's platform and communication strategy, which de-emphasized regional autonomy and anti-clericalism in favor of a more generic ethnic nationalistic message directly broadcast from Salvini's social media accounts.

These predictions help us distinguish our setting from a model where challenges to democracy are driven by the incumbent's desire to test the (uncertain) strength of the public's opposi-

tion to backsliding. In that case, the likelihood of backsliding (i) does not respond to citizens' uncertainty about the incumbent's preferences, and (ii) should be very low when politicians, through social media and big data, have access to powerful tools to learn citizens' attitudes.

Second, our results show that mass polarization can simultaneously decrease the likelihood of backsliding and increase its expected severity. The reason is that mass polarization weakens the responsiveness of citizens' support to the behavior of the incumbent. This reduces the disciplining effect of public opinion on autocrats but also the value of lowering citizens' expectations for opportunistic authoritarians.

Third, our theory provides a mechanism that simultaneously accounts for citizens' intrinsic commitment to democracy (Voeten, 2016), their increased dissatisfaction with democratic governance (Foa and Mounk, 2016), and the popularity of leaders who gradually erode democratic norms observed in Turkey, Poland, Hungary, and—on a smaller scale—in the United States, the United Kingdom, and other Western democracies. In Section 5.1, we illustrate how the evolution of public opinion during several episodes of democratic backsliding matches the predictions of our theory.

2. Related Literature

Our paper contributes to the literature on the causes of democratic backsliding and to the study of context-dependent preferences in formal political theory.

Over the last decade, scholars have become increasingly concerned about democratic backsliding, i.e., violations of the limits on the ability of the executive to use the power of the government (Waldner and Lust, 2018). These actions encompass the breach of traditionally respected norms, the testing of the boundaries of the law, and its outright violation (Howell and Wolton, 2018; Howell, Shepsle and Wolton, 2019; Helmke, Kroeger and Paine, 2019; Versteeg et al., 2019).

Scholars have focused on two main culprits: the rise of mass polarization and the weakening of political parties. Recent work formally and experimentally shows how higher mass polarization leads fewer voters to sanction violations of democratic norms (Chiopris, Nalepa

and Vanberg, 2021; Luo and Przeworski, 2020; Graham and Svolik, 2020; Carey et al., 2020; Miller, 2020). A common premise of these theories is that elected incumbents have authoritarian ambitions (Svolik, 2019) in pursuit of which they are willing to sacrifice popular support. Our theory, instead, shows that backsliding can occur even when most incumbents share voters' affinity for democratic norms. It also suggests that the relationship between polarization and backsliding is more subtle than previously theorized.[3]

Another line of literature links the weakening of parties' programmatic identity to the twin phenomena of backsliding (Rosenbluth and Shapiro, 2018; Urbinati, 2019; Levitsky and Cameron, 2003) and populism (Berman and Snegovaya, 2019; Prato and Wolton, 2018). These authors argue that deep societal changes (increases in income dispersion, immigration, and the importance of social media) have stifled parties' ability to mediate between government and society (Stokes, 1999; Rosenblum, 2010), thereby producing voter confusion. This paper formalizes a mechanism through which voter confusion improves the appeal of opportunistic political entrepreneurs with authoritarian stances.

Finally, our model contributes to the formal literature on context-dependent preferences in political science (Callander and Wilson, 2006, 2008). In our model, citizens evaluate the performance of an incumbent relative to their expectations, captured by a reference point (see Kahneman and Tversky, 1979 and Bell, 1985 for seminal contributions). This paper follows the work of Kőszegi and Rabin (2006, 2007), where the reference point is endogenously derived from the players' equilibrium behavior. A growing literature, pioneered by Lindstädt and Staton's (2012) reduced-form approach, applies reference dependence to international relations (Acharya and Grillo, 2019), electoral competition (Alesina and Passarelli, 2019; Lockwood and Rockey, 2020; Panunzi, Pavoni and Tabellini, 2020; Karakas and Mitra, 2021), political protests (Passarelli and Tabellini, 2017), and campaigns (Grillo, 2016).

[3]See Grossman et al. (2020) for a different argument mitigating the relationship between polarization and backsliding.

3. Baseline Model

A unit mass of citizens indexed by i ("she") is ruled by an incumbent I ("he").

First, I chooses whether to respect ($c = 0$) or challenge democratic norms ($c = 1$), e.g., by challenging the prerogatives of the legislature or the head of state, or by announcing a measure restricting the constitutional rights of certain groups.[4] Subsequently, he chooses a policy y from the interval $\mathcal{Y}(c) \subset \mathbb{R}$. For simplicity, $\mathcal{Y}(0) = 1$: if I does not challenge, his subsequent policy choice is $y = 1$. Instead, if I challenges democratic norms, he can achieve more extreme policies: $\mathcal{Y}(1) = [1 + \delta, 2]$. Hence, we can write $y(c, d) = 1 + cd$, where the choice variable $d \in [\delta, 1]$ captures the severity of the escalation against democratic norms. For instance, during his clash with President Mattarella, Salvini could choose from a set of options ranging from trying to get Savona appointed to a less important cabinet position to full-blow impeachment proceedings against Mattarella.

When $d = 1$, the incumbent fully escalates. When $d = \delta$, he holds off. The parameter $\delta \in (0, 1)$ is inversely related to the strength of institutional *checks and balances*. Lower δ captures an increase in the the power (and/or willingness) of various institutional actors (e.g., the judicial) to curtail the incumbent's initial challenge, thereby reducing its consequences.[5] Figure 1 summarizes the incumbent's sequence of actions.

[FIGURE 1 ABOUT HERE]

Citizens vary in their policy preferences but share a common intrinsic aversion to violations of democratic norms (see, e.g., Graham and Svolik, 2020; Carey et al., 2020). For instance, citizens disagree about how tight immigration restrictions should be, but they all prefer that due process and rule of law be respected. This aversion can be justified by the presence of future periods in which democratic backsliding reduces constraints on office-holders or makes it harder to remove them from office (Luo and Przeworski, 2020).

[4] In Supplemental Appendix C.2.2 (page 9), we show that the binary nature of the decision to challenge is without loss of generality.

[5] In Supplemental Appendix C.2.3 (page 12), we allow checks and balances to affect the whole range of possible escalation levels d.

Each citizen i evaluates policy outcomes $y(c,d)$ in light of her ideology θ_i, reflected in the payoff $\theta_i y(c,d)$. Citizens' ideology parameters are distributed according to the cumulative density function F. Citizens with a positive (negative) ideology favor (oppose) the incumbent's direction of policy change (which, without any loss of generality, is towards the right). The common aversion to backsliding is captured by the payoff $-cd$. As a result, a citizen opposes (favors) democratic backsliding if and only if her ideology is below (above) 1. Let $\mathbf{q} = (c,d)$ be the outcome of the incumbent's behavior. Citizens' *material utility* is given by $\theta_i y(c,d) - cd$, i.e.,

$$u(\mathbf{q};\theta_i) = \underbrace{\theta_i(1+cd)}_{\text{Policy preference}} \qquad \underbrace{-cd}_{\text{Aversion to backsliding}} \tag{1}$$

Assumption 1. *F is uniform over the interval $\left[-\frac{1}{2\psi}, \frac{1}{2\psi}\right]$ with $\frac{1}{2\psi} > 1$.*

The parameter ψ captures the degree of ideological homogeneity in society: lowering ψ increases the share of citizens with extreme policy preferences.[6] We then interpret an increase in mass polarization as a reduction in ψ. Assumption 1 implies that a majority of citizens, but not all, oppose democratic backsliding. Some citizens are willing to accept it for ideological reasons or due to economic distress, a sense of disenfranchisement, or ethnoracial prejudice (Hahl, Kim and Sivan, 2018; Pettigrew, 2017; Smith and Hanley, 2018).

Like citizens, the incumbent has an ideology θ_I. In addition, I values citizens' support (for example, because of reelection motives). His utility function is

$$u_I(\mathbf{q};\theta_I) = u(\mathbf{q};\theta_I) + R\pi(\mathbf{q}), \tag{2}$$

where $\pi(\mathbf{q})$ is the share of citizens who support I and $R \in \mathbb{R}_+$ is the relative importance of support (e.g., the strength of his electoral concern). The incumbent observes his ideology θ_I, but citizens do not. Their prior is given by the cumulative density function F_I.

[6] The uniform assumption is for tractability, but our insights qualitatively extend to other distributions (e.g., $1-\psi$ could measure the probability mass on the extremes of a fixed support). $2\psi < 1$ guarantees that the incumbent's support (see below) is always interior.

Assumption 2. *F_I is uniform over the interval $\left[\tau - \frac{1}{2\phi}, \tau + \frac{1}{2\phi}\right]$, with $\tau \in (0,1)$ and $\frac{1}{2\phi} > \max\left\{\frac{R}{\delta} + \tau - 1, \frac{R}{1-\delta} + 1 - \tau\right\}$.*

τ is the incumbent's average ideology, and $\frac{1}{2\phi}$ measures citizens' uncertainty about it. $\tau < 1$ implies that most incumbents oppose backsliding. Although our results extend to the case of $\tau > 1$, Assumption 2 ensures that leaders' autocratic tendencies cannot produce a substantial likelihood of backsliding. The lower bound on $\frac{1}{2\phi}$, instead, ensures that some incumbents are immune to public opinion, so their behavior is entirely driven by their policy payoff.

