ISSN 2279-9362

Collegio Carlo Alberto

Reference Points and Democratic Backsliding

Edoardo Grillo and Carlo Prato

No. 602
December 2019

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Reference Points and Democratic Backsliding*

Edoardo Grillo† Carlo Prato‡

January 31, 2020

Abstract

In recent years, the behavior of elected leaders has raised concerns about the state of liberal democracy. Yet, evidence shows that electorates remain largely committed to democratic norms. We propose a theory of democratic backsliding where citizens evaluate incumbent performance relative to a reference point that is endogenous to the incumbent's behavior. We show that democratic backsliding can occur even when most citizens *and most politicians* share a primitive aversion to it. By challenging norms of democracy, incumbents can lower citizens' reference points—only to partially back down and beat this lowered standard. As a result, gradual backsliding actually enhances an incumbent's popular support. We show that this mechanism can only arise when, owing to programmatically weak parties, citizens are uncertain about leaders' policy preferences. Conversely, mass polarization and citizens' information have an ambiguous effect on the occurrence of backsliding.

*We are grateful to Avi Acharya, Sheri Berman, Peter Buisseret, Jon Eguia, Teresa Esteban-Casanelles, Tim Frye, Diego Gambetta, John Huber, Giovanna Invernizzi, Kimuli Kasara, John Marshall, Anne Meng, Monika Nalepa, Jacopo Perego, Chiara Superti, Michael Ting and Stephane Wolton, and seminar and conference participants at UC San Diego for their helpful comments and conversations. This paper was previously circulated under the title "Opportunistic Authoritarians, Reference-Dependent Preferences, and Democratic Backsliding."

†Collegio Carlo Alberto. Email: edoardo.grillo@carloalberto.org

‡Columbia University. Email: cp2928@columbia.edu

1. Introduction

In the summer of 2019, after withdrawing his party from the cabinet in which he was serving as Italy's interior 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 authored an executive order effectively denying asylum seekers access to public services. When those provisions were struck by the courts, he accused the judges of left-wing and pro-migrant bias and ordered an audit of their public comments to force their recusal. He also threatened to remove police protection from a journalist who criticized him. While the verdicts and his critic's security detail eventually remained in place, these attempts to weaken judicial independence and silence the media brought substantial gains in the polls.

Salvini's actions are hardly exceptional. From the U.K. Prime Minister's prorogation of the Parliament to the U.S. President's attempts to stonewall congressional oversight, from the forced retirement of judges in Poland to the purges of public employees in Turkey, scholars and observers are increasingly concerned about democratic backsliding (Waldner and Lust, 2018; Levitsky and Ziblatt, 2018; Przeworski, 2019): democratically elected leaders use the power of government to undermine constraints of accountability on their actions.

These features are at odds with extensive observational and experimental evidence that most voters all else equal dislike challenges to democratic norms (Voeten, 2016; Graham and Svolik, 2019). Thus, even if electoral institutions have lost part of their sanctioning power over politicians' autocratic ambitions, we should expect challenges to democratic norms to reduce politicians' public support—not improve it.

In this paper, we propose a theory of democratic backsliding where challenges to democracy can arise even when most citizens and most incumbents share a common intrinsic aversion to them. We study the emergence of *opportunistic authoritarians*—incumbents who attack democratic institutions to enhance their popularity. The theory provides a unified framework to analyze the two leading explanations for backsliding identified by the literature: mass polarization (Svolik, 2019; Nalepa, Vanberg and Chiopris, 2018) and the programmatic weakening of political parties (Rosenbluth and Shapiro, 2018; Berman and Snegovaya, 2019).

Our results suggest that weak parties are necessary for the emergence of opportunistic authoritarians, while mass polarization can actually reduce the likelihood of backsliding.

Our theory is built on the premise that (i) voters 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 citizens' assessment of the incumbent is not based solely on an absolute standard—her performance in office—, but also on the comparison with 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 expectation, the support decreases.

The idea of context-dependent preferences has a long history in social and behavioral sciences and a large body of evidence showing its importance in various settings: attitudes towards the legislative (Kimball and Patterson, 1997), the executive (Waterman, Jenkins-Smith and Silva, 1999) and democratic institutions (Corazzini et al., 2014), but also labor markets (Farber, 2008), sports (Pope and Schweitzer, 2011), gambling behavior (Lien and Zheng, 2015), contractual environments (Fehr, Hart and Zehnder, 2011).

In our setting, citizens form their reference point after the incumbent's choice of challenging democratic norms, but before the choice of doubling down.[1] Thus, voters' reference point responds to incumbent's behavior. In particular, if citizens believe, based on his early actions, that the incumbent will engage in some serious dismantling of democratic norms, they become pessimistic and, as a result, their reference point will be low. If the incumbent then ends up not (or only partially) doubling down, his performance will exceed the reference point, thereby improving his popularity. Owing to voters' reference-dependent preferences,

[1]This timing in the formation of the reference point is crucial, but also quite natural. See the discussion in Section 3.

an incumbent can challenge democratic norms *and* enjoy substantial support not *despite* citizens' aversion to democratic backsliding, but *precisely because* of it.

Citizens' expectations are not arbitrary. In line with a rational-expectation approach (Kőszegi and Rabin, 2007), the behavior of the incumbent affects the citizens' reference point through their correct conjectures about his future (equilibrium) behavior. As a result, we obtain sharp predictions about the emergence of opportunistic authoritarians.

We show that citizens' uncertainty about the incumbent's ideology increases the likelihood of democratic backsliding. We relate these conditions to the documented disintermediation of political representation by political parties: challenging democracy is a more viable strategy when citizens' expectations about leaders' future actions are no longer anchored to parties' programmatic identities and the fact-based reporting of traditional media outlets.

In that respect, our theory underscores the importance of intermediation by parties and media—and how their weakening in recent years led to populist authoritarianism (Mair, 2002; Rosenblum, 2010). Indeed, going back to our initial example, Salvini's tenure as leader of the *Lega Nord* party coincided with a large shift from the party's traditional platform, which emphasized regional autonomy, anti-clericalism and economic freedom to a more general nationalist message, often directly broadcast from Salvini's own social media accounts.

Our theory does not suggest that democratic backsliding *always* improves an incumbent's popularity. Responsiveness to public opinion (henceforth, political responsiveness) does rein in the authoritarian impulses of truly autocratic incumbents: as in Svolik (2019) and Nalepa, Vanberg and Chiopris (2018), citizens sanctions severe violations of democratic norms (i.e., challenging and doubling down). As a result, in addition to opportunistic authoritarians, our model also produces *restrained autocrats*, namely incumbents who, despite their true ideology, do not attack democratic norms because they value citizens' support.

The contemporaneous presence of opportunistic authoritarians and retrained autocrats complicates the relationship between the occurrence of democratic backsliding and several key factors identified in the literature. For instance, we show that mass polarization and a less informed citizenry can actually decrease the likelihood of democratic backsliding (but not its expected severity). The reason is that mass polarization and worse information both

decrease the responsiveness of citizen support to incumbent behavior. This mitigates *both* the disciplining effect of public opinion that motivates restrained autocrats *and* the incentive to manipulate citizens' reference points that drives opportunistic authoritarians.

In addition to these novel empirical implications, our theory provides a mechanism that simultaneously accounts for citizens' intrinsic commitment to democracy documented by Voeten (2016), their increased dissatisfaction with democratic governance (Foa and Mounk, 2016), and the simultaneous popularity of leaders who gradually erode democratic norms observed in Turkey, Poland, Hungary, and—on a smaller scale—in United States and other Western democracies. In the absence of strong ideological and programmatic commitments, these rulers simply try to cling to power, gradually lower the expectations of large segments of the electorate without ever disappointing them and, this way, consolidate their support.

2. Related Literature

Our paper contributes to two main strands of literature: one on the causes of democratic backsliding, the other on context-dependent preferences in formal political theory.

After several decades of spreading and consolidation of liberal democracy, in recent years scholars had to contend with a reversal in these trends. Democratic backsliding refers to violations of the constraints that limit the executive's ability to use the tools of government (Waldner and Lust, 2018). They encompass the breaking of traditionally respected norms, outright violations of the law, and more nuanced testing of its boundaries. Examples include the circumvention of term limits (Versteeg et al., 2019) and the forced expansion of executive authority, often referred to as executive absolutism (Howell and Wolton, 2018; Howell, Shepsle and Wolton, 2019) or constitutional hardball (Helmke, Kroeger and Paine, 2019).

Recent explanations for democratic backsliding focus on two phenomena: the rise of polarization and the weakening of political parties. A first line of literature formally and experimentally shows how growing mass polarization leads to fewer voters sanctioning violations of democratic norms (Nalepa, Vanberg and Chiopris, 2018; Luo and Przeworski, 2019; Graham and Svolik, 2019; Carey et al., 2019; Miller, 2020). As an explanation for backsliding, these

theories require that (i) many democratically elected incumbents must have "authoritarian ambitions" (Svolik, 2019) in pursuit of which they are willing to sacrifice popular support and (ii) that changes in backsliding should track changes in mass polarization across a number of countries. Our theory does not require either premise. Importantly, it can also account for voters' intrinsic aversion to violations *and the increased support for violators*.

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). Deep societal changes (increase in income dispersion, immigration, and the importance of social media) have stifled parties' ability to intermediate between government and society (Stokes, 1999; Rosenblum, 2010), thereby resulting in voter confusion. This paper provides a formalization of the mechanism through which voter confusion can end up boosting support for opportunistic political entrepreneurs with authoritarian stances.

Our model also contributes to the formal literature on context-dependent preferences in political science (Callander and Wilson, 2006, 2008). In particular, in our model individuals evaluate outcomes based not only on absolute standards, but relative to their expectations—captured by a reference point (see Kahneman and Tversky, 1979 and Bell, 1985, for seminal contributions).[2] Specifically, this paper follows the work of Kőszegi and Rabin (2006, 2007, 2009), where the reference point is endogenously determined in equilibrium given the behavior and related expectations of players. A smaller but growing literature pioneered by Lindstädt and Staton's (2012) reduced-form approach, applies the idea of reference dependence to international relations (Acharya and Grillo, 2019), electoral competition (Alesina and Passarelli, 2019; Lockwood and Rockey, 2019), and political campaigns (Grillo, 2016).

[2]For axiomatized models of reference-dependence, see Gul (1991), Sugden (2003) or Ok, Ortoleva and Riella (2015).

3. Baseline Model

A polity is composed of a unit mass of citizens indexed by i ("she"), and is ruled by an incumbent I ("he"). The incumbent chooses whether or not to challenge democracy, then chooses how much to double down on his initial challenge. More aggressive doubling down results in more drastic policy changes. Citizens form their retrospective assessment of the incumbent and choose whether or not to support him.