Once the incumbent has made his choices ($\mathbf{q}$), citizens decide whether to support him or not. Citizens' behavior depends on their *total utility*, which is the sum of their material utility, $u(\mathbf{q}; \theta_i)$, and an additional psychological component capturing reference-dependence. The psychological component depends on how much the utility experienced by citizen i exceeds or falls short of her reference point $\underline{u}$. When this gap is positive, citizen i experiences a psychological gain (relief); when it is negative, she suffers a psychological loss (disappointment). The parameter $\eta \in \mathbb{R}_+$ captures the importance of this psychological component relative to material utility:

$$v(\mathbf{q}; \theta_i | \underline{u}) = u(\mathbf{q}; \theta_i) + \eta\big[u(\mathbf{q}; \theta_i) - \underline{u}\big] \tag{3}$$

In line with Kőszegi and Rabin (2006, 2007), the reference point is determined endogenously: it equals the citizen's expected utility *following the incumbent's decision to challenge or not*. Formally, let $\theta_I \mapsto \hat{\mathbf{q}}(\theta_I) = (\hat{c}(\theta_I), \hat{d}(\theta_I))$ describe the incumbent's behavior. Then, the reference point of a citizen with ideology θ_i when she observes c is given by:

$$\underline{u}(c; \hat{\mathbf{q}}, \theta_i) = E\left[u(\hat{\mathbf{q}}; \theta_i) \mid c\right]. \tag{4}$$

As a result, the incumbent's decision to challenge democratic norms has two consequences. First, it changes the set of policy choices available to him. Second, it prompts citizens to closely consider the ultimate consequences of the incumbent's actions, which leads to the formation of their reference points.

An equilibrium is a profile $(\hat{\mathbf{q}}, \underline{u}(0; \hat{\mathbf{q}}, \theta_i), \underline{u}(1; \hat{\mathbf{q}}, \theta_i))$ specifying a sequentially rational strategy $\hat{\mathbf{q}}$ for each incumbent's type and a reference point for each choice of c and each citizen's

type θ_i. Reference points have the fixed-point structure typical of rational expectations. On the one hand, reference points affect support—and, thus, the behavior of the incumbent. On the other hand, the behavior of the incumbent feeds back into reference points.

3.1 Discussion

Before proceeding with the analysis, we highlight two key features of the model.

First, citizens form their reference points after the incumbent's decision to challenge democracy but before the choice of how much to double down. If citizens' reference points were entirely determined before the incumbent's actions, they could not respond to his behavior. If citizens' reference points were entirely determined after the incumbent's actions, material payoffs would always coincide with reference points, and reference-dependence would play no role. Our results would be qualitatively unaffected if reference points were determined not only by the incumbent's actions, but also by exogenous factors such as the duration of democratic institutions or the behavior of previous incumbents. Our results would also continue to hold if citizens formed their reference point at the beginning of the game and updated it after every non-terminal history.

Second, in line with experimental (Woon, 2012) and empirical (for a review, see Healy and Malhotra, 2013) evidence, the baseline model assumes that citizens' assessments of the incumbent are purely *retrospective*. Yet, in light of an influential critique of retrospection in models of electoral accountability (Fearon, 1999), Supplemental Appendix C.4 (page 17) shows that our results extend to situations in which citizens' evaluations are *prospective*, i.e., their support for the incumbent depends on their conjectures about his future performance.

4. Analysis

Given retrospective evaluations, a citizen with ideology θ_i supports the incumbent if and only if $v(\mathbf{q};\theta_i) \geq 0$.[7] The incumbent's support is thus equal to

$$\pi(\mathbf{q}) = \int_{-\frac{1}{2\psi}}^{\frac{1}{2\psi}} \mathbb{1}_{\{v(\mathbf{q};z)\geq 0\}} dF(z) \tag{5}$$

Incumbent behavior affects his policy utility (policy concerns) and his popular support (popularity concerns), which depends on the impact of such behavior on citizens' material and psychological payoffs. To understand how these three channels operate, we introduce them sequentially. We begin with the benchmark case of no popularity concern ($R = 0$). We then introduce popularity concerns in the absence of psychological payoffs associated with reference dependence ($R > 0$ and $\eta = 0$). Finally, we analyze the novel incentives generated by reference dependence.

4.1 Policy Concerns

When $R = 0$, the incumbent's behavior does not respond to public opinion. Instead, he simply maximizes his policy utility $\theta_I(1 + cd) - cd$. If θ_I exceeds one, the value of a more extreme policy exceeds the loss from weakening democratic norms, so I chooses $c = 1$ and then fully doubles down ($d = 1$). We refer to incumbents with $\theta_I > 1$ as *autocrats*. If instead θ_I is below one, the incumbent prefers not to challenge democratic norms and sets $c = 0$. We refer to incumbents with $\theta_I \leq 1$ as *democrats*.

Proposition 1. *When the incumbent has no popularity concerns ($R = 0$),*
(i) if the incumbent is an autocrat ($\theta_I > 1$), $c = 1$ and $d = 1$;
(ii) otherwise ($\theta_I \leq 1$), $c = 0$, and there is no backsliding.

Proof. Proofs of all formal statements are in the Appendix. □

[7] The specific way in which citizens resolve an indifference does not affect the analysis, nor does the choice of zero as a threshold for supporting the incumbent.

4.2 Popularity Concerns without Reference Dependence

Suppose that the incumbent values citizens' support ($R > 0$), but citizens do not exhibit reference dependence ($\eta = 0$). In this case, popularity concerns are entirely driven by citizens' material payoffs. Only citizens with $u(\mathbf{q}; \theta_i) \geq 0$ support the incumbent. Since most citizens oppose backsliding (by Assumption 1, $u(\mathbf{q}; \theta_i)$ is decreasing in both c and d for a majority of them), challenges to democratic norms necessarily reduce the incumbent's popular support. When the incumbent respects democratic norms, his support equals $\pi(0, 0) = \Pr(\theta_i \geq 0) = 1 - F(0) = \frac{1}{2}$. When he challenges them, more citizens abandon him, and the loss in support is *increasing* in the level of subsequent escalation:

$$\pi(1, d) = 1 - F(\theta_i + d\theta_i - d) = \frac{1}{2} - \psi \frac{d}{1+d}.$$

The incumbent's payoff can then be written as:

$$u_I(c, d; \theta_I) = \theta_I + (\theta_I - 1)cd + \frac{R}{2} - R\psi c \frac{d}{1+d}. \tag{6}$$

Since democratic backsliding reduces popular support, all democratic incumbents choose to respect democratic norms. Autocratic incumbents instead face a trade-off between popular support and their own policy preferences. Only autocratic types that are extreme enough (*extreme authoritarians*) choose to violate norms, and when they do, they always double down.[8] Autocratic incumbents with less extreme ideologies, conversely, are unwilling to accept the loss in public support associated with backsliding. These *restrained autocrats* choose to respect democratic norms despite their intrinsic preferences ($u_I(1, 1; \theta_I) \leq u_I(0, 0; \theta_I)$). Their ideologies fall in the interval $\theta_I \in (1, \theta^\dagger]$, with

$$\theta^\dagger := 1 + \frac{R\psi}{2} \tag{7}$$

[8]Because the loss in support is concave in the level of escalation, conditional on challenging these norms, extreme authoritarians choose full escalation.

Proposition 2. *When the incumbent has popularity concerns ($R > 0$), but citizens do not exhibit reference dependence ($\eta = 0$),*
(i) if the incumbent's autocratic tendencies are strong enough ($\theta_I > \theta^\dagger$), $c = 1$ and $d = 1$;
(ii) otherwise, ($\theta_I \leq \theta^\dagger$), $c = 0$, and there is no backsliding.

$\theta^\dagger$ captures the disciplining power of popularity concerns (for example, electoral incentives). This force has a long intellectual history and it directly links to a key argument for the centrality of electoral institutions in democratic regimes (Schumpeter, 1942; Popper, 1945). By institutionalizing the contingency of a ruler's power on popular support, elections protect societies from unpopular governance outcomes.

Proposition 2 is not inconsistent with the notion that democratic backsliding unfolds over time, but it predicts that incumbents should always double down, which is at odds with empirical accounts of recent episodes of democratic backsliding, with attacks often followed by sudden retreats and significant setbacks (Levitsky and Ziblatt, 2018).

Moreover, $\theta^\dagger$ (and, thus, the frequency of restrained autocrats) is decreasing in $\frac{1}{\psi}$: mass polarization reduces the drop in support associated to backsliding. Hence, Proposition 2 implies that polarization should increase the likelihood of backsliding (Chiopris, Nalepa and Vanberg, 2021; Svolik, 2019).

In the next section, we show that reference dependence (i) induces incumbent behaviors that are more consistent with observed patterns, (ii) creates incentives for democrats to engage in democratic backsliding, and (iii) alters the way in which mass polarization and checks and balances affect the occurrence and severity of backsliding.

4.3 Reference Dependence and Opportunistic Authoritarians

We now consider the case in which an incumbent with popularity concerns ($R > 0$) faces citizens who exhibit reference dependence ($\eta > 0$). As discussed above, reference points are determined by citizens' expectations $\underline{u}(0; \hat{\mathbf{q}}, \theta_i)$ and $\underline{u}(1; \hat{\mathbf{q}}, \theta_i)$ (which in equilibrium are correct) about the incumbent's behavior following $c \in \{0, 1\}$. Given that utilities are linear in policy choices, reference points are determined by $\underline{d}_c := E[\hat{d} \mid c]$, the expected level of

escalation following the choice of c:

$$\underline{u}(c; \hat{\mathbf{q}}, \theta_i) = \theta_i + (\theta_i - 1)\underline{d}_c,$$

where $\underline{d}_0 = 0$ and $\underline{d}_1 \in [\delta, 1]$.

If the incumbent does not challenge (i.e., $c = 0$), citizens face no uncertainty regarding the policy choice. Hence, the total utility of a citizen is equal to her ideology, $v(0, d; \theta_i) = \theta_i$, the incumbent's support is equal to $1/2$, and his utility equals

$$u_I(0, 0; \theta_I) = \theta_I + \frac{R}{2}. \tag{8}$$

If instead I challenges democratic norms, citizens' behavior depends on the expected level of escalation, $\underline{d}_1$. Fixing an expected ($\underline{d}_1$) and actual (d) level of escalation, a citizen with ideology θ_i supports the incumbent if and only if

$$\begin{aligned} v(1, d; \theta_i) &= \theta_i + (\theta_i - 1)d + \eta\Big[\theta_i + (\theta_i - 1)d - \theta_i - (\theta_i - 1)\underline{d}_1\Big] \\ &= \theta_i + (\theta_i - 1)\Big[d + \eta(d - \underline{d}_1)\Big] \geq 0. \end{aligned} \tag{9}$$

To guarantee that a citizen's propensity to support the incumbent after a challenge is increasing in her ideology, in the main text we assume that δ is large enough. (In Supplemental Appendix B (page 1), we provide a complete characterization and we show that a failure of Assumption 3 strengthens our main result.)

Assumption 3. *Institutional checks and balances are not too strong:*

$$\delta > \frac{\eta - 1/2}{1 + \eta}$$

As a result, when $c = 1$ the incumbent's support equals

$$\pi(1, d) = \frac{1}{2} - \psi\frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}. \tag{10}$$

Notice that support is strictly decreasing and strictly convex in d. Since the median citizen dislikes democratic backsliding, doubling down entails a loss in support whose size increases in the level of escalation d. Substituting (10) into the incumbent's utility, we obtain

$$u_I(1, d; \theta_I) = \theta_I + (\theta_I - 1)d + R\left[\frac{1}{2} - \psi\frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}\right]. \tag{11}$$

Crucially, $\pi(1, d)$ is not necessarily lower than $\pi(0, 0)$. To see why, consider Figure 2. Following a challenge to democratic norms, citizens expect at least some incumbents (those with extreme policy preferences) to double down. Since most citizens dislike backsliding, their reference points will go down (cf. the dotted gray line in Figure 2). Because "it could have been worse," the decision not to double down produces relief. The resulting positive psychological payoff may offset the negative material payoff from partial backsliding.

In line with the attribution bias (see, Haggag et al. 2018, Bushong and Gagnon-Bartsch 2019, and the references therein), this mental process ought not to be explicit nor sentient. Citizens may misattribute their positive attitude toward the incumbent to the material payoff, deeming it better than it actually is, and still behave in line with our model. In fact, the mechanism we describe is also in line with (and can provide a behavioral foundation for) the notion of *blind retrospection* (Healy and Malhotra, 2013; Achen and Bartels, 2017).

Comparing (8) and (11) reveals the potential trade-off faced by an incumbent when choosing c. Challenging democratic norms shifts policy outcomes but might reduce popular support:

$$u_I(1, d; \theta_I) - u_I(0, 0; \theta_I) = \underbrace{(\theta_I - 1)d}_{Policy\ \ Drift} - \underbrace{R\psi\frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}}_{Public\ \ Opinion\ \ Feedback}. \tag{12}$$

For an autocrat ($\theta_I > 1$), the value of shifting policy outcomes fully offsets the cost of backsliding, i.e., the *policy drift* is positive. However, challenging also changes citizens' retrospective evaluations. On the one hand, it lowers the policy payoff of most citizens. On the other hand, it reduces their reference points from θ_i to $\theta_i + (\theta_i - 1)\underline{d}_1$. Depending on the importance of psychological factors in citizens' assessments, there are two cases.

Proposition 3. *When reference dependence has little impact on citizens' utility ($\eta < \frac{\delta}{2-\delta}$), the incumbent's equilibrium behavior is identical to the one described in Proposition 2.*

When psychological factors are not too important, all incumbents who challenge democratic institutions fully escalate.[9] Because citizens' reference points are determined in equilibrium, $\underline{d}_1 = 1$. Hence, in equilibrium, citizens do not experience any disappointment or relief, and the cutoff type of the incumbent who is indifferent between challenging and not challenging is still $\theta^\dagger$. Since $\underline{d}_1 = 1$, the incumbent's support conditional on challenging is

$$\theta_I + (\theta_I - 1)d + R\left(\frac{1}{2} - \frac{d + \eta(d-1)}{1 + d + \eta(d-1)}\right). \tag{13}$$

Because (13) is decreasing in d, choosing $d = \delta$ *enhances* the incumbent's popular support: if citizens are expecting full escalation, the choice not to escalate comes as a positive surprise for (a majority of) citizens, as illustrated in Figure 2. Exploiting this fear-and-relief mechanism is especially tempting for autocratic incumbents with less extreme ideologies, i.e., with θ_I close to $\theta^\dagger$. The condition $\eta < \delta/(2-\delta)$ ensures that the increase in support generated by citizens' relief is not too large, so that every incumbent with type around (and including) $\theta^\dagger$ prefers to play according to the equilibrium strategy described in Proposition 3.

Now suppose that the psychological payoffs are important enough, $\eta \geq \delta/(2-\delta)$. By the argument above, if citizens expected full escalation after a challenge ($\underline{d}_1 = 1$), then some incumbents would find it profitable to partially retreat and enjoy the increase in support from the associated relief. Convexity of the incumbent's support in d implies that if challenges occur in equilibrium, incumbents will either choose no further escalation ($d = \delta$) or full escalation ($d = 1$). Since the incumbent's utility satisfies the single crossing condition, the equilibrium escalation level must also be weakly increasing in ideology. The incumbent's equilibrium behavior is then described by two cutoffs $\underline{\theta}$ and $\overline{\theta}$, jointly determined with the expected escalation $\underline{d}_1$ (see equations (17)-(19) in the Appendix).

Proposition 4. *When reference dependence is important enough ($\eta \geq \frac{\delta}{2-\delta}$), there exists $\underline{\theta}$ and $\overline{\theta} > 1$ such that:*

[9] Notice that if the condition on η in Proposition 3 holds, Assumption 3 holds as well.

(i) if $\theta_I > \overline{\theta}$, $c = 1$ *and* $d = 1$*;*

(ii) if $\theta \in (\underline{\theta}, \overline{\theta}]$, $c = 1$ *and* $d = \delta$*;*

(iii) otherwise ($\theta_I \leq \underline{\theta}$), $c = 0$, *and there is no backsliding.*

Moreover, the expected level of escalation following a challenge equals

$$\underline{d}_1 = 1 - (1 - \delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi}. \tag{14}$$

[FIGURE 2 ABOUT HERE]

Opportunistic authoritarians. The behavior of incumbents with ideology $\theta_I \in (\underline{\theta}, \overline{\theta}]$ is driven by the interaction between reference dependence and political incentives. Compared to Proposition 2, incumbents in the interval $(\theta^\dagger, \overline{\theta}]$ back down after the initial challenge because they want to benefit from the increase in support associated with citizens' relief. For these incumbents, reference dependence strengthens the disciplining effect of public opinion and limits the severity of democratic backsliding.

Incumbents in the interval $(\underline{\theta}, \theta^\dagger]$, conversely, challenge democratic norms even though they would have respected them in the absence of reference dependence. The benefit of the additional support generated by citizens relief more than offsets the loss in material utility associated with backsliding. For these incumbents, reference dependence weakens the disciplining effect of public opinion and increases the likelihood of democratic backsliding.

What drives the frequency of opportunistic authoritarians? Below, we show that it increases with the uncertainty about the incumbent's ideology, i.e., it decreases in ϕ. Moreover, when ϕ is small enough (so that extreme autocrats are likely enough), the relief associated with a partial retreat may fully offset the loss in citizens' material payoff from backsliding and push even *democratic* incumbents to challenge democracy.

Proposition 5. *The likelihood of opportunistic authoritarians increases with the uncertainty concerning incumbents' ideology:* $\partial\underline{\theta}/\partial\phi > 0$ *and* $\partial\overline{\theta}/\partial\phi < 0$. *Furthermore, there exists* $\phi^* \in \mathbb{R}$*, such that if* $\phi < \phi^*$ *and reference dependence is important enough, opportunistic authoritarians also include some democrats,* $\underline{\theta} < 1$.

Proposition 5 implies that the empirical relevance of opportunistic authoritarians increases in citizens' uncertainty about the incumbent's programmatic preferences. Leaders who frequently depart from their parties' traditional platforms and resort to personalistic appeals are most susceptible to challenging democratic norms. Matteo Salvini and Boris Johnson exemplify this pattern. Salvini is a former far-left militant who re-branded a regional independentist party into a ethnonationalist movement centered around his leadership. Johnson rose to power by building a reputation for pragmatism and personal charm as mayor of London—a city where Labor voters have been outnumbering Conservatives since 1992.

In practice, citizen uncertainty can be reduced by strong political parties that anchor their leaders' programmatic commitments and by a robust, independent media system. Our results then formalize the idea that the weakened intermediation by parties and media is a key prerequisite for populist authoritarianism (Mair, 2002; Rosenblum, 2010).

[FIGURE 3 ABOUT HERE]

Figure 3 summarizes the incumbent's equilibrium behavior.[10] If reference dependence is not important enough (i.e., if $\eta \leq \frac{\delta}{2-\delta}$), the equilibrium behavior of the incumbent is identical to the case of no reference dependence and only autocrats with sufficiently high ideology ($\theta_I > \theta^\dagger$) challenge democratic norms (and then fully escalate).

If, instead, reference dependence is sufficiently important (i.e., if $\eta > \frac{\delta}{2-\delta}$) opportunistic authoritarians emerge. Incumbents with ideology in $(\underline{\theta}, \overline{\theta}]$ challenge democratic norms and then partially retreat in order to exploit the fear-and-relief mechanism described above. Compared to the case of low reference dependence, incumbents with ideologies between $\theta^\dagger$ and $\overline{\theta}$ (highlighted in dark gray in Figure 3) choose partial retreat ($d = \delta$) instead of full escalation ($d = 1$). Incumbents with ideologies between $\underline{\theta}$ and $\theta^\dagger$ (highlighted in light gray in Figure 3) conversely choose to challenge democratic norms instead of respecting them.

As η further increases, the fear-and-relief mechanism becomes so strong that restrained autocrats disappear (i.e., $\underline{\theta}$ falls below one), and some democrats challenge democracy. When

[10] Recall that Assumption 3 puts an upper bound on η. See Supplemental Appendix B (page 1) for a characterization of the equilibrium when Assumption 3 fails.

this happens, changes in the incumbent's incentives have counter-intuitive effects. For instance, stronger political responsiveness (either as an increase in the relative importance of popular support, R, or in the responsiveness of citizens' behavior to their realized payoff, ψ) decreases $\underline{\theta}$, thereby increasing the likelihood of backsliding. Stronger electoral incentives can then encourage democratic incumbents to behave in an authoritarian manner. This not only goes against the intrinsic preferences of these incumbents, but is also counter to the interests of citizens.