First, I chooses whether to challenge democratic norms ($c = 1$, for example announcing a *prima facie* unconstitutional measure, or that judicial review will be ignored, or that minority rights will be restricted) or not ($c = 0$).[3] Subsequently, he chooses a policy y from the interval $\mathcal{Y}(c)$, which corresponds to how much he will double down on his initial challenge. Challenging democratic norms expands the range of policy outcomes available to the incumbent. For simplicity, we assume that $\mathcal{Y}(0) = 1$: if he chooses not to challenge democratic norms, I's subsequent policy choice will be constrained to $y = 1$. Conversely, $\mathcal{Y}(1) = [1 + \delta, 2]$: if he challenges, more severe doubling down will allow him to achieve more drastic policy change. The choice variable $d \in [\delta, 1]$ captures the severity of the escalation following a challenge. $y(c, d) = 1 + cd$ then denotes the policy outcome as a function of the incumbent's actions.

When $d = 1$, the incumbent chooses full escalation. When $d = \delta$, the incumbent chooses no further escalation (i.e., to partially back down). The parameter $\delta \in (0, 1)$ captures the strength of institutional checks and balances—how much push-back the incumbent receives from other institutions (e.g., the judiciary system, or independent agencies) after his initial attempt to force constitutional boundaries. The lower δ, the lower is the erosion of democratic norms that the incumbent achieves without further escalating his challenge of democratic norms.[4]

[3] In Appendix B.4 we show that the binary nature of the decision to challenge is without loss of generality provided that voters dislike more severe violations of democratic norms increasingly more, which is compatible with the model we present in this Section.

[4] The assumption that challenging democracy can expand the set of achievable policy outcomes is substantively important. It also captures the idea that authoritarian backsliding

Figure 1 below, summarizes the incumbent's sequential decision problem.

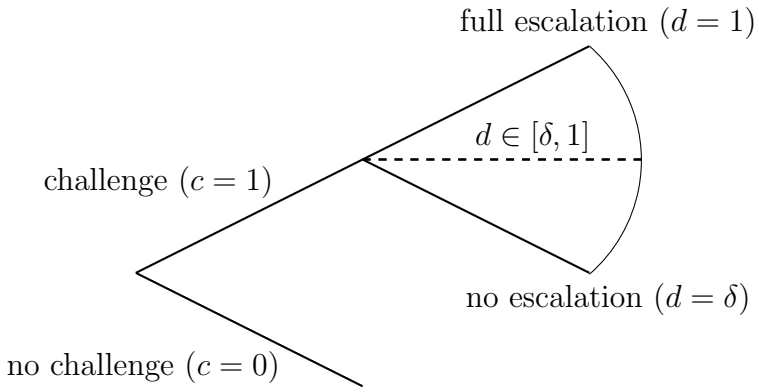


Figure 1: Incumbent's Sequential Decision Problem.

Citizens vary in their policy preferences and share a common intrinsic aversion for challenges against democratic institutions (see, e.g., Graham and Svolik, 2019; Carey et al., 2019). Each citizen i evaluates policy outcomes $y(c, d)$ in light of her ideology θ_i, reflected in the payoff $\theta_i y(c, d)$. We assume that θ_i is distributed in the population according to a cumulative density function F. Citizens with a positive (negative) ideology favor (oppose) policy changes. The intrinsic aversion to democratic norms, instead, is captured by the payoff $-cd$. Let $\mathbf{q} = (c, d)$ be the outcome of the incumbent's behavior. Then, citizens' *material utility* is given by $\theta_i y(c, d) - cd$, that is

$$u(\mathbf{q}; \theta_i) = \theta_i(1 + cd) - cd. \tag{1}$$

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.[5] Hence, we interpret a reduc-

is a gradual process that often starts with institutional reforms (e.g., *de jure* or *de facto* weakening of the judicial or other independent authorities), which, if successful, paves the way to more extreme policy measures. However, our results would continue to hold if we allowed full reversibility ($\delta = 0$).

[5]The specific distributional assumption is made for analytic tractability. Our results would extend to other distributions as long as the density f is flat enough.

tion in ψ as an increase in mass polarization. Assumption 1 implies that most citizens are intrinsically averse to democratic backsliding, while the upper bound on ψ guarantees the existence of a minority of autocratic citizens who support full escalation against democracy.[6]

Like citizens, the incumbent experiences a policy payoff and has 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 incumbent's support, i.e. the share of citizens who support him, and $R \in \mathbb{R}_+$ measures the importance of support (for example, the strength of her electoral concern). The incumbent knows his ideology θ_I, but the citizens do not. Their uncertainty is captured by the (common) cumulative density function F_I.

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. The assumption that $\tau < 1$ implies that most incumbents are against authoritarian backsliding. The upper bound on ϕ, instead, ensures that some incumbent types are immune to public opinion, namely their behavior is entirely driven by their policy payoff (see footnote 13 for a discussion of what happens when this latter assumptions fails).

Once the incumbent has chosen the policy vector $\mathbf{q}$, citizens decide whether to support him or not (e.g., voting in his favor, albeit there are other situations in which incumbents value citizens support). In the baseline model, these evaluations are purely retrospective. 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

[6]The assumption also simplifies the exposition by ensuring that the share of citizens supporting the incumbent is always interior.

parameter $\eta \in \mathbb{R}_+$ captures the relative importance of this psychological component relative to a citizen's 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, 2009), we assume that the reference point is determined endogenously: it is equal to the citizen's expected utility *following the incumbent's decision to challenge or not*. Formally, let the behavior of the incumbent be summarized by a strategy $\theta_I \mapsto \hat{\mathbf{q}}(\theta_I) = (\hat{c}(\theta_I), \hat{d}(\theta_I))$. Then, the reference point of a citizen with ideology θ_i when she observes c is equal to:

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

Thus, the incumbent's decision to challenge democratic norms has two consequences: (i) it changes the set of policy choices available to him, and (ii) it triggers a thought process among citizens about the ultimate consequences of the incumbent's actions, which leads to the formation of their reference point.

An equilibrium is a profile $(\hat{\mathbf{q}}, \underline{u}(0; \hat{\mathbf{q}}, \theta_i), \underline{u}(1; \hat{\mathbf{q}}, \theta_i))$ that specifies a sequentially rational strategy $\hat{\mathbf{q}}$ for each incumbent's type and a reference point for each observed choice of c. The equilibrium reference points are endogenous objects possessing the fixed-point structure typical of rational expectations: on the one hand, the reference point affects support—and thus the behavior of the incumbent—, on the other hand, the behavior of the incumbent feeds back into the reference point.

3.1 Discussion

Before proceeding with the analysis, we briefly discuss two important assumptions of the model: how the the reference point is formed and how citizens evaluate the incumbent.

We assume that citizens form their reference point after the incumbent's decision to challenge democracy but before the final choice of how much to double down. If citizens' reference

point was entirely determined before the incumbent's actions, it could not respond to the incumbent's behavior. If citizens' reference point was entirely determined after the incumbent's actions, citizens' material payoff will always coincide with their reference point, thereby leaving no room for exceeding or falling short of expectations.[7] In line with Acharya and Grillo (2019), we assume that the reference point is *entirely* determined after the incumbent's first choice. Our results, however, would be qualitatively unaffected if we assumed that the reference point was a weighted average between (i) an ex-ante exogenous standard, (ii) the expected payoff determined after the choice of c and (iii) the final payoff determined after the choice of d by the incumbent.

In line with experimental (Woon, 2012) and empirical (for a review, see Healy and Malhotra, 2013) evidence, the baseline model also assumes that citizens' assessment of the incumbent are purely *retrospective* Yet, in light of an influential critique of retrospection inmodels of electoral accountability (Fearon, 1999), it is important to show that our results extend to situations in which citizens' evaluation are also *prospective*: their support for the incumbent depends on their conjectures about the incumbent's *future* performance. Appendix B.2 shows that opportunistic authoritarians also emerge in an extension in which a citizen's assessment of the incumbent is strictly decreasing in their perceived ideological distance.

4. Analysis

Given how retrospective evaluations are formed, a citizen with ideology θ_i supports the incumbent if and only if $v(\mathbf{q};\theta_i) \geq 0$.[8] The incumbent's support is thus equal to

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

[7]The model would then be identical to one where citizens do not display reference-dependent preferences.

[8]The specific way in which citizens break an indifference does not affect the analysis. Also, the threshold of zero is without loss of generality and our results would be unchanged if zero was replaced by a constant $\underline{v}$.

In our model, incumbent behavior is driven by two sets of concerns: (i) policy concerns, i.e., how his behavior affects his policy utility, and (ii) popularity concerns, i.e., how his behavior affects his support. Popularity concerns, in turn, respond to two distinct mechanisms: (a) how the incumbent's behavior affects citizens' material payoff and (b) how it affects citizens' *psychological payoff*. To clearly 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 the psychological payoff determined by reference dependence ($R > 0$ and $\eta = 0$), and we finally describe the novel incentives that reference dependence generates.

4.1 The Incumbent's Policy Concerns

When $R = 0$, the incumbent's behavior does not respond to public opinion. In the absence of political responsiveness, the incumbent simply maximizes his policy utility $\theta_I + cd\theta_I - cd$. When θ_I exceeds one, the value of a more extreme policy exceeds the loss from weakening democratic norms, so the incumbent chooses $c = 1$ and then fully doubles down on this initial challenge ($d = 1$). Conversely, when θ_I is below one, the incumbent prefers not to violate constitutional boundaries and sets $c = 0$.

Since challenges to democratic institutions are initiated only by incumbents with $\theta_I > 1$—who then fully escalate—we refer to these types as *autocrats*. Conversely, we refer to incumbents with type $\theta_I \leq 1$ as emphdemocrats.We summarize this discussion in the next proposition. (Proofs of all formal statements are in Appendix A.)

Proposition 1. *Suppose that the incumbent is not office-motivated ($R = 0$). Then,*
(i) if the incumbent is a democrat ($\theta_I \leq 1$), then $c = 0$ and $y(c, d) = 1$;
(ii) if the incumbent is an autocrat ($\theta_I > 1$), then $c = 1$ and $y(c, d) = 2$.

4.2 Popularity Concerns without Reference Dependence

Now, suppose that the incumbent is office motivated ($R > 0$), but citizens do not exhibit reference dependence ($\eta = 0$). In this case, popularity concerns are entirely driven by

citizens' material payoffs. Only citizens with $u(\mathbf{q};\theta_i) \geq 0$ will support the incumbent. Since a majority of citizens oppose authoritarian backsliding (i.e., for a majority of citizens $u(\mathbf{q};\theta_i) = \theta_i + (\theta_i - 1)cd$ is decreasing in both c and d), challenges to democratic norms necessarily reduce the incumbent's popular support. When the incumbent respects democratic norms, his support equals $\pi(0,0) = F(\theta_i \geq 0) = 1 - F(0) = \frac{1}{2}$. When he chooses to challenge them, more citizens abandon him, and this loss in popular support is *increasing* in the level of subsequent escalation:

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

We can then write down the incumbent's payoff as a function of his choices of c and d:

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

Since democratic backsliding entails a drop in popular support, all liberal types choose to respect democratic norms. Autocratic types, conversely, face a trade off, since democratic backsliding increases their policy utility. As a result, only autocratic types that are extreme enough (*extreme authoritarians*) will choose to violate norms, and will then double down in full.[9]

Conversely, citizens act as a check on autocratic leaders with less extreme ideologies. The threat of losing support generates a measure of *restrained autocrats*—autocratic types that are induced to respect democratic norms. These are essentially the same driving forces described in the existing formal literature on democratic backsliding (Svolik, 2019; Nalepa, Vanberg and Chiopris, 2018). This idea has deep roots: it directly links to key argument for the centrality of electoral institutions in a democratic regime (Schumpeter, 1942; Popper, 1945): by institutionalizing the contingency of a ruler's power on popular support, elections inoculate societies from unpopular governance outcomes.