5. Implications

5.1 The Dynamics of Public Opinion

Our theory's key prediction is that backsliding results from a series of challenges and retreats with three key features. First, the challenge is unpopular among the majority of the citizenry. Second, the incumbent becomes less popular following the challenge. Third, the retreat ends up restoring or even boosting his popularity relative to the pre-challenge level.

This dynamic is at play in several recent episodes of backsliding. Consider first the U.K. Parliament prorogation controversy. After ascending to premiership in July 2019 with a platform centered on "getting Brexit done," Boris Johnson found himself without a parliamentary majority to achieve that goal. To avoid a vote that would have extended the October deadline for Britain's exit from the E.U. in the absence of an agreement, on August 28, 2019, the Cabinet obtained a five-week prorogation of parliament.

Due to its timing and unusual length, the suspension was denounced as an illegal attempt to curtail the authority of Parliament. On September 24, 2019, the U.K. Supreme Court ruled the prorogation unlawful and, therefore, null. While disagreeing with the verdict, the following day Boris Johnson vowed to respect it and not to seek a second prorogation.

Prorogation was unpopular: polls show that only 30% of the British public supported it, while 46% opposed it.[11] Furthermore, support for Boris Johnson dropped substantially. His

[11]Ipsos MORI Online Brexit Polling `https://bit.ly/3nrGpIe`, accessed on 11/18/2020.

net approval fell from -7% in late July to -18% in mid-September.[12] Nevertheless, in the aftermath of the ruling, his popularity quickly returned to its late July level. By the end of October, his net approval was already at +2%. Less than two months later, his party handily won the general election.

Boris Johnson's prorogation and Matteo Salvini's efforts to force the appointment of Paolo Savona (described in the Introduction) illustrate how challenges to democratic norms often target institutions (the U.K. Parliament and the Italian Presidency) endowed with oversight authority over the executive. However, attacks can also be leveled against electoral institutions or citizens' individual rights, as illustrated by the 2019 Istanbul mayoral election and the 2016 attempted abortion ban in Poland.

On March 31, 2019, the nationally ruling Justice and Development Party (AKP) narrowly lost the mayoral election in Istanbul—Turkey's economic and financial center—to the opposition candidate Ekrem Imamoglu. The AKP immediately sought to invalidate the vote by alleging minor administrative irregularities. In early May, the Supreme Electoral Council nullified the result and called a new election for June 23. Imamoglu won again, by a substantially larger margin. While the AKP leader and President of Turkey Recep Tayyip Erdogan was in a position to force through his candidate, he chose to respect the electoral result.

Turkish citizens opposed the Supreme Electoral Council's ruling by a margin of 36 percentage points. Nevertheless, Erdogan's popularity did not suffer much. After dropping from -2.2% in March to -6.6% in April, his net approval rating began to improve over the summer and by September was already positive (+1.3%).[13]

Another instance of the public opinion dynamics described by the model comes from Poland. After winning the 2015 election, in September 2016 the ruling Law and Justice (PiS) party

[12]Source: Ipsos MORI Political Monitor `https://bit.ly/3f8gAKl` accessed on 11/18/2020. Other available polls for the period show the same dynamics. See Supplemental Appendix E (page 23).

[13]Sources: MetroPOLL Center for Strategic and Social Research *Turkey's Pulse*, June 2019 `https://bit.ly/2HOpYTx` (page 48) and October 2019 `https://bit.ly/3fjEDG8`, accessed on 11/13/2020.

allowed a draft law to reach the final stage of debate in the Polish Parliament. The proposed bill, resulting from a civil initiative that enjoyed the support of many PiS MPs (BBC, 2016), sought to tighten restrictions on abortion by banning all elective abortions and punishing doctors performing them with jail time.

A poll in September 2016 showed overwhelming support for the existing legislation, with only 11% in favor of tighter restrictions (Chrzczonowicz, 2017). Following the preliminary vote, on October 3, 2016, more than 100,000 people across the country took to the streets to protest against the bill (Korolczuk, 2016). In response to the protests, the PiS withdrew its support to the bill, and, on October 6, the Parliament rejected the proposal. Despite these events, the PiS suffered only a minor reduction in public support, which fell from about 40% in July/August to 35% in mid-September, before returning to 38% by mid-October.[14] This pattern is not unique to this episode: as Marcinkiewicz and Stegmaier (2017) document, the PiS enjoyed a remarkably stable support, despite several (and unpopular) attempts to weaken judicial independence by lowering the retirement age of judges.

[FIGURE 4 ABOUT HERE]

Our theory also produces implications on the evolution of support across different ideological groups. Support from citizens who are programmatically aligned with the incumbent and ideologically extreme (i.e., those with $\theta_i \geq 1$) should first increase after a challenge and then decrease after a retreat. Conversely, support among citizens who (i) are programmatically aligned with the incumbent but have a moderate ideology ($\theta_i \in (0, 1)$), or (ii) are not programmatically aligned with the incumbent ($\theta_i < 0$) should decrease after a challenge and then rebound after a retreat. Although the Trump administration's norm-violating initiatives (e.g., the Muslim Ban) made some GOP supporters uneasy, criticism rarely developed into open opposition and faded quickly following the legal setbacks that these initiatives encountered. These patterns, illustrated in Figure 4, imply that when it comes to backsliding, the main political cleavage is not between citizens who are programmatically aligned with

[14]Data from CBOS Public Opinion Research Center's *Polish Public Opinion 1/2017* `https://bit.ly/35P67Ah`, accessed on 11/13/2020.

the incumbent and citizens who are not. Rather, it is between those with extreme enough ideologies and everyone else.

5.2 The Effect of Mass Polarization

Previous scholarship has singled out mass polarization as a key determinant of democratic backsliding (Chiopris, Nalepa and Vanberg, 2021; Svolik, 2019): as polarization increases, citizens' voting decisions become less responsive to the behavior of incumbents, who can then try to short-circuit democratic norms with relative impunity. While our theory does not contradict this idea, it does highlight that the role for polarization it subtler.

When either reference dependence is sufficiently weak or citizens' uncertainty about the incumbent's ideology is limited (i.e., ϕ is large, so that the expected escalation level $\underline{d}_1$ is sufficiently low), the same force described in Svolik (2019) operates in our model. Higher polarization (i.e., lower ψ) weakens the disciplining effect of public opinion. As a result, fewer autocrats are deterred from engaging in democratic backsliding.

However, when reference dependence is strong enough and citizens are sufficiently uncertain about the incumbent's ideology, polarization reduces the frequency of opportunistic authoritarians. By weakening citizens' response to the incumbent's actions, polarization reduces the incentive to exploit the fear-and-relief mechanism that drives the behavior of opportunistic authoritarians. As a result, mass polarization can decrease the overall likelihood of backsliding and the likelihood of mild episodes of backsliding (i.e., challenges followed by holding off), while increasing the likelihood of severe episodes of backsliding (i.e., full escalation).

Our analysis suggests that the link between polarization and democratic backsliding is less straightforward than previously theorized. This can help explain the lack of correlation between polarization and the occurrence of democratic backsliding in the U.S. states recently documented by Grumbach (2021). An equally thorough analysis of cross-country data is beyond the scope of this paper. However, the preliminary "reality check" of Figure 5 confirms that polarization and backsliding are related in a subtle way. In line with our theory, Draca and Schwarz (2020)'s measure of polarization seems negatively associated with below-median

(left plot) yearly reductions in the *V-Dem*'s Liberal Democracy Index (Coppedge et al., 2020) and positively associated with above-median yearly reductions in the same index (right plot).

[FIGURE 5 ABOUT HERE]

5.3 Institutional Checks and Balances

Conventional wisdom dating back at least to the Madisonian idea that "ambition must be made to counteract ambition" (Hamilton, Madison and Jay, 2008, no. 51) holds that stronger checks and balances should protect democracy from challenges from within. Our model suggests that this intuition is incomplete.

Proposition A.1 in Supplemental Appendix A (page 1) shows that stronger checks and balances (lower δ) may increase the likelihood of backsliding (a challenge is more likely) *and* decrease its expected severity (full escalation is less likely). On the one hand, lower δ reduces the damage of a challenge followed by a retreat (higher material utility). On the other hand, lower δ increases the relief that citizens experience when an incumbent retreats, thereby increasing the appeal of the fear-and-relief strategy and the likelihood of opportunistic authoritarians.

The above results do not depend on the specific way in which we model checks and balances. In Supplemental Appendix C.2.3 (page 12), we allow for a more flexible approach in which δ increases both the upper bound and the lower bound of the range of possible escalation levels. As long as the upper bound is not significantly more responsive to δ than the lower bound, our results continue to hold. For example, they extend to the case of $d \in [\delta, 1+\delta]$.[15] Intuitively, lower δ improves the material utility of most citizens and thus enhances the appeal of the fear-and-relief mechanism described above.

[15] If checks and balances only affect the upper bound (e.g., $d \in [0, 1+\delta]$) an immediate generalization of the argument in footnote 16 implies that all incumbents would challenge, most would back down, and only incumbents with extreme θ_I would fully escalate.

6. Alternative Explanations

In Supplemental Appendix D (page 19), we formally study two alternative explanations for backsliding: desensitization and boundaries testing. Under desensitization, challenges to democratic norms weaken citizens' emotional response to subsequent violations, thereby facilitating backsliding. Supplemental Appendix D.1 (page 19) shows that desensitization cannot account for two key results in this paper: retreats (which desensitization discourages) and challenges by democratic incumbents.

Backsliding can also be driven by an incumbent's willingness to test citizens' opposition to the dismantling of democracy (modeled as the realization of a noisy signal after a challenge). Supplemental Appendix D.2 (page 20) shows that this mechanism can generate challenges to democratic norms followed by partial retreats. However, the likelihood of partial retreats is increasing in the incumbent's uncertainty about citizens' preferences and negligible when the latter is small. The implication is that more accurate public opinion tools should reduce both the probability of a challenge and the probability of a retreat. This is at odds with the widespread use of micro-targeting, social media, and big data that have improved politicians' ability to track citizens' preferences. Perhaps even more important, retreats in the boundaries testing model should be associated with a sharp decline in the incumbent's popularity, which goes against the evidence motivating this paper.

7. Robustness

Several assumptions in our baseline model can be relaxed without affecting our results. For instance, our mechanism continues to operate in a model of non-ideological challenges or "power grabs" formally studied in Supplemental Appendix C.3 (page 13). In this setting, challenges do not expand the set of achievable policy outcomes, but they directly improve the incumbent's chances of staying in power, thereby insulating him from public opinion (as in Luo and Przeworski, 2020; Gratton and Lee, 2020).

Our model also assumes that the incumbent knows the distribution of citizens' preferences (while citizens are uncertain about his preferences). Assuming uncertainty about the average of the distribution of θ_i simply shifts the incumbent's expected support (equation 5) by a constant, thereby leaving the values of the equilibrium thresholds unaffected.

Introducing risk aversion in material payoffs does not qualitatively affect our results. Given the timing of our model—with the decision to support the incumbent occurring when uncertainty has been resolved—, risk aversion would only lower reference points and thus boost the relief associated with beating expectations.

Supplemental Appendix C.2.1 (page 8) shows that when a second, fully reversible challenge (i.e., one with $d \in [0, 1]$) is available, in equilibrium no incumbent strictly prefers it (provided that δ is large enough).[16] Supplemental Appendix C.2.2 (page 9) also shows that our results continue to hold if we allow an arbitrarily large number of intermediate challenges. From a methodological standpoint, these extensions also show how to adapt the notion of sequential equilibrium to our environment.

In the baseline model, the incumbent's support (and his overall utility) is convex in the level of escalation d. Convexity might not hold under different assumptions about the importance of reference dependence (η), citizens' material utility ($u(\mathbf{q}; \theta)$), or the distribution of ideologies among citizens (F). In this case, some incumbents may choose an interior level of escalation. Supplemental Appendix B (page 1) shows that our insights continue to hold as long as the the incumbent's utility satisfies the single-crossing property, so that more extreme incumbents choose higher levels of escalation. Allowing for interior solutions that vary continuously with the incumbent's type significantly increases analytic complexity, and additional assumptions might be needed to guarantee the existence of an equilibrium.[17] If an

[16] With a single fully reversible challenge (i.e., $\delta = 0$), all incumbents would choose $c = 1$, and most with then back down. As a result, there would no longer be a behavioral distinction between restrained autocrats and opportunistic authoritarians (see Figure 3).

[17] This is common with rational expectations equilibria. Given the definition of $\underline{d}_1$, the fixed point problem in the proof Proposition 4 would have to account for a continuous cdf over the range of d, $[\delta, 1]$, rather than the average between its two extreme values. Then,

equilibrium exists, the effects of polarization and checks and balances do not change relative to the baseline model.

8. Conclusion

This paper presents a theory of democratic backsliding in which, despite the fact that most citizens and most incumbents intrinsically dislike violations of democratic norms, these violations arise frequently and can even increase an incumbent's popular support.

When citizens' reference dependence is weak or their uncertainty about the incumbent's preferences is negligible, our theory implies that polarization and weak checks and balances contribute to the emergence of backsliding, in line with existing scholarship.

When instead citizens' reference dependence is strong and their uncertainty about the incumbent is substantial—as it has been recently in the U.S., the U.K., and Italy, where leaders have abandoned traditional programmatic campaigns in favor of direct and personalistic appeals—, these insights become subtler. Our work can then reconcile otherwise puzzling empirical patterns in politicians' behavior and citizens' attitudes. Challenging democratic norms allows incumbents to move the goal posts to their advantage. As a recent Washington Post column suggests (Hiatt, 2019), these actions lead citizens to focus on the fact that "it could have been worse," all the while things continue to get worse.

Our theory highlights how politicians can manipulate citizens' emotional reactions—in particular, relief—to their advantage. Recent events in U.S. politics suggest that disappointment can also affect citizens' response to politicians' behavior. The deep disappointment experienced by core Trump supporters towards key members of the Republican establishment—who after endorsing four years of norm-violating behavior chose to accept the outcome of the 2020 election—likely contributed to the storming of the United States Capitol on January 6, 2021.

Brouwer's fixed point theorem would no longer apply. However, our existence argument would extend to any arbitrarily large but finite number of possible escalation levels. In this case, in equilibrium some incumbents may choose intermediate levels of escalation.

9. Appendix: Proofs

[Table 1 ABOUT HERE]

Proof of Proposition 1. Absent popularity concerns, the utility of the incumbent is given by $u_I(\mathbf{q};\theta_I) = \theta_I + (\theta_I - 1)cd$. Hence, incumbents with ideology $\theta_I > 1$ choose the pair (c,d) that maximizes the product cd, namely $c = 1$ and $d = 1$. Instead, incumbents with ideology $\theta_I < 1$ choose the pair (c,d) that minimizes the product cd, namely $c = 0$ and $d = 0$. Incumbents with ideology exactly equal to 1 are indifferent among all feasible pairs (c,d); since such incumbents have measure zero, we assume without loss of generality that they choose $c = 0$ and $d = 0$. □

Proof of Proposition 2. The utility of the incumbent is given by

$$u_I(c,d;\theta_I) = \theta_I + (\theta_I - 1)cd + \frac{R}{2} - R\psi c\frac{d}{1+d}.$$

Note that, when $c = 1$, the incumbent's utility is strictly convex in d. Because $d \in [\delta, 1]$, this implies that, conditional on choosing $c = 1$, I will choose either $d = \delta$ or $d = 1$. In the former case, his utility is

$$u_I(1,\delta;\theta_I) = \theta_I + (\theta_I - 1)\delta + \frac{R}{2} - R\psi\frac{\delta}{1+\delta}.$$

In the latter case, his utility is

$$u_I(1,1;\theta_I) = \theta_I + (\theta_I - 1) + \frac{R}{2} - R\psi\frac{1}{2}.$$

Observe that $u_I(1,\delta;\theta_I) > u_I(0,0;\theta_I)$ if and only if $\theta_I \geq 1+R\psi/(1+\delta)$ and that $u_I(1,\delta;\theta_I) > u_I(1,1;\theta_I)$ if and only if $\theta_I \leq 1 + R\psi/(2(1+\delta))$. Hence, whenever the incumbent is better off choosing $(1,\delta)$ instead of $(0,0)$, he strictly prefers $(1,1)$ to $(1,\delta)$. As a consequence, $d = \delta$ is never optimal when the incumbent prefers $c = 1$ to $c = 0$. Comparing $u_I(1,1;\theta_I)$ with $u_I(0,0;\theta_I)$, we can then conclude that incumbents with ideology $\theta_I < 1 + R\psi/2$ choose $(c,d) = (0,0)$, while those with ideology $\theta_I > 1 + R\psi/2$ choose $(c,d) = (1,1)$. Incumbents

with ideology $\theta_I = 1 + R\psi/2$ are indifferent between $(0,0)$ or $(1,1)$ and we assume without loss of generality that they choose $(0,0)$. $\square$

Proof of Proposition 3. The incumbent's utility in this case is given by

$$u_I(c,d;\theta_I) = \theta_I + (\theta_I - 1)cd + \frac{R}{2} - R\psi c \frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}.$$

Since $\frac{\delta}{2-\delta} < \frac{1+\delta}{1-\delta}$, when

$$\eta \leq \frac{\delta}{2-\delta}, \tag{15}$$

the expression is convex in d. Hence, we can follow the reasoning of the proof of Proposition 2 to conclude that the behavior described in this proposition (which implies $\underline{d}_1 = 1$) is an equilibrium as long as the incumbent $\theta^\dagger$ prefers $d = 1$ to $d = \delta$ (even though δ generates a positive surprise equal to $1 - \delta$). The existence of the equilibrium then requires

$$\theta^\dagger + (\theta^\dagger - 1) + \frac{R}{2} - R\psi\frac{1}{2} \geq \theta^\dagger + (\theta^\dagger - 1)\delta + \frac{R}{2} - R\psi\frac{\delta + \eta(\delta - 1)}{1 + \delta + \eta(\delta - 1)}$$
$$(\theta^\dagger - 1) \geq \frac{R\psi(1+\eta)}{2[1 + \delta + \eta(\delta - 1)]}$$

Substituting for $\theta^\dagger$, the previous inequality becomes (15). Hence, if reference dependence is not too important, the behavior described in the proposition is part of an equilibrium. To prove uniqueness, assume that $\eta \leq \delta/(2-\delta)$ and note that the incumbent's utility conditional on choosing $c = 1$ is increasing in $\underline{d}_1$ for any value of d. Since by Proposition 2 an incumbent with ideology $\theta_I \leq \theta^\dagger$ strictly prefers $(0,0)$ to $(1,d)$ for all $d \in [\delta, 1]$ when $\eta \leq \delta/(2-\delta)$ and $\underline{d}_1 = 1$, the same must be true when $\eta \leq \delta/(2-\delta)$ and $\underline{d}_1 < 1$. Furthermore, given any $\underline{d}_1 < 1$, an incumbent with ideology θ_I prefers $(1,\delta)$ to $(1,1)$ if and only if

$$\theta_I \leq 1 + R\psi \frac{1+\eta}{(2 + \eta - \eta\underline{d}_1)(1 + \delta + \delta\eta - \underline{d}_1\eta)}. \tag{16}$$

Since expression (15) implies that

$$(2 + \eta - \eta\underline{d}_1)(1 + \delta + \delta\eta - \underline{d}_1\eta) \geq 2(1 + \delta + \delta\eta - \eta) \geq 2(1 + \eta),$$

the right-hand side of (16) is below $\theta^\dagger = 1 + R\psi/2$. As a consequence, in equilibrium only incumbents with $\theta_I > \theta^\dagger$ choose $c = 1$, and when they do, they prefer $(1, 1)$ to $(1, \delta)$ (and, by convexity, to all $d \in (\delta, 1)$). This implies that, in equilibrium, $\underline{d}_1 = 1$. □