[9]Because the loss in support is concave on the level of escalation, conditional on challenging these norms, they will choose full escalation.

Restrained autocrats are those for which $\theta_I > 1$ and $u_I(1,1;\theta_I) \leq u_I(0,0;\theta_I)$, namely $\theta_I \in (1, \theta^\dagger]$ with

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

Proposition 2. *Suppose the Incumbent is office-motivated ($R > 0$), but citizens do not exhibit reference dependence ($\eta = 0$). Then,*
(i) $c = 1$ if and only if the incumbent's autocratic tendencies are strong enough, i.e., $\theta_I > \theta^\dagger$, in which case $d = 1$;
(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). Crucially, this restraining effect is increasing with the importance of support (R) and decreasing with citizens' ideological dispersion. In line with Nalepa, Vanberg and Chiopris (2018) and Svolik (2019), mass polarization limits citizens' responsiveness and thus reduces the drop in support associated with democratic backsliding.

While Proposition 2 is consistent with the notion that democratic backsliding unfolds over time, it also predicts that incumbents should always double down on their challenges, which is at odds with the prevailing accounts of how democratic backsliding proceeded in Venezuela, Turkey, Poland and Hungary—where attacks were often followed by sudden retreats and significant setbacks.

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 some form of democratic backsliding and (iii) affects non-trivially the way in which factors such as mass polarization and citizen information acquisition affect incumbent's behavior.

4.3 Reference Dependence and Opportunistic Authoritarians

We now consider the case in which an office-motivated incumbent ($R > 0$) faces citizens who exhibit reference dependence ($\eta > 0$). As discussed above, reference points are determined by citizens expectations following the incumbent's decision on whether to challenge or not, $\underline{u}(0;\hat{\mathbf{q}},\theta_i)$ and $\underline{u}(1;\hat{\mathbf{q}},\theta_i)$—which in equilibrium are correct. Given the structure of our game

and the linearity of utilities with respect to policy choices, these expectations are fully identified by the expected level of escalation given the initial choice of c:

$$\underline{u}(c;\hat{\mathbf{q}},\theta_i) = \begin{cases} \theta_i & c=0 \\ \theta_i + (\theta_i - 1)\mathbb{E}[\hat{d} \mid c=1] & c=1 \end{cases}.$$

where $E[\hat{d} \mid c=1] := \underline{d}_1 \in [\delta, 1]$.

If the incumbent chooses not to escalate (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

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

Instead, if the incumbent challenges democratic institutions, citizens' behavior depends on the expected level of escalation, $\underline{d}_1$, which is determined in equilibrium. In particular, 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

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

In the body of the paper, we assume that institutional checks and balances are not too strong. (In the Appendix, we provide a complete characterization and show that the assumption below effectively stacks the deck against our main result.)

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

$$\delta > \frac{\eta - 1/2}{1+\eta} \tag{10}$$

Substantively, the assumption guarantees that in equilibrium a citizen's propensity to support the incumbent after a challenge is increasing in her ideology.

Hence, when $c = 1$ the Incumbent's support is interior and equals

$$\pi(1, d) = \frac{1}{2} - \psi \frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}. \tag{11}$$

This support is strictly decreasing and strictly convex in d: because citizens (on average) dislike democratic backsliding, doubling down on democratic institutions entails a loss in support that gets increasingly higher as the escalation d goes up. Substituting for the support in the Incumbent's utility, we get

$$u_I(1, d; \theta_I) = (\theta_I - 1)d + R\left[\frac{1}{2} - \psi \frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}\right]. \tag{12}$$

Notice that $\pi(1, d)$ is not necessarily lower than $\pi(0, 0) = \frac{1}{2}$. The reason is that that citizens' reference point might go down following the incumbent's decision to challenge democratic norms. Comparing (8) and (12), we can identify the potential trade-off faced by an incumbent when he decides whether to challenge or not. If he challenges institutions, he might shift the policy (which he likes if $\theta_I > 1$), but this also entails a loss in popular support.

$$u_I(1, d; \theta_I) - u_I(0, 0; \theta_I) = \underbrace{(\theta_I - 1)d}_{Policy\ \ Drift} - \underbrace{R\psi \frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}}_{Public\ \ Opinion\ \ Feedback} \tag{13}$$

Due to the policy drift, a challenge to democratic norms allows the incumbent to set a more extreme policy, but it also changes citizens' retrospective evaluation through a feedback mechanism: it lowers the policy payoff of most citizens, but it also shifts their reference point—from θ_i to $\theta_i - (1 - \theta)\underline{d}_1$.

Then we can consider two different cases, depending on the relative importance of reference dependence in determining citizens' utility.

Proposition 3. *Suppose reference dependence has little impact on citizens' utility:*

$$\eta < \frac{\delta}{2 - \delta}.$$

Then, the incumbent's equilibrium behavior is unique and identical to the one described in Proposition 2.

In the settings covered by Proposition 3,[10] all incumbents who challenge democratic institutions also fully escalate. Because citizens' reference point is determined in equilibrium, $\underline{d}_1 = 1$. Hence, citizens are do not experience any disappointment or relief and the cutoff type $\overline{\theta}_I$, who is indifferent between challenging and not challenging, is still $\theta^\dagger$. As a result, the equilibrium utility of the incumbent is then equal to:

$$u_I^\star(\theta_I) = \begin{cases} \theta_I + \frac{R}{2} & \text{if } \theta_I < \theta^\dagger \\ 2\theta_I - 1 + \frac{R}{2}\left[1 - \psi\right] & \text{if } \theta_I \geq \theta^\dagger \end{cases} \tag{14}$$

To understand why this equilibrium requires reference dependence not to be too important for citizens' utility, note that conditional on challenging, 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 and thus increases the incumbent's support. This behavior is particularly tempting for autocratic incumbents with less extreme ideologies, namely incumbents with θ_I close to $\theta^\dagger$.

Inequality $\eta < \delta/(2-\delta)$ guarantees exactly that type $\theta^\dagger$ strictly prefers to play according to the equilibrium strategy rather than to reap the benefits associated with the above described fear-and-relief mechanism. Indeed, when reference dependence has a small impact on citizens' utility (i.e., when η is low), the extent of citizens' relief and the resulting increase in support is limited. Thus, incumbents do not engage in this strategic behavior.[11]

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

[11]Note that the cutoff for η is increasing in δ. This is intuitive: as the strength of institutional checks and balances decreases (i.e. δ increases), the extent of the positive surprise that the incumbent can generate decreases as well. Hence, the strategic behavior becomes less profitable and the incumbent will not engage in it also for relatively high values of η.

Now, consider the case in which the importance of reference dependence is not too low, $\eta > \delta/(2-\delta)$. Convexity of the incumbent's utility function with respect to d implies that if challenges occur in equilibrium, incumbents will either choose not to escalate further, $(d = \delta)$, or full escalation $(d = 1)$. Moreover, because the incumbent's utility satisfies the single crossing condition, the level of escalation chosen by the incumbent must be weakly increasing in his ideology. Hence, the following proposition holds.

Proposition 4. *Suppose that reference dependence is important enough:*

$$\eta \geq \frac{\delta}{2-\delta}. \tag{15}$$

Then, we can identify two levels of ideology $\underline{\theta}$ *and* $\overline{\theta}$ *such that*
(i) $c = 1$ *if and only if* $\theta_I > \underline{\theta}$
(ii) $d = \delta$ *if* $\theta \in (\underline{\theta}, \overline{\theta}]$ *and* $d = 1$ *if* $\theta_I > \overline{\theta}$.
In this equilibrium, the citizens' reference point following a challenge is given by

$$\underline{d}_1 = 1 - (1-\delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} \tag{16}$$

The behavior of the incumbent in this equilibrium is then characterized by cutoffs $\underline{\theta}$ and $\overline{\theta}$ that are jointly determined in equilibrium together with $\underline{d}_1$—see equations (A-2) and (A-3) in the Appendix. In particular, $\overline{\theta}$ is always greater than 1, while $\underline{\theta}$ can also be lower than 1.

Opportunistic authoritarians. Proposition 4 implies that the behavior of incumbents with ideology in the interval $(\underline{\theta}, \overline{\theta}]$ is driven by the interaction between reference dependence and political incentives. In particular, compared to Proposition 2 and 3, types in the interval $(\theta^\dagger, \overline{\theta}]$ do not double down on the initial challenge because they want to benefit from the increase in support that voters' relief generate. In other words, reference dependence strengthens the disciplining effect of public opinion and limits the severity of the attack against democracy.

However, incumbents in the interval $(\underline{\theta}, \theta^\dagger]$ end up challenging democratic institutions even though they would have respected them in the absence of reference dependence. The logic is similar, but opposite in sign, to the one highlighted above: if voters are expecting the

incumbent to escalate, and this does not happen—i.e., the incumbent chooses $(c, d) = (1, \delta)$—citizens experience relief and support increases with respect to the one associated with no challenge. In this case, reference dependence weakens the disciplining effect of popularity concerns (e.g., electoral incentives) and increases the likelihood of democratic backsliding.

When the ideology of the incumbent is sufficiently uncertain (i.e., ϕ small), incumbents with extreme ideologies are substantially likely. Hence, a citizen who observes a challenge will expect full escalation with high probability (i.e., $\underline{d}_1 \simeq 1$). When this happens, the increase in support associated with citizens' relief may push even democratic incumbents to challenge democracy, i.e. $\underline{\theta} < 1$. Importantly, and paradoxically, when this last phenomenon occurs, stronger political responsiveness (measured either as an increase in the relative importance of popular support, R, or in the responsiveness of citizens' behavior to their realized payoff, ψ) may lead to a further decrease in $\underline{\theta}$.