Proof of Proposition 4. The single crossing property of the incumbent's utility (see equation 11) implies that the level of escalation chosen by the incumbent must be non-decreasing in his ideology. The convexity of the incumbent's utility further implies the existence of the cutoffs introduced in the statement of the proposition. In particular ideology $\underline{\theta}$ makes the incumbent indifferent between not challenging and challenging and then choosing $d = \delta$. Similarly, ideology $\overline{\theta}$ makes the incumbent indifferent between challenging and then choosing not to escalate or challenging and then choosing full escalation. Hence, the expected level of escalation will be given by the expectation of d conditional on $c = 1$, namely conditional on $\theta_I \geq \underline{\theta}$. This yields (14). Furthermore, $\underline{\theta}$ satisfies

$$\delta(\underline{\theta} - 1) = \frac{R\psi[\delta(1+\eta) - \eta\underline{d}_1]}{1 + \delta(1+\eta) - \eta\underline{d}_1} \tag{17}$$

while $\overline{\theta}$ satisfies:

$$(\overline{\theta} - 1) = \frac{R\psi(1+\eta)}{[1 + 1 + \eta(1 - \underline{d}_1)][1 + \delta + \eta(\delta - \underline{d}_1)]}. \tag{18}$$

We then obtain that

$$\underline{d}_1 = 1 - (1-\delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} = \delta + (1-\delta)\frac{1 + 2(\tau - \overline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} \tag{19}$$

Obviously, this can be an equilibrium only if $\underline{\theta} \leq \overline{\theta}$ or equivalently

$$\frac{R\psi}{1 + \delta + \eta(\delta - \underline{d}_1)}\left[\eta\frac{\underline{d}_1}{\delta} - (1+\eta)\frac{1 + \eta(1 - \underline{d}_1)}{1 + 1 + \eta(1 - \underline{d}_1)}\right] \geq 0. \tag{20}$$

The first term in (20) is positive by Assumption 3; thus the sign of the left-hand side of (20) is equal to the sign of the squared bracket.

In the reminder of the proof, we show that the system of equations defined by (17)-(19) (i) has a solution, and (ii) all solutions are such that $\tau - \frac{1}{2\phi} < \underline{\theta} \leq \overline{\theta} < \tau + \frac{1}{2\phi}$.

By Assumption 2, there exist θ^l and θ^h with $\tau - \frac{1}{2\phi} < \theta^l < \theta^h < \tau + \frac{1}{2\phi}$ such that for all

possible $\pi(1,d)$, (i) for all $\theta_I < \theta^l$, $\arg\max_{\{0,1\}\times[\delta,1]} u_I(c,d;\theta_I) = (0,\delta)$ and (ii) for all $\theta_I > \theta^h$, $\arg\max_{\{0,1\}\times[\delta,1]} u_I(c,d;\theta_I) = (1,1)$. Hence, the solution of the system (17)-(19) is the fixed point of $\mathcal{F}(\underline{\theta},\overline{\theta},\underline{d}_1)$, which maps the set

$$[\theta^l,\theta^h]^2 \times \left[\delta + (1-\delta)\frac{1+2(\tau-\theta^h)\phi}{1+2(\tau-\theta^l)\phi}, \delta + (1-\delta)\frac{1+2(\tau-\theta^l)\phi}{1+2(\tau-\theta^h)\phi}\right]$$

into itself as follows

$$\mathcal{F}(\underline{\theta},\overline{\theta},\underline{d}_1) = \begin{bmatrix} \frac{1}{\delta}\frac{R\psi[\delta(1+\eta)-\eta\underline{d}_1]}{1+\delta(1+\eta)-\eta\underline{d}_1} + 1 \\ \\ \frac{R\psi(1+\eta)}{[1+1+\eta(1-\underline{d}_1)][1+\delta+\eta(\delta-\underline{d}_1)]} + 1 \\ \\ \delta + (1-\delta)\frac{1+2(\tau-\overline{\theta})\phi}{1+2(\tau-\underline{\theta})\phi} \end{bmatrix}$$

Since the mapping is continuous, Brouwer's Theorem ensures the existence of a fixed point. Suppose that the fixed point is such that $\underline{\theta} > \overline{\theta}$. Then expression (20) must fail, that is

$$\eta\frac{\underline{d}_1}{\delta} < (1+\eta)\frac{1+\eta(1-\underline{d}_1)}{1+1+\eta(1-\underline{d}_1)}. \tag{21}$$

Moreover, (19) implies that $\underline{d}_1 > 1$. $\underline{d}_1 > 1$, in turns, implies that (i) $\eta\frac{1}{\delta} < \eta\frac{\underline{d}_1}{\delta}$ and (ii) the right hand side of (21), being increasing in $1-\underline{d}_1$, is strictly smaller than $\frac{1+\eta}{2}$. Putting everything together yields

$$\eta\frac{1}{\delta} < \eta\frac{\underline{d}_1}{\delta} < (1+\eta)\frac{1+\eta(1-\underline{d}_1)}{1+1+\eta(1-\underline{d}_1)} < \frac{1+\eta}{2}.$$

which contradicts the premise of the proposition $\eta \geq \frac{\delta}{2-\delta}$. □

Proof of Proposition 5. The first statement follows from applying the implicit function theorem to the system:

$$\begin{aligned} \underline{\theta} - \frac{1}{\delta}\frac{R\psi[\delta(1+\eta) - \eta\underline{d}_1]}{1+\delta(1+\eta)-\eta\underline{d}_1} - 1 &= 0 \\ \overline{\theta} - \frac{R\psi(1+\eta)}{[1+1+\eta(1-\underline{d}_1)][1+\delta+\eta(\delta-\underline{d}_1)]} - 1 &= 0 \\ \underline{d}_1 - \delta - (1-\delta)\frac{1+2(\tau-\overline{\theta})\phi}{1+2(\tau-\underline{\theta})\phi} &= 0 \end{aligned}$$

Because Assumption 3 holds, $\partial\underline{\theta}/\partial\phi$ has the same sign of $R\psi\eta/[\delta(1+\delta(1+\eta)-\eta\underline{d}_1)^2]$, while $\partial\overline{\theta}/\partial\phi$ has the same of $-[R\psi\eta(1+\eta)(3+\delta(1+\eta)+\eta-2\eta\underline{d}_1)]/[(2+\eta-\eta\underline{d}_1)^2(1+\delta(1+\eta)-\eta\underline{d}_1)^2]$. As a result, the first derivative is positive and the second (again by Assumption 3) is negative.

To show the second statement, observe that Proposition 4 requires that (i)

$$\delta > \frac{\eta\underline{d}_1 - (1+2\psi)^{-1}}{1+\eta}$$

and (ii) $\eta \geq \frac{\delta}{2-\delta}$, or equivalently $\delta \leq \frac{2\eta}{1+\eta}$. In addition, some democrats become opportunistic authoritarians when (iii) $\underline{\theta} < 1$, that is, using equation (17), $\delta < \frac{\eta}{1+\eta}\underline{d}_1$. To prove the proposition, notice that as $\phi \to 0$, $\underline{d}_1 \simeq 1$. Then conditions (i) and (ii) can be combined into

$$\delta \in \left(\max\left\{0, \frac{\eta-(1+2\psi)^{-1}}{1+\eta}\right\}, \min\left\{1, \frac{2\eta}{1+\eta}\right\}\right],$$

while condition (iii) becomes $\delta < \frac{\eta}{1+\eta}$. By inspection,

$$\frac{\eta}{1+\eta} \in \left(\max\left\{0, \frac{\eta-(1+2\psi)^{-1}}{1+\eta}\right\}, \min\left\{1, \frac{2\eta}{1+\eta}\right\}\right].$$

As a consequence, when (i) and (ii) hold, the proposition holds as long as $\delta < \frac{\eta}{1+\eta}$, which is true if η is sufficiently high.[18] □

[18]Note that an excessively high η, however, may lead to the violation of condition (i) above. See Supplemental Appendix B (page 1) for details on what happens in this case.

References

Acharya, Avidit and Edoardo Grillo. 2019. "A Behavioral Foundation for Audience Costs." *Quarterly Journal of Political Science* 14(2):159–190.

Achen, Christopher H and Larry M Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government*. Vol. 4 Princeton University Press.

Alesina, Alberto and Francesco Passarelli. 2019. "Loss aversion in politics." *American Journal of Political Science* 63(4):936–947.

BBC. 2016. "Poland abortion: Parliament rejects near-total ban." *BBC News* . https://www.bbc.com/news/world-europe-37573938.

Bell, David E. 1985. "Disappointment in decision making under uncertainty." *Operations Research* 33(1):1–27.

Berman, Sheri and Maria Snegovaya. 2019. "Populism and the decline of social democracy." *Journal of Democracy* 30(3):5–19.

Bushong, Benjamin and Tristan Gagnon-Bartsch. 2019. "Learning with misattribution of reference dependence." *Working Paper* . https://scholar.harvard.edu/files/gagnonbartsch/files/misattribution_bushong_gagnonbartsch_jan31.pdf.

Callander, Steven and Catherine H. Wilson. 2006. "Context-dependent Voting." *Quarterly Journal of Political Science* 1(3):227–254.

Callander, Steven and Catherine H. Wilson. 2008. "Context-dependent voting and political ambiguity." *Journal of Public Economics* 92(3):565 – 581.

Carey, John, Gretchen Helmke, Mitchell Sanders, Katherine Clayton, Brendan Nyhan and Susan Stokes. 2020. "Who Will Defend Democracy? Evaluating Tradeoffs in Candidate Support Among Partisan Donors and Voters." *Working Paper* . https://preprints.apsanet.org/engage/apsa/article-details/5e8e2bc81e9caa0012076ec0.