Thus, in our model, electoral incentives may encourage some democratic incumbents to behave in an authoritarian manner. This goes against not only their intrinsic preferences, but also the interests of citizens. We summarize this discussion in the next proposition

Proposition 5. *There exists $\phi^* \in \mathbb{R}$, such that if $\phi < \phi^*$ and reference dependence is important enough, opportunistic authoritarians also include some democratic incumbents ($\underline{\theta} < 1$).*

Proposition 5 implies that democratic politicians may become opportunistic authoritarians only if (i) reference dependence is sufficiently strong and (ii) citizens are sufficiently uncertain about politicians' intrinsic policy positions. In practice, this uncertainty can be reduced by strong political parties (which can "certify" their leaders' programmatic commitments) and a robust, independent media system. Our results then provide a formalization to the idea that the weakening of the intermediation by parties and media is a key prerequisite for populist authoritarianism (Mair, 2002; Rosenblum, 2010). It also highlights a natural complementarity between democratic backsliding and populism—defined as a governing strategy based on a direct, unmediated relation between a leader and "the people." Indeed, one could interpret a transition to less mediated communication as an increase in the ability of the incumbent to affect citizens' utility through the reference-dependent component. In Appendix B.3, we

confirm this intuition in a more rigorous way by allowing citizens to rationally choose their level of attention (or inattention) toward the incumbent's behavior.

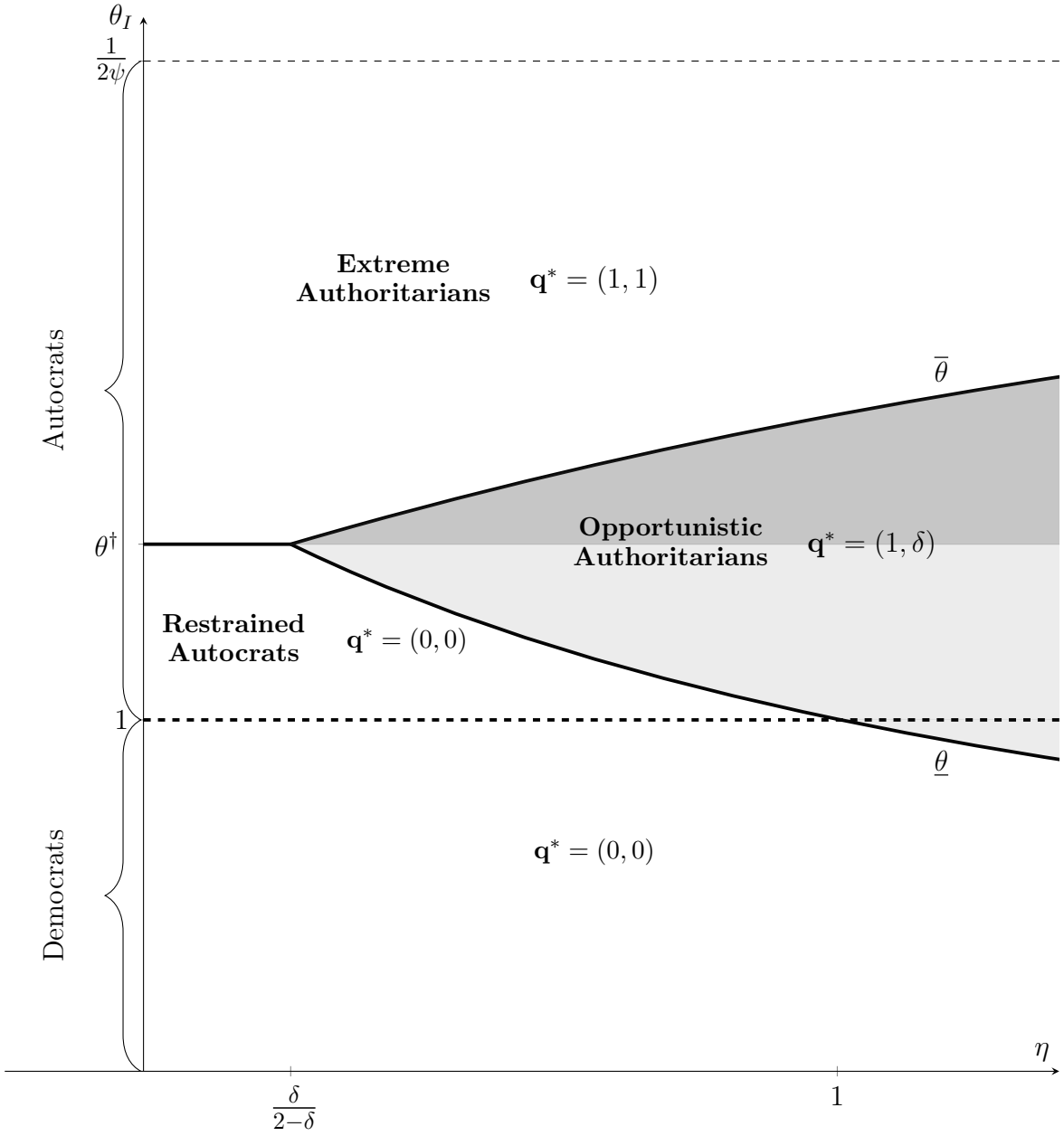


Figure 2: Incumbent's equilibrium behavior as a function of his type θ and the importance of reference dependence η (parameter values $\psi = 0.2$, $\tau = 0.5$, $\phi = 0.25$, $R = 4$ and $\delta = 0.35$).

Figure 2 below summarizes the incumbent's equilibrium behavior under the assumptions.[12] If the importance of reference dependence is sufficiently low (i.e., if $\eta \leq \frac{\delta}{2-\delta}$) the equilibrium behavior of the incumbent is identical to the case of no reference dependence (cf. Proposition 3). Only autocrats with sufficiently high ideology ($\theta_I > \theta^\dagger$) challenge democratic norms (and they fully escalate), while autocrats with less extreme ideology ($\theta_I \in (1, \theta^\dagger]$) are restrained and behave as democrats.

However, if the importance of reference dependence is sufficiently large (i.e., if $\eta > \frac{\delta}{2-\delta}$), a subset of relatively moderate autocrats with ideology in the interval $(\underline{\theta}, \overline{\theta}]$ finds it optimal to challenge democratic norms and then refrain from escalating in order to enhance his support using the fear-and-relief mechanism described above (cf. Proposition 4): opportunistic authoritarians emerge. When compared to the case of low reference dependence, this choice has two implications. On the one hand, some extreme authoritarians are partially disciplined: incumbents with ideology between $\theta^\dagger$ and $\overline{\theta}$ (highlighted in dark gray in Figure 2) choose not to escalate ($d = \delta$) instead of full escalation ($d = 1$). On the other hand, previously restrained autocrats are encouraged to behave in a more authoritarian way: incumbents with ideology in $(\theta^\dagger, \underline{\theta}]$ (highlighted in light gray in Figure 2) begin to challenge democratic norms. As the importance of reference dependence keeps increasing, restrained autocrats disappear and some democratic incumbents—those with ideology in the interval $(\underline{\theta}, 1]$—challenge democracy because they want to gain from citizens' relief (cf. Proposition 5). When this happens, changes in the incumbent's incentives generate subtle and counter-intuitive comparative static results.

[12]Recall that Assumption 3 puts an upper bound on η. See Section B.1 for a characterization of the equilibrium when Assumption 3 fails.

5. Implications

5.1 Challenges without Doubling Down

Proposition 4 implies that opportunistic authoritarians emerge if reference dependence is sufficiently important. The behavior of these politicians differ from the one described in Proposition 3 in an important dimension: when a challenge occurs, it does not lead to full escalation. This is because the fear-and-relief mechanism described above motivates autocrats on both sides of $\theta^\dagger$ to challenge democracy and then partially back down. Hence, reference dependence strengthens the disciplining effect of popular control for autocrats with ideology above $\theta^\dagger$ (e.g., the interval $(\theta^\dagger, \overline{\theta}]$) expands) and weakens it for those with ideology below $\theta^\dagger$ (i.e., the interval $(\underline{\theta}, \theta^\dagger]$ also expands). When $\underline{\theta} < 1$, reference dependence leads to a complete reversal of such disciplining effect.[13]

The effect of mass polarization. Previous scholarship has singled out mass polarization as a key enabling factor of democratic backsliding (Nalepa, Vanberg and Chiopris, 2018; Svolik, 2019). The logic is that in a highly polarized environment, citizens' voting decision are relatively unresponsive to the behavior of incumbents, who can then try to short-circuit democratic norms to achieve their policy goals with relative impunity. While not entirely contradicting this idea, mass polarization plays a more subtle role in our theory.

When either (i) reference dependence is sufficiently weak, or (ii) citizens are not too uncertain about the incumbent's policy positions (i.e., ϕ is not too small), mass polarization increases the likelihood of democratic backsliding. The reason is that higher mass polarization (i.e., lower ψ) reduces the punishment associated with violating democratic norms,

[13] At the other extreme of the ideology spectrum, absent any restriction on parameters, the incentive to choose an escalation level $d = \delta$ (as opposed to $d = 1$) can be so strong that even the incumbent with the highest ideology ($\tau + \frac{1}{2\phi}$) chooses this action. Under this scenario, which is ruled out by Assumption 2, $\underline{d}_1 = \delta$ and the incumbent who is indifferent between choosing $c = 0$ or choosing $c = 1$ (and then $d = \delta$) would have ideology $1 + \frac{\psi R}{1+\delta} > \theta^\dagger$. Hence, even when Assumption 2 does not hold, opportunistic authoritarians arise.

thereby reducing incumbents' accountability: fewer autocrats are deterred from triggering democratic backsliding.

However, when reference dependence is strong enough and citizens are sufficiently uncertain about the incumbent's policy positions, (i..e, the assumptions of Proposition 5 hold), higher mass polarization reduces the likelihood of opportunistic authoritarians.[14] The reason is that weakening citizens' responses to incumbent behavior, mass polarization reduces politicians' incentives to try to lower citizens' expectations (and then profit from their relief) by challenging democratic norms. Hence, when opportunistic authoritarians arise, mass polarization decreases the overall likelihood of democratic backsliding *and* it increases its severity conditional on occurring (i.e., it increases the likelihood of full escalation conditional on a challenge occurring).

5.2 Institutional Checks and Balances

Our model also illustrates how the strength of institutional checks and balances (i.e., lower δ) affects the occurrence of democratic backsliding. Conventional wisdom—traced 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. While our model generally confirms this intuition, it also cautions about the limitations of this protection.

To highlight this implication, we focus on the more innovative part of our theory: the case in which reference dependence is sufficiently strong (i.e., when Proposition 4 holds).[15] The first consequence of stronger checks and balances is that challenges to democracy are in expectation less damaging: conditional on incumbents not doubling down, citizens are better off as δ goes down.

[14]Formally, lower ψ pushes both $\underline{\theta}$ and $\overline{\theta}$ closer to 1. This implies that, if opportunistic authoritarians exist, an increase in mass polarization reduces their likelihood.

[15]This case requires relatively strong checks and balances, so this part of our theory is more likely to apply to relatively more advanced democracies.

Proposition 6 in Appendix A, however, shows that checks and balance also affect the likelihood and intensity of these challenges. Also in line with the conventional wisdom (yet, through a novel mechanism), stronger checks and balances generally increase the disciplining effect of popular control and lower the likelihood of full escalation: the relief that citizens experience when an incumbent backs down from a challenge is higher—and so is the resulting gain in citizens' support.[16] However, for the very same reason—and contrary to the conventional wisdom—, stronger checks and balances also encourage democrats to become opportunistic authoritarians and, as a result, they also increase the likelihood of democratic backsliding.

5.3 Rational Inattention

In Appendix B.3, we analyze a simplified extension of the model with rationally inattentive citizens. Specifically, we assume that citizens can choose their level of attention to politics, which in turn increases the probability that they observe the incumbent's actions (we impose general assumptions on the exact way in which this happens)

Holding the behavior of the incumbent constant, attention is always valuable for the citizen: it improves her ability to estimate the ex-post payoff from supporting the incumbent, and thus improves her choice. However, citizen attention also feeds back into incumbents' incentives, and its effect is crucially mediated by reference dependence. Generally speaking, more attention strengthens the relationship between incumbent behavior and public opinion, similar to a decrease in polarization. The importance of reference dependence governs how this increased responsiveness shape incumbent behavior, but its overall effect on the citizen's *ex-ante* payoff is ambiguous. On the one hand, higher attention increases the likelihood that the citizen detects and punishes severe democratic backsliding (i.e., a challenge followed by doubling down). On the other hand, attention increases the likelihood that an incumbent

[16]Notice that δ also affects the reference point: holding incumbents' strategies fixed, higher δ increases the reference point, thereby partially offsetting the gain from backing down. This effect, however, is second order because it vanishes as ϕ approaches zero.

who challenges but does not escalate will manage to lower citizens' expectations and benefit from the boost in support identified in Proposition 4.[17]

The extension implies that increased availability of information (and attention to politics) can be a double-edged sword: while providing a stronger protection against the authoritarian tendencies of autocratic incumbents, very attentive citizens generate stronger incentive for opportunistic authoritarians and creates space for a gradual erosion of democratic norms. As in Prato and Wolton (2016), the best-case scenario for electoral incentives are "Goldilocks voters" who pay some attention—but not *too much* attention—to politics.

6. Conclusion

This paper presents a theory of democratic backsliding in which most citizens and most incumbents intrinsically dislike violations of democratic norms and yet, these violations do not always reduce popular support.

When (i) citizens are not too uncertain about incumbent's intrinsic policy preferences or (ii) the standard to which they evaluate them is not too responsive to politicians' initial actions, the implications of the theory about the role of mass polarization, the strength of checks and balances, and citizen information mirror conventional scholarly wisdom as well as the insights of a more recent formal theoretical literature. When instead these conditions fail, a lot of these insights are almost flipped on their heads, and they help reconcile some otherwise puzzling empirical patterns in politicians' behavior and citizens' attitudes: even if most citizens intrinsically dislike democratic backsliding, challenging norms of democracy allows incumbents to effectively moving 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.

[17]Notice that this result creates a potential benefit from information avoidance. The channel, however, is distinct from previously documented results that rely, e.g., on anticipatory utility (Kőszegi, 2006).

References

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

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

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.

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. 2019. "Who Will Defend Democracy? Evaluating Tradeoffs in Candidate Support Among Partisan Donors and Voters." *Unpublished manuscript* .

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.

Farber, Henry S. 2008. "Reference-dependent preferences and labor supply: The case of New York City taxi drivers." *American Economic Review* 98(3):1069–82.

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.

Fehr, Ernst, Oliver Hart and Christian Zehnder. 2011. "Contracts as reference points—experimental evidence." *American Economic Review* 101(2):493–525.

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

Graham, Matthew and Milan Svolik. 2019. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *Partisanship, Polarization, and the Robustness of Support for Democracy in the United States (March 18, 2019)* .

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

Gul, Faruk. 1991. "A theory of disappointment aversion." *Econometrica* 59(3):667–686.

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.".

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

Howell, William G, Kenneth Shepsle and Stephane Wolton. 2019. "Executive Absolutism: A Model." *Available at SSRN 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.

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

Kőszegi, Botond. 2006. "Emotional agency." *The Quarterly Journal of Economics* 121(1):121–155.

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.

Kőszegi, Botond and Matthew Rabin. 2009. "Reference-dependent consumption plans." *American Economic Review* 99(3):909–36.

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.

Lien, Jaimie W and Jie Zheng. 2015. "Deciding when to quit: Reference-dependence over slot machine outcomes." *American Economic Review* 105(5):366–70.

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

Lockwood, Ben and James Rockey. 2019. "Negative voters: Electoral competition with loss-aversion." *Economic Journal* Forthcoming.

Luo, Zhaotian and Adam Przeworski. 2019. "Subversion by Stealth: Elementary Dynamics of Democratic Backsliding." *Unpublished manuscript* .

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

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

Nalepa, Monika, Georg Vanberg and Caterina Chiopris. 2018. "Authoritarian Backsliding." *Unpublished manuscript, University of Chicago and Duke University* .

Ok, Efe A, Pietro Ortoleva and Gil Riella. 2015. "Revealed (p) reference theory." *American Economic Review* 105(1):299–321.

Pope, Devin G and Maurice E Schweitzer. 2011. "Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes." *American Economic Review* 101(1):129–57.

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

Prato, Carlo and Stephane Wolton. 2016. "The voters' curses: why we need Goldilocks voters." *American Journal of Political Science* 60(3):726–737.

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.

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.

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

Sugden, Robert. 2003. "Reference-dependent subjective expected utility." *Journal of Economic Theory* 111(2):172–191.

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?" *Available at SSRN 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.

A. Appendix: Proofs

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 $d = 1$. On the contrary, 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 θ_I 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 $c = 1$, the incumbent 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 $(1,\delta)$. In other words, $d = \delta$ 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 $\theta_I < 1 + R\psi/2$ will choose $(c,d) = (0,0)$, while those with ideology $\theta_I > 1 + R\psi/2$. Incumbents with ideology $\theta_I = 1 + R\psi/2$ are indifferent between choosing $(0,0)$ or $(1,1)$ and we assume without loss of generality that they choose $(0,0)$. □

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

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

Following the reasoning of the proof of Proposition 2, we can conclude that the behavior described in the proposition is an equilibrium as long as the incumbent $\theta^\dagger$ prefers to play $d = 1$, rather than $d = \delta$ even though this latter action would generate a positive surprise equal to $(1 - \delta)$. In other words, the existence of the equilibrium 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:

$$\eta \leq \frac{\delta}{2 - \delta}, \tag{A-1}$$

Hence, if reference dependence is not too important, the behavior described in the proposition is part of an equilibrium. To prove that such behavior is the unique one compatible with equilibrium, 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. Hence, if $\underline{d}_1 < 1$ and $\eta \leq \delta/(2 - \delta)$, an incumbent with ideology $\theta^\dagger$ strictly prefers $(0, 0)$ to $(1, d)$ for any $d \in [\delta, 1]$. Furthermore, given any $\underline{d}_1 < 1$, an incumbent with ideology θ_I prefers $(1, \delta)$ to $(1, 1)$ if and only if

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

Since expression (A-1) 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 the previous inequality is below $\theta^\dagger = 1 + R\psi/2$, we conclude that $(\delta, 1)$ is not optimal for any incumbent. Therefore, $\underline{d}_1$ cannot occur in equilibrium if $\eta \leq \delta/(2-\delta)$. □

Proof of Proposition 4. The single crossing property of the incumbent's utility (i.e., Equation 12) implies that the level of escalation chosen by the incumbent must be increasing in her 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

$\theta_I \geq \overline{\theta}$. This yields (16). Furthermore, $\underline{\theta}$ satisfies

$$\delta(\underline{\theta} - 1) = \frac{R\psi[\delta(1+\eta) - \eta\underline{d}_1]}{1 + \delta(1+\eta) - \eta\underline{d}_1} \tag{A-2}$$

while $\overline{\theta}$ satisfies:

$$(\overline{\theta} - 1) = \frac{R\psi(1+\eta)}{[1+1+\eta(1-\underline{d}_1)][1+\delta+\eta(\delta-\underline{d}_1)]}. \tag{A-3}$$

In this case, we immediately get that

$$\underline{d}_1 = 1 - (1-\delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} = \delta + (1-\delta)\frac{1 + 2(\tau - \overline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} \tag{A-4}$$

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

$$\frac{R\psi}{1 + \delta + \eta(\delta - \underline{d}_1)}\left[\eta\frac{\underline{d}_1}{\delta} - (1+\eta)\frac{1+\eta(1-\underline{d}_1)}{1+1+\eta(1-\underline{d}_1)}\right] \geq 0. \tag{A-5}$$

The first term in (A-5) is positive by Assumption 3; thus the sign of (A-5) is determined by the sign of the squared bracket.

In the reminder of the proof, we show that the system of equations defined by (A-2)-(A-4) (i) has a solution, and (ii) all solutions are such that $\tau - \frac{1}{2\phi} < \underline{\theta} \leq \overline{\theta} < \tau + \frac{1}{2\phi}$.

Also notice that 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 (A-2)-(A-4) 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

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

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 (A-5) must fail, that is

$$\eta\frac{\underline{d}_1}{\delta} < (1+\eta)\frac{1+\eta(1-\underline{d}_1)}{1+1+\eta(1-\underline{d}_1)}. \tag{A-6}$$

Moreover, (A-4) 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 (A-6), 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. Proposition 4 requires that (i) $\delta > d_h^\circ(\underline{d}_1)$, or

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

and (ii) $\eta \geq \frac{\delta}{2-\delta}$, or

$$\delta \leq \frac{2\eta}{1+\eta}.$$

In addition, some democrats become opportunistic authoritarians when (iii) $\underline{\theta} < 1$, that is, using equation (A-2),

$$\delta < \frac{\eta}{1+\eta}\underline{d}_1.$$

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] □

Proposition 6. *When ϕ is small enough,*
(i) $\overline{\theta}$ is strictly decreasing in δ
(ii) when $\underline{\theta} < 1$, $\underline{\theta}$ is strictly increasing in δ.

Proof of Proposition 6. As ϕ approaches zero, the reference point $\underline{d}_1$ approaches one. In this case δ affects the thresholds $\underline{\theta}$ and $\overline{\theta}$ only via its direct effect. The first result then follows

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

by inspection of equation (A-3). To prove the second result, observe that differentiating $\underline{\theta}$ in (A-2), yields

$$\frac{\partial \underline{\theta}}{\partial \delta} \propto \frac{\partial}{\partial \delta}\left(\frac{1+\eta-\eta\frac{\underline{d}_1}{\delta}}{1+\delta(1+\eta)-\eta\underline{d}_1}\right) \propto -[\delta(1+\eta)-\eta\underline{d}_1]^2+\eta\underline{d}_1 \approx -[\delta(1+\eta)-\eta]^2+\eta \quad \text{(A-7)}$$

Assumption 3 requires that $\delta(1+\eta)-\eta > -1/2$ and the fact that $\underline{\theta} < 1$ requires that $\delta(1+\eta)-\eta < 0$. Together, they imply $-[\delta(1+\eta)-\eta]^2 > -\eta+\delta(1+\eta)$, which implies that (A-7) is positive. □

B. Supplemental Appendix (for Online Publication)

B.1 General Characterization

In the main text we analyzed the model assuming that institutional checks and balances are sufficiently strong, namely under Assumption 3. This guarantees that a challenge to democratic norms yields a sizable move toward extreme policies. In this Section we show that our qualitative insights extend to settings in which this is not the case.