Carnagey, Nicholas L, Craig A Anderson and Brad J Bushman. 2007. "The effect of video game violence on physiological desensitization to real-life violence." *Journal of experimental social psychology* 43(3):489–496.

Chiopris, Caterina, Monika Nalepa and Georg Vanberg. 2021. "A Wolf in sheep's clothing: citizen uncertainty and democratic backsliding." *Working Paper* . https://bit.ly/3djDcYZ.

Chrzczonowicz, Magdalena. 2017. "Już 42 Proc: Polakow Za Liberalizacja Ustawy Antyaborcyjnej. Nowy Sondaz OKO. press." *OKO. press* . https://bit.ly/38NwkBf.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Anna Luhrmann, Kyle L. Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Steven Wilson, Agnes Cornell, Nazifa Alizada, Lisa Gastaldi, Haakon Gjerløw, Garry Hindle, Nina Ilchenko, Laura Maxwell, Valeriya Mechkova, Juraj Medzihorsky, Johannes von Römer, Aksel Sundström, Eitan Tzelgov, Yi ting Wang, Tore Wig and Daniel Ziblatt. 2020. "V-Dem Dataset V10." *Varieties of Democracy (V-Dem) Project* . https://doi.org/10.23696/vdemds20.

Corazzini, Luca, Sebastian Kube, Michel André Maréchal and Antonio Nicolo. 2014. "Elections and deceptions: an experimental study on the behavioral effects of democracy." *American Journal of Political Science* 58(3):579–592.

Draca, Mirko and Carlo Schwarz. 2020. "How polarized are citizens? Measuring ideology from the ground-up." *Working Paper* . https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3154431.

Fearon, James D. 1999. Electoral accountability and the control of politicians: selecting good types versus sanctioning poor performance. In *Democracy, accountability, and representation*, ed. Adam Przeworski, Susan C Stokes and Bernard Manin. Cambridge University Press pp. 55–61.

Foa, Roberto Stefan and Yascha Mounk. 2016. "The danger of deconsolidation: The democratic disconnect." *Journal of democracy* 27(3):5–17.

Graham, Matthew H and Milan 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.

Gratton, Gabriele and Barton E Lee. 2020. "Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies." *UNSW Economics Working Paper 2020-13* .

Grillo, Edoardo. 2016. "The hidden cost of raising voters' expectations: Reference dependence and politicians' credibility." *Journal of Economic Behavior & Organization* 130:126–143.

Grossman, Guy, Dorothy Kronick, Matthew Levendusky and Marc Meredith. 2020. "The Majoritarian Threat to Liberal Democracy." *Working Paper* . `https://bit.ly/2ITsA6c`.

Grumbach, Jake. 2021. "Laboratories of Democratic Backsliding." *Working Paper* . `https://rb.gy/o3oxhp`.

Haggag, Kareem, Devin G Pope, Kinsey B Bryant-Lees and Maarten W Bos. 2018. "Attribution Bias in Consumer Choice." *The Review of Economic Studies* 86(5):2136–2183.

Hahl, Oliver, Minjae Kim and Ezra W. Zuckerman Sivan. 2018. "The Authentic Appeal of the Lying Demagogue: Proclaiming the Deeper Truth about Political Illegitimacy." *American Sociological Review* 83(1):1–33.

Hamilton, Alexander, James Madison and John Jay. 2008. *The federalist papers*. Oxford University Press.

Healy, Andrew and Neil Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of Political Science* 16(1):285–306.

Helmke, Gretchen, Mary Kroeger and Jack Paine. 2019. "Exploiting Asymmetries: A Theory of Democratic Constitutional Hardball." *Working Paper* . `https://www.gretchenhelmke.com/uploads/7/0/3/2/70329843/exlpoiting_asymmetries.pdf`.

Hiatt, Fred. 2019. "'It could have been worse' is the foundation of Trump's presidency." *The Washington Post* . https://wapo.st/2p8WM3K.

Howell, William G, Kenneth Shepsle and Stephane Wolton. 2019. "Executive Absolutism: A Model." *Working Paper* . https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3440604.

Howell, William G and Stephane Wolton. 2018. "The Politician's Province." *Quarterly Journal of Political Science* 13(2):119–146.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect theory: An analysis of decision under risk." *Econometrica* 47(2):363–391.

Karakas, Leyla D. and Devashish Mitra. 2021. "Electoral competition in the presence of identity politics." *Journal of Theoretical Politics* Forthcoming.

Kimball, David C and Samuel C Patterson. 1997. "Living up to expectations: Public attitudes toward Congress." *The Journal of Politics* 59(3):701–728.

Korolczuk, Elzbieta. 2016. "Explaining mass protests against abortion ban in Poland: the power of connective action." *Zoon politikon* 7:91–113.

Kőszegi, Botond and Matthew Rabin. 2006. "A model of reference-dependent preferences." *The Quarterly Journal of Economics* 121(4):1133–1165.

Kőszegi, Botond and Matthew Rabin. 2007. "Reference-dependent risk attitudes." *American Economic Review* 97(4):1047–1073.

Levitsky, Steven and Daniel Ziblatt. 2018. *How democracies die*. Broadway Books.

Levitsky, Steven and Maxwell Cameron. 2003. "Democracy without Parties? Political Parties and Regime Change in Fujimori's Peru." *Latin American Politics and Society* 45(3):1–33.

Lindstädt, René and Jeffrey K Staton. 2012. "Managing expectations." *Journal of Theoretical Politics* 24(2):274–302.

Lockwood, Ben and James Rockey. 2020. "Negative voters: Electoral competition with loss-aversion." *The Economic Journal* 130(632):2619–2648.

Luo, Zhaotian and Adam Przeworski. 2020. "Democracy and its Vulnerabilities: Dynamics of democratic backsliding." *Working Paper* . https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3469373.

Mair, Peter. 2002. Populist democracy vs party democracy. In *Democracies and the populist challenge*. Springer pp. 81–98.

Marcinkiewicz, Kamil and Mary Stegmaier. 2017. "Despite its anti-democratic agenda, Poland's Law and Justice party still enjoys public support." *Democratic Audit* . https://rb.gy/mgrof8.

Miller, Michael K. 2020. "A Republic, If You Can Keep It: Breakdown and Erosion in Modern Democracies." *Journal of Politics* p. forthcoming.

Panunzi, Fausto, Nicola Pavoni and Guido Tabellini. 2020. "Economic Shocks and Populism." *Working Paper* . http://www.igier.unibocconi.it/files/PPT_December_2019.pdf.

Passarelli, Francesco and Guido Tabellini. 2017. "Emotions and Political Unrest." *Journal of Political Economy* 125(3):903–946.

Pettigrew, Thomas F. 2017. "Social psychological perspectives on Trump supporters." *Journal of Social and Political Psychology* 5(1):107–116.

Popper, Karl. 1945. *The open society and its enemies*. Routledge.

Prato, Carlo and Stephane Wolton. 2018. "Rational ignorance, populism, and reform." *European Journal of Political Economy* 55:119–135.

Przeworski, Adam. 2019. *Crises of democracy*. Cambridge University Press.

Quattrone, George A. and Amos Tversky. 1988. "Contrasting Rational and Psychological Analyses of Political Choice." *American Political Science Review* 82(3):719–736.

Rosenblum, Nancy L. 2010. *On the side of the angels: an appreciation of parties and partisanship*. Princeton University Press.

Rosenbluth, Frances and Ian Shapiro. 2018. *Responsible Parties: Saving Democracy from Itself*. Yale University Press.

Schumpeter, Joseph A. 1942. *Capitalism, socialism and democracy*. Harper & Brothers.

Smith, David Norman and Eric Hanley. 2018. "The anger games: Who voted for Donald Trump in the 2016 election, and why?" *Critical Sociology* 44(2):195–212.

Stokes, Susan C. 1999. "Political parties and democracy." *Annual Review of Political Science* 2(1):243–267.

Svolik, Milan W. 2019. "Polarization versus Democracy." *Journal of Democracy* 30(3):20–32.

Urbinati, Nadia. 2019. *Me the People: How Populism Transforms Democracy*. Harvard University Press.

Versteeg, Mila, Timothy Horley, Anne Meng, Mauricio Guim and Marilyn Guirguis. 2019. "The Law and Politics of Presidential Term Limit Evasion." *Columbia Law Review* 2020.

Voeten, Erik. 2016. "Are people really turning away from democracy?" *Working Paper* . `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882878`.

Waldner, David and Ellen Lust. 2018. "Unwelcome change: Coming to terms with democratic backsliding." *Annual Review of Political Science* 21:93–113.

Waterman, Richard W, Hank C Jenkins-Smith and Carol L Silva. 1999. "The expectations gap thesis: Public attitudes toward an incumbent president." *The Journal of Politics* 61(4):944–966.

Woon, Jonathan. 2012. "Democratic Accountability and Retrospective Voting: A Laboratory Experiment." *American Journal of Political Science* 56(4):913–930.

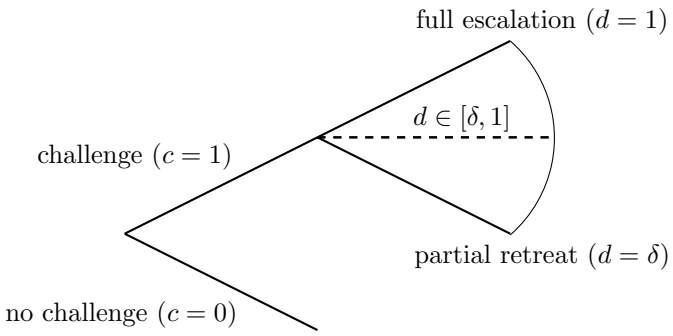

Figure 1: The Incumbent's choices. The Incumbent first chooses whether or not to challenge democratic norms, and then how much to double down against them.