To this goal, let $d^\circ(\underline{d}_1) \equiv (\eta \underline{d}_1 - 1)/(1+\eta)$ and recall the definition of $v(\mathbf{q}; \theta_i \mid \underline{u})$ in (3). If $d > (<)\, d^\circ(\underline{d}_1)$, (9) implies that $v(\mathbf{q}; \theta_i \mid \underline{u})$ is increasing (decreasing) in θ_i. Instead, if $d = d^\circ(\underline{d}_1)$, $v(\mathbf{q}; \theta_i \mid \underline{u}) = 1$ and thus the support for the incumbent is equal to 1.[19] By continuity, we can then define an interval around $d^\circ(\underline{d}_1)$ such that when d falls in this interval, then the support of the incumbent is equal to 1. To characterize the support of the incumbent, let

$$\theta^*(d, \underline{d}_1) = \min\left\{\max\left\{\frac{d + \eta(d - \underline{d}_1)}{1 + d + \eta(d - \underline{d}_1)}, -\frac{1}{2\psi}\right\}, \frac{1}{2\psi}\right\}. \tag{B-1}$$

Assumption 1 implies that $\theta^*(1, \underline{d}_1) \in (0, 1/(2\psi))$. Further define $d^\circ_\ell(\underline{d}_1)$ to be the smallest solution of $\theta^\star(d, \underline{d}_1) = 1/(2\psi)$, namely

$$d^\circ_\ell(\underline{d}_1) = \frac{\eta \underline{d}_1 - (1 - 2\psi)^{-1}}{1 + \eta}, \tag{B-2}$$

Similarly, define $d^\circ_h(\underline{d}_1)$ to be the largest solution of $\theta^\star(d, \underline{d}_1) = -1/(2\psi)$, namely

$$d^\circ_h(\underline{d}_1) = \frac{\eta \underline{d}_1 - (1 + 2\psi)^{-1}}{1 + \eta}. \tag{B-3}$$

Then, the following proposition holds.

Proposition B.1. *Let $\underline{d}_1$ be the reference point of the citizens. Then, the incumbent's support is equal to*

$$\pi(1, d \mid \underline{d}_1) = \begin{cases} \frac{1}{2} + \psi\theta^\star(d, \underline{d}_1) & d < d^\circ_\ell(\underline{d}_1); \\ 1 & d \in [d^\circ_\ell(\underline{d}_1), d^\circ_h(\underline{d}_1)]; \\ \frac{1}{2} - \psi\theta^\star(d, \underline{d}_1) & d > d^\circ_h(\underline{d}_1). \end{cases} \tag{B-4}$$

[19] Note that we are ignoring the constraint $d \geq \delta$. This is irrelevant for the discussion that follows.

The support is strictly increasing and strictly convex in d in the interval $[\delta, d_\ell^\circ(\underline{d}_1)]$ and strictly decreasing and strictly convex in d in the interval $[d_h^\circ(\underline{d}_1), 1]$. Finally, the incumbent's utility, u_I, is also convex on d.

Proof. The first part of the statement follows from (9) and (B-1). Instead, the properties of $\pi(1, d \mid \underline{d}_1)$ follow from observing that $\theta^\star(d, \underline{d}_1)$ is increasing and concave in d when $d > d_\ell^\circ(\underline{d}_1)$, increasing and strictly convex in d if $d < d_h^\circ(\underline{d}_1)$ and constant in d in the interval $[d_h^\circ(\underline{d}_1), d_\ell^\circ(\underline{d}_1)]$. Hence, $\pi(1, d, y \mid \underline{d}_1)$ is strictly increasing and strictly convex in d in the interval $[0, d_\ell^\circ]$. Instead, it is strictly decreasing and strictly convex in d in the interval $[d_h^\circ, 1]$. The convexity of u_I with respect to d follows from the fact that the policy-related component of the incumbent's utility is linear in d for any θ_I. □

In light of Proposition B.1, Assumption 3 in the main text restricts attention to the case in which the support of the incumbent is decreasing (and convex) in δ, namely the case in which $d > d_h^\circ(\underline{d}_1)$ for any $\underline{d}_1$.

Differently from the case analyzed in the main text, if $d \leq d_\ell^\circ(\underline{d}_1)$, the incumbent's support is increasing in the level of extremism. To understand why, observe that when the incumbent chooses policies that are not too extreme, all citizens with low ideology will support him. Instead, citizens with high ideology will not because they would rather pick higher values of d; thus, if d increases, some of these citizens will support the incumbent yielding an increase in his support.

Because the level of escalation is bounded below by δ and Proposition B.1 holds, if we fix expectations at $\underline{d}_1$, the optimal behavior (c^*, d^*) of any incumbent with ideology different from $\theta_I = 1$ belongs to a finite set, D^*.[20] Depending on the value of δ, D^* is one of three possible sets. Figure B-1 depicts the set of citizens supporting the incumbent (shaded area), function $\theta^*(d, \underline{d}_1)$ (solid black line) and the possible equilibrium levels of escalation (black dots) in each of these three cases.

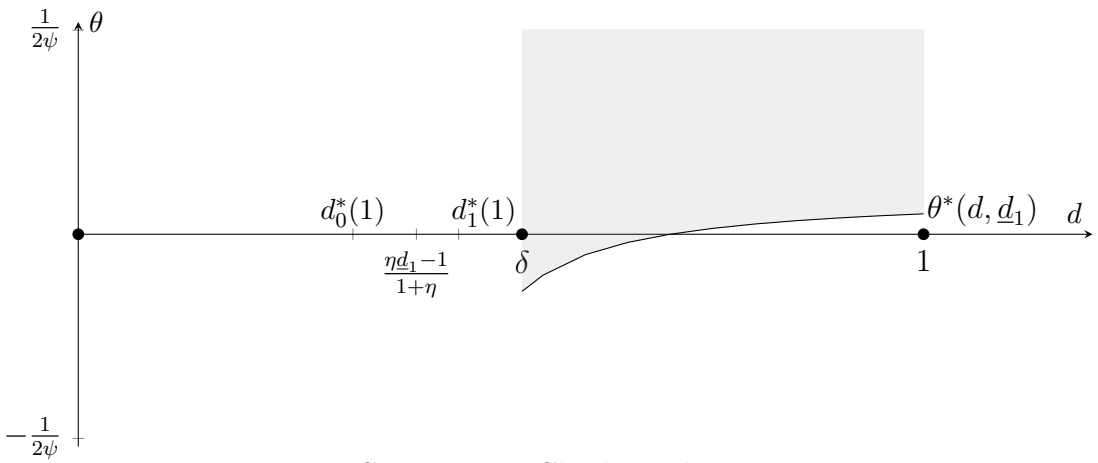
Case 1. If $\delta > d_h^\circ$, then $D^* = \{(0,0), (1,\delta), (1,1)\}$.

Case 2. If $\delta \in (d_\ell^\circ, d_h^\circ]$, then $D^* = \{(0,0), (1,\delta), (1,d_h^\circ), (1,1)\}$.

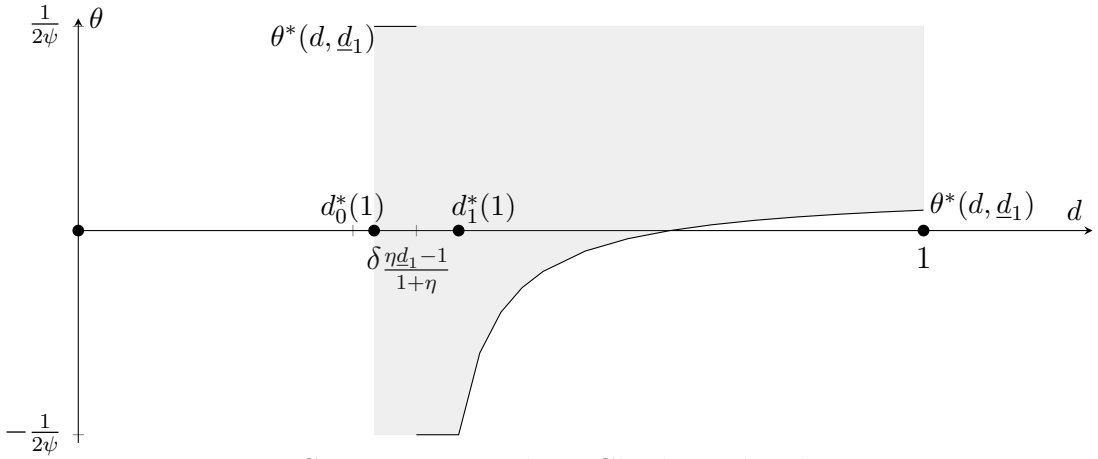
Case 3. If $\delta \leq d_\ell^\circ$, then $D^* = \{(0,0), (1,\delta), (1,d_\ell^\circ), (1,d_h^\circ), (1,1)\}$.

When δ is sufficiently large (i.e., $\delta \geq (2\eta - 1)/(2(1+\eta))$), (B-3) implies that $\delta > d_h^\circ(\underline{d}_1)$ independently of δ and of the citizens' expectations. Hence, the relevant case is the first one, which is analyzed in the main text. We will now consider the other two possible cases.

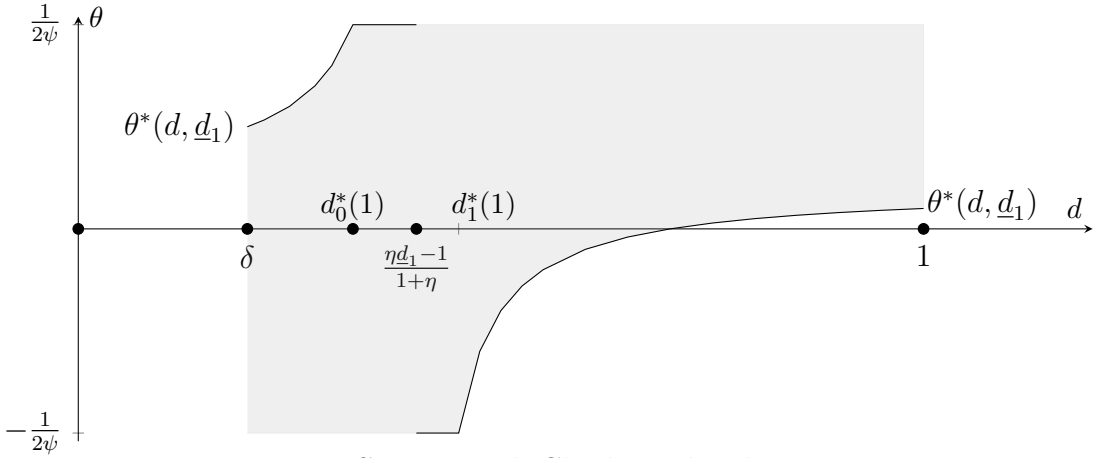
[20] An incumbent with ideology equal to $\theta_I = 1$ may have a continuum of optimal strategies. However, because these types have measure zero, we can ignore them in what follows.



Case 1: Low Checks and Balances



Case 2: Intermediate Checks and Balances



Case 3: High Checks and Balances

Figure B-1: Support for the Incumbent as a function of d in the three cases when $\underline{d}_1 = 1$, $\psi = 0.1$, $\eta = 7/3$ and $\delta = 0.525$ (Case 1), $\delta = 0.35$ (Case 2), $\delta = 0.2$ (Case 3)

Suppose we are in case 2, thus $\delta \in (d_\ell^\circ, d_h^\circ]$. Abstracting from popular support, Incumbents with ideology equal to 1 would be indifferent between all levels of escalation. Indeed all such

levels of escalation of $\delta \in [\delta, d_h^\circ]$ would maximize the Incumbent's support. By continuity it is thus immediate to conclude that incumbents with ideology close but lower than 1 will choose $d = \delta$, while incumbents with ideology close and higher than 1 will choose $d_h^\circ(\underline{d}_1)$. Hence, if $\delta \in (d_\ell^\circ(\underline{d}_1), d_h^\circ(\underline{d}_1)]$, we can define two cutoffs, $\underline{\theta} < \tilde{\theta}$, such that the equilibrium is characterized as follows:

- if $\theta_I \leq \underline{\theta}$, then the incumbent chooses $(0, 0)$;
- if $\theta_I \in (\underline{\theta}, 1]$, then the incumbent chooses $(1, \delta)$;
- if $\theta_I \in (1, \tilde{\theta}]$, then the incumbent chooses $(1, d_h^\circ(\underline{d}_1))$;
- if $\theta_I > \tilde{\theta}$, then the incumbent chooses $(1, 1)$.

Obviously, this equilibrium exists as long as the resulting expected escalation $\underline{d}_1$ is such that $\delta \in (d_\ell^\circ(\underline{d}_1), d_h^\circ(\underline{d}_1)]$. Also, $\underline{\theta}$ is defined as the ideology of the incumbents who are indifferent between (i) not challenging and (ii) challenging and then not doubling down. Similarly, $\tilde{\theta}$ is defined as the ideology of the incumbents who are indifferent between choosing a level of escalation equal to d_h° and one equal to 1. Putting together these observations, we can conclude that this equilibrium is characterized by the following set of equations:

$$\begin{aligned}
\underline{\theta} &= 1 - \frac{R}{2\delta} \\
\tilde{\theta} &= 1 + \frac{R}{1 - d_h^\circ(\underline{d}_1)} \left[\frac{1}{2} + \frac{1 + \eta(1 - \underline{d}_1)}{2 + \eta(1 - \underline{d}_1)} \right] \\
\underline{d}_1 &= 1 - \frac{(\tilde{\theta} - \underline{\theta}) + \delta(\underline{\theta} - 1) - d_h^\circ(\underline{d}_1)(\tilde{\theta} - 1)}{1/2 + (\tau - \underline{\theta})\phi} \phi
\end{aligned}$$

Observe that $\underline{\theta}$ lies below 1. Hence, in this case, reference dependence always turns some democrats into opportunistic authoritarians.

Finally, suppose that institutional checks and balances are sufficiently low to fall in Case 3. Then, the same reasoning described above implies that incumbents with ideology close but below (above) $\theta_I = 1$ will choose the lowest (highest) level of escalation that guarantees full support, $d_\ell^\circ(\underline{d}_1)$ ($d_h^\circ(\underline{d}_1)$). Differently from Case 2, however, incumbents with ideology lower than 1 who decides to challenge democratic norms can now choose two possible levels of d: δ or $d_\ell^\circ(\underline{d}_1) > \delta$. Given reference point $\underline{d}_1$, the utility that an incumbent with ideology θ_I gets

from playing $(0,0)$, $(1,\delta)$, and $(1,d_\ell^\circ(\underline{d}_1))$ are respectively equal to

$$\begin{aligned}
u_I(0,0;\theta_I) &= \theta_I + \frac{R}{2} \\
u_I(1,\delta;\theta_I) &= \theta_I + (\theta_I - 1)\delta + R\left[\frac{1}{2} + \psi\frac{\delta + \eta(\delta - \underline{d}_1)}{1 + \delta + \eta(\delta - \underline{d}_1)}\right] \\
u_I(1,d_\ell^\circ(\underline{d}_1);\theta_I) &= \theta_I + (\theta_I - 1)d_\ell^\circ(\underline{d}_1) + R
\end{aligned}$$

We can then have two possible equilibrium configurations. In the first one, there is no incumbent is choosing $d = \delta$ after $c = 1$. In this case, there are two cutoff types, $\underset{\sim}{\theta} < \tilde{\theta}$, and the incumbent behavior is as follows:

- if $\theta_I \leq \underset{\sim}{\theta}$, then the incumbent chooses $(0,0)$;
- if $\theta_I \in (\underset{\sim}{\theta}, 1]$, then the incumbent chooses $(1, d_\ell^\circ(\underline{d}_1))$;
- if $\theta_I \in (1, \tilde{\theta}]$, then the incumbent chooses $(1, d_h^\circ(\underline{d}_1))$;
- if $\theta_I > \tilde{\theta}$, then the incumbent chooses $(1,1)$.

The cutoff types as well as the reference points are defined by the following system of equations in three unknowns:

$$\begin{aligned}
\underset{\sim}{\theta} &= 1 - \frac{R}{2d_\ell^\circ(\underline{d}_1)} \\
\tilde{\theta} &= 1 + \frac{R}{1 - d_h^\circ(\underline{d}_1)}\left[\frac{1}{2} + \frac{1 + \eta(1 - \underline{d}_1)}{2 + \eta(1 - \underline{d}_1)}\right] \\
\underline{d}_1 &= 1 - \frac{(\tilde{\theta} - \underset{\sim}{\theta}) + d_\ell^\circ(\underline{d}_1)(\underset{\sim}{\theta} - 1) - d_h^\circ(\underline{d}_1)(\tilde{\theta} - 1)}{1/2 + (\tau - \underset{\sim}{\theta})\phi}\phi.
\end{aligned}$$

In the second equilibrium configuration, instead, a positive mass of incumbents chooses δ. In this case, there are three cutoffs, $\underline{\theta} < \underset{\sim}{\theta} < \tilde{\theta}$ and the incumbent behavior is the following:

- if $\theta_I \leq \underline{\theta}$, then the incumbent chooses $(0,0)$;
- if $\theta_I \in (\underline{\theta}, \underline{\theta}]$, then the incumbent chooses $(1,\delta)$;
- if $\theta_I \in (\underline{\theta}, 1]$, then the incumbent chooses $(1, d_\ell^\circ(\underline{d}_1))$;
- if $\theta_I \in (1, \tilde{\theta}]$, then the incumbent chooses $(1, d_h^\circ(\underline{d}_1))$;
- if $\theta_I > \tilde{\theta}$, then the incumbent chooses $(1,1)$.

In particular, the cutoff types and the reference point are defined by the following system equations in 4 unknowns:

$$\begin{aligned}
\underset{\sim}{\theta} &= 1 - \frac{R}{2\delta} \\
\underset{\sim}{\theta} &= 1 - \frac{R}{d_\ell^\circ(\underset{\sim}{d}_1) - \delta}\left[\frac{1}{2} - \psi\frac{\delta + \eta(\delta - \underset{\sim}{d}_1)}{1 + \delta + \eta(\delta - \underset{\sim}{d}_1)}\right] \\
\tilde{\theta} &= 1 + \frac{R}{1 - d_h^\circ(\underset{\sim}{d}_1)}\left[\frac{1}{2} + \psi\frac{1 + \eta(1 - \underset{\sim}{d}_1)}{2 + \eta(1 - \underset{\sim}{d}_1)}\right] \\
\underset{\sim}{d}_1 &= 1 - \frac{(\tilde{\theta} - \underset{\sim}{\theta}) - \delta(\underset{\sim}{\theta} - \underset{\sim}{\theta}) - d_\ell^\circ(\underset{\sim}{d}_1)(1 - \underset{\sim}{\theta}) - d_h^\circ(\underset{\sim}{d}_1)(\tilde{\theta} - 1)}{1/2 + (\tau - \underset{\sim}{\theta})\phi}\phi.
\end{aligned}$$

Independently of the actual equilibrium configuration, in Case 3, there is also a mass of incumbents who, despite being democrats, challenge democratic institutions because of popularity concerns. In other words, also in this case reference dependence turns some democrats into opportunistic authoritarians. Such mass is given by $(\underset{\sim}{\theta}, 1]$ in the first equilibrium and by $(\underline{\theta}, 1]$ in the second equilibrium.

B.2 Extension: Prospective voting

In the main text, we assumed that voters' decision to support is retrospective. As a result, their support for the incumbent is an increasing function of the utility experienced while the incumbent was in office. In this section, we show that, because of reference dependence, opportunistic authoritarians may arise also when voters are prospective, namely they condition their behavior on the expected utility that the incumbent could yield if reappointed. Since this assumption is especially meaningful in the context of voting, in this section we explicitly interpret the decision to support as a voting decision. We will show that, even under prospective voting, reference dependence can generate opportunistic authoritarians, and even some democratic types might behave in that way.

To model prospective voting, consider the following two-periods modification of our model. Suppose there is only one voter with ideology $\theta < 1$ and that the incumbent can be one of three possible types $\theta_I \in \{\theta, \theta_m, \theta_h\}$, with $\theta < \theta_m < \theta_h$ and $\theta_h \simeq \frac{1}{2\phi} + \tau$. The existence of a finite number of types simplifies the analysis, but it is without loss of generality. The incumbent's type is his private information and the voter holds a uniform prior, $\Pr(\theta_I) = 1/3$ for all θ_I. Further assume that the voter's utility of having the incumbent in office is equal to the opposite of the distance between her and his ideology, $u(\theta_I) = -(\theta_I - \theta)$. Hence, ideological differences are associated with differences in utility.

The timing of the game is as follows. In period 1, the incumbent first chooses whether to challenge norms or not, $c \in \{0,1\}$ and then picks a level of escalation $d \in \{0,\delta,1\}$ or equivalently a policy $y \in \mathcal{Y}(c)$ with the following constraint: $\mathcal{Y}(0) = 1$ and $\mathcal{Y}(0) = \{1+\delta, 2\}$.[21] As in the main text, we assume that the reference point is determined upon observing the decision on whether to challenge democratic norms or not. At the end of the first period elections are held and the incumbent is reappointed if the expected total utility of the voter conditional on the information available at the end of period 1 is greater than a random variable ξ uniformly distributed in the interval $\left[-\frac{1}{2\chi}, \frac{1}{2\chi}\right]$. ξ is realized after the incumbent chooses the vector (c,d) and its realization is independent of it. In particular, we can interpret ξ as the a proxy of the ideological closeness of the challenger with the voter and assume that if the voter appoints a challenger, his reference point fully adjusts to the type of the challenger.[22] In period 2, payoffs are experienced.