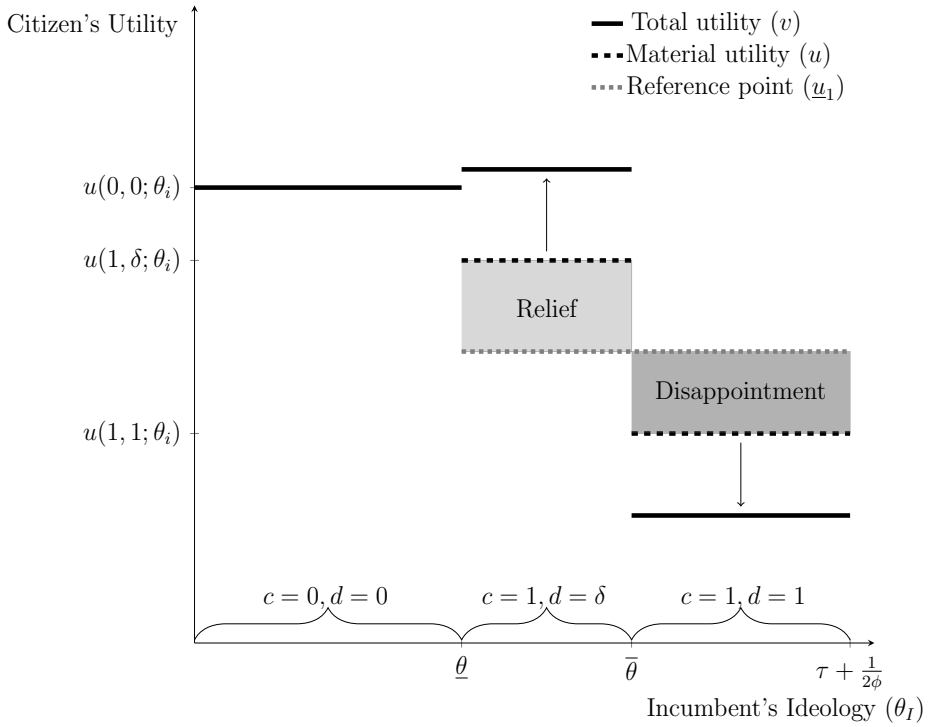

Figure 2: A citizen's material and psychological payoffs. Relative to respecting democratic norms, challenging and not doubling down reduces moderate citizens' ($\theta_i < 1$) material payoff, but improves their total payoff.

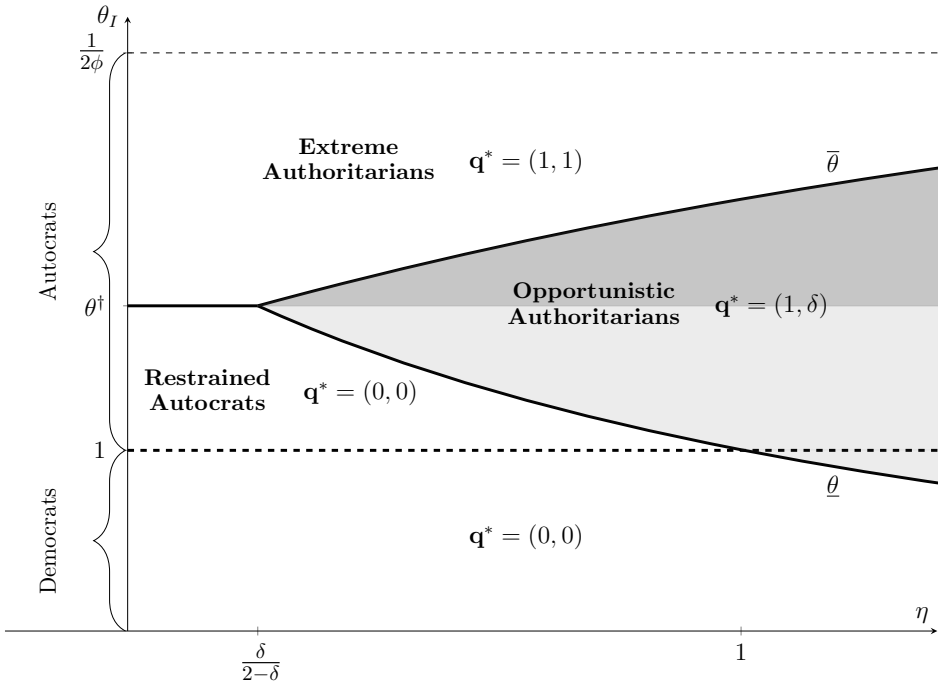

Figure 3: How incumbent equilibrium behavior varies with his ideology θ_I and the importance of reference dependence η (parameter values: $\psi = 0.2$, $\tau = 0.5$, $\phi = 0.25$, $R = 4$, $\delta = 0.35$).

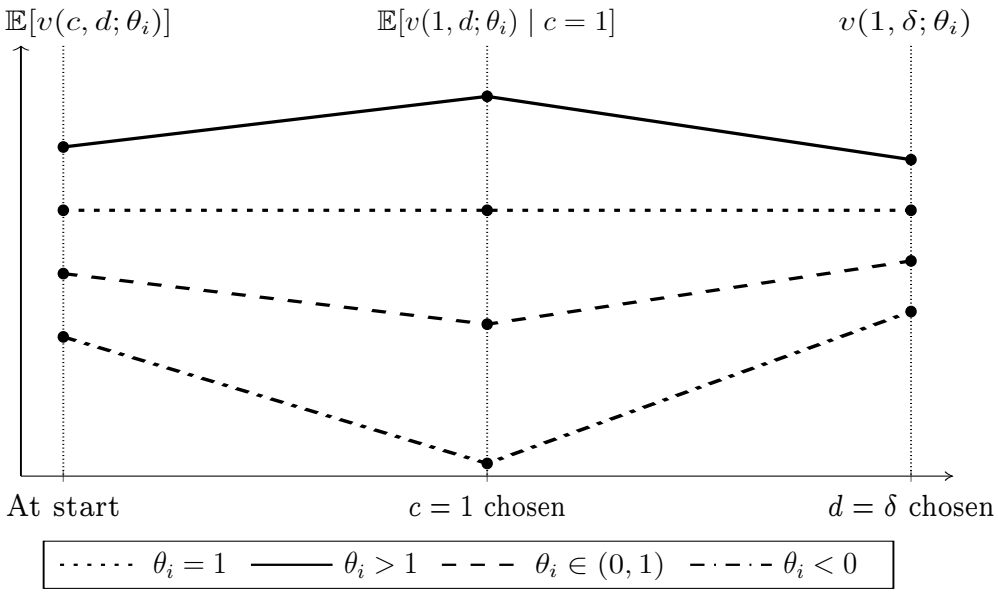

Figure 4: The dynamics of incumbent approval. Expected support among citizens of different ideologies before the incumbent's choice of c (left), after the choice of $c = 1$ (middle), and after the choice of $d = \delta$ (right).

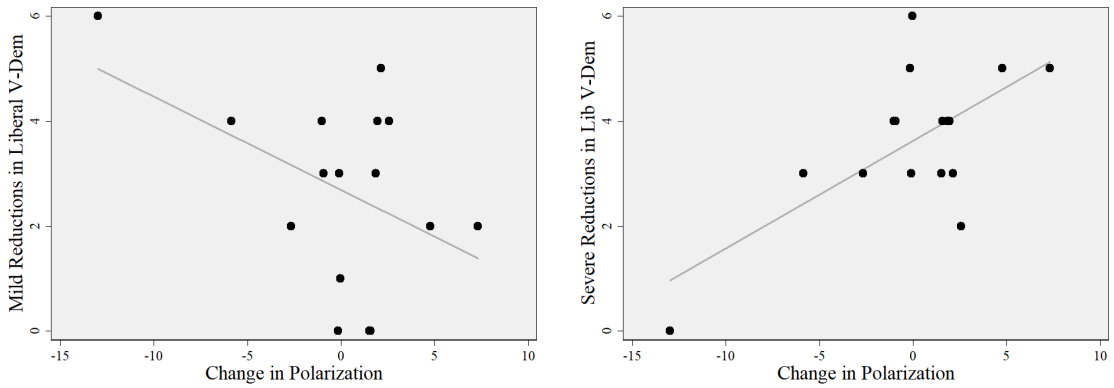

Figure 5: *Polarization* is the average of the eight (standardized) indexes of polarization computed by Draca and Schwarz using waves 4 (2000-2004) and 5 (2005-2009) of the World Value Survey. *Mild Reductions in V-Dem* is the number of below-median yearly reductions in the Liberal Democracy Index over the period 2009-2019. *Severe Reductions in V-Dem* is the number of above-median yearly reductions in the Liberal Democracy Index over the period 2009-2019. The sample includes Austria, Belgium, Canada, Denmark, Finland, France, UK, Germany, Iceland, Ireland, Italy, Malta, Netherlands, Portugal, Spain, and the US.

Choice variables	
$c \in \{0, 1\}$	Challenge decision
$d \in [\delta, 1]$	Escalation following a challenge
$\mathbf{q} = (c, d)$	Incumbent's behavior
$y(c, d) = 1 + cd$	Policy outcome
Parameters	
η	Weight on citizens' psychological utility
θ_i	Ideology of citizen i
R	Weight on support in incumbent's utility
θ_I	Ideology of the incumbent
ψ^{-1}	Mass Polarization
τ	Incumbent's average ideology
ϕ^{-1}	Programmatic uncertainty about the incumbent
δ	Weakness of institutional checks and balances
Functions	
$u(\mathbf{q}; \theta_i)$	Citizen's material utility
$\underline{u}(c; \hat{\mathbf{q}}, \theta_i)$	Citizen's reference point after c
$\underline{d}_c$	Expected escalation after c
$v(\mathbf{q}; \theta_i \mid \underline{u}) = u(\mathbf{q}; \theta_i) + \eta[u(\mathbf{q}; \theta_i) - \underline{u}]$	Citizen's total utility
$\pi(\mathbf{q})$	Support for the incumbent after $\mathbf{q}$
$u_I(\mathbf{q}; \theta_I) = u(\mathbf{q}; \theta_I) + R\pi(\mathbf{q})$	Incumbent's utility

Table 1: Summary of choice variables, parameters and functions.