Because the game has only two periods, we assume that neither the voter, nor the incumbent discount the future and we avoid discussing the dynamic updating of reference points. The expected utility of the voter when she has to decide who to vote for is thus given by:

$$E[u(c,d)] = -(E[\theta_I \mid c,d] - \theta) + \eta\left\{-(E[\theta_I \mid c,d] - \theta) + (E[\theta_I \mid c] - \theta)\right\}.$$

The incumbent utility is equal to

$$E[u_I(c,d)] = \theta_I + (\theta_I - 1)cd + R\hat{\pi}(c,d),$$

where $\hat{\pi}(c,d)$ is the probability with which the incumbent gets reappointed.[23]

In this setting, we can prove the following result:

Proposition B.2. *If the voter votes prospectively and exhibits reference dependence, there exists equilibria in which a democrat incumbent with ideology $\theta_m < 1$ behaves as an opportunistic authoritarian and chooses $(c,d) = (1,\delta)$.*

[21] Contrary to the baseline model, we assume that after a challenge the incumbent can only choose between $d = \delta$ or $d = 1$. This enables us to avoid discussing out-of-equilibrium beliefs, yet none of our results hinges on this restriction.

[22] Our results would hold also if the reference point used to evaluate the challenger was a weighted average of the type of the challenger and the type of the incumbent as perceived after the choice of c, provided that the weight on the latter is not too high.

[23] Given that there is only one voter, the incumbent's support in the main text is replaced by the probability of reappointment.

Proof. Suppose the incumbent with ideology θ does not challenge, while incumbents with ideology θ_m and θ_h challenge, but they differ in the choice of d, with the former choosing δ and the latter choosing 1. Because this strategy is separating, the voter can infer the type of the incumbent from his behavior. Hence,

$$E[\theta_I \mid c] = \begin{cases} \theta & \text{if } c = 0 \\ \frac{\theta_m + \theta_h}{2} & \text{if } c = 1 \end{cases} \qquad E[\theta_I \mid c, d] = \begin{cases} \theta & \text{if } (c,d) = (0,0) \\ \theta_m & \text{if } (c,d) = (1,\delta) \\ \theta_h & \text{if } (c,d) = (1,1) \end{cases}$$

Exploiting the distributional assumption of ξ, we can thus conclude that the probability with which the incumbent is reappointed can be written as follows

$$\hat{\pi}(c,d) = \begin{cases} \frac{1}{2} & \text{if } (c,d) = (0,0) \\ \frac{1}{2} + \chi\left[-(\theta_m - \theta) + \eta\left(\frac{\theta_h - \theta_m}{2}\right)\right] & \text{if } (c,d) = (\delta,1) \\ \frac{1}{2} + \chi\left[-(\theta_h - \theta) - \eta\left(\frac{\theta_h - \theta_m}{2}\right)\right] & \text{if } (c,d) = (1,1) \end{cases} \tag{B-5}$$

In order for the postulated incumbent's behavior to be incentive compatible, we need to satisfy the following three conditions:

$$\begin{aligned} 1 - \theta &\geq \frac{R}{\delta}\chi\left[\eta\left(\frac{\theta_h - \theta_m}{2}\right) - (\theta_m - \theta)\right], \\ (1 - \theta_m) &\in \left[\frac{-R}{1-\delta}\chi(\theta_h - \theta_m)(1+\eta), \frac{R}{\delta}\chi\left(\eta\left(\frac{\theta_h - \theta_m}{2}\right) - (\theta_m - \theta)\right)\right] \\ (\theta_h - 1) &\geq \frac{R}{1-\delta}\chi(\theta_h - \theta_m)(1+\eta) \end{aligned}$$

Because $\theta_m < 1$, it is immediate to verify that if $\eta = 0$, the second condition cannot be satisfied. Hence, there is no fully separating equilibrium in which a democrat challenges democratic norms and then choose $d = \delta$. In other words, in the absence of reference dependence, democrats cannot turn into opportunistic authoritarians.

Suppose instead that $\eta > 0$. In this case, the first and second conditions are satisfied if

$$1 - \theta_m < \frac{R}{\delta}\chi\left[\eta\left(\frac{\theta_h - \theta_m}{2}\right) - (\theta_m - \theta)\right] < 1 - \theta.$$

Because $\theta < \theta_m$, this condition is satisfied for a non-empty range of ηs. Furthermore, the third condition can be satisfied by taking (for instance) χ to be sufficiently low. □

B.3 Extension: Rational Inattention

Consider the following simplified extension of our baseline model (as in the previous extension, we interpret the decision to support as a voting decision):

1. There is a single citizen (voter), with ideology $\theta_v = 0$;
2. the choice of d is binary: $d \in \{\delta, 1\}$;
3. the voter v re-elects the incumbent if and only if his payoff exceeds the realization of a zero-mean uniform popularity shock $\xi \in \left[-\frac{1}{2\chi}\frac{1}{2\chi}\right]$—higher realizations implying a more charismatic/popular opponent and higher values of χ a less volatile electoral environment;
4. the probability that the voter observes I's choices depends on her level of attention, which equals $a \in [0, 1]$–chosen before the incumbent makes his choices;
5. a is associated with a cognitive cost $\frac{a^2}{\alpha 2}$, reflecting the voter's opportunity cost of acquiring and processing political information

Recall that $\mathbf{q} = \{c, d\}$ identifies a sequence of choices by the incumbent (the policy $y(c, d)$ is uniquely determined by c and d). Under the assumptions, the voter's material payoff simplifies to

$$u(\mathbf{q}) = -cd$$

Specifically, the voter observes two reports: $r_1 \in \{\emptyset, c\}$ (realized after c is chosen) and $r_2 \in \{\emptyset, cd\}$ (realized after d is chosen) and attention effort increases the probability of observing an informative report.

In this extension, the voter's interim payoff from supporting the incumbent is a function of the report observed by the voter and the incumbent's actions. We write it as

$$\hat{v}(r_1, r_2; \mathbf{q}) = E\left[u(\mathbf{q}) \mid r_1, r_2\right] + \eta\Big\{E\left[u(\mathbf{q}) \mid r_1, r_2\right] - E\left[u(\mathbf{q}) \mid r_1\right]\Big\} \tag{B-6}$$

In particular,

$$\begin{aligned}
\hat{v}(c, cd; \mathbf{q}) &= -cd + \eta\big(-cd + E[cd \mid c]\big),\\
\hat{v}(c, \emptyset; \mathbf{q}) &= -E[cd \mid c],\\
\hat{v}(\emptyset, cd; \mathbf{q}) &= -cd + \eta\big(-cd + E[cd \mid \emptyset]\big),\\
\hat{v}(\emptyset, \emptyset; \mathbf{q}) &= -E[cd \mid \emptyset].
\end{aligned}$$

To conserve space, let

$$\begin{aligned}\rho_{12} &\equiv \Pr(r_1 = c, r_2 = cd)\\ \rho_2 &\equiv \Pr(r_1 = \emptyset, r_2 = cd)\\ \rho_1 &\equiv \Pr(r_1 = c, r_2 = \emptyset)\\ \rho_\emptyset &\equiv \Pr(r_1 = \emptyset, r_2 = \emptyset)\end{aligned}$$

and recall that all this quantities are function of the voter's attention level.

Because ξ is independent of the voter's information and its density is linear, the incumbent's reelection probability is given by

$$\pi(\mathbf{q}; a) = \frac{1}{2} + \chi \hat{V}(\mathbf{q}, a) \tag{B-7}$$

where

$$\hat{V}(\mathbf{q}, a) = \left\{ \begin{array}{l} \rho_{12}\big(- cd(1+\eta) + \eta E[cd \mid c]\big) + \rho_1 E[-cd \mid c] + \\ +\rho_2\big(- cd(1+\eta) + \eta E[cd \mid \emptyset]\big) + \rho_\emptyset E[-cd] \end{array} \right\} - \frac{a^2}{2\alpha}$$

Notice that in any equilibrium $\hat{V}(\mathbf{q}, a) \geq \hat{V}(\mathbf{q}, 0) \geq -1$ and $\hat{V}(\mathbf{q}, a) \leq 1+\eta$. Hence, imposing $\frac{1}{2\chi} \geq 1 + \eta$ ensures that π is interior. To ensure a positive measure types choosing $(0, 0)$, we impose

$$\tau - \frac{1}{2\phi} \geq \tau - \frac{1}{2\phi} - \delta\left(1 - \tau + \frac{1}{2\phi}\right) + R,$$

that is $\frac{1}{2\phi} \geq \frac{R}{\delta} + 1 - \tau$. To ensure a positive measure types choosing $(1, 1)$, we impose

$$\tau + \frac{1}{2\phi} - \left(1 - \tau - \frac{1}{2\phi}\right) \geq \tau + \frac{1}{2\phi} - \delta\left(1 - \tau - \frac{1}{2\phi}\right) + R,$$

that is $\frac{1}{2\phi} \geq \frac{R}{1-\delta} - 1 + \tau$.

The incumbent now faces two dimensions of uncertainty when it comes to the voter's behavior: the realization of the shock ξ and the realization of the voter's information—i.e., whether or not she observed her choices *at the time in which they were chosen*.

Characterization of π. We begin with some notation: holding the strategy of the incumbent fixed, let $\underline{d}_1 = E[d \mid c = 1]$ and $p_0 = \Pr(c = 0)$. Since $E[-cd] = -(1 - p_0)\underline{d}_1$, (B-8) π

can be rewritten as

$$\pi(\mathbf{q};a) = \frac{1}{2} - \chi\frac{a^2}{2\alpha} + \chi \left[\begin{array}{l} \rho_{12}\eta c\underline{d}_1 + (\rho_2\eta - \rho_\emptyset)(1-p_0)\underline{d}_1 \\ -(\rho_{12}+\rho_2)cd(1+\eta) - \rho_1 c\underline{d}_1 \end{array} \right] \tag{B-8}$$

Notice that, conditional on choosing $c = 1$, higher values of d lead to a lower vote share. This effect operates through two channels: (i) increased disappointment when voters learn both c and cd, and (ii) reduced material payoff whenever voters learn their material payoff (i.e., when they learn the value of cd). The expression also reveals that initial pessimism about the incumbent's actions (i.e., higher $\underline{d}_1$) has an ambiguous effect on her vote share: when the voter observes *both* incumbent's actions or just her material payoff, higher $\underline{d}_1$ decreases the standard to which the incumbent is held—and thus improves his standing. Conversely, when the voter does not observe anything or when she only observes the incumbent's initial action c but not her choice of doubling down, higher $\underline{d}_1$ decreases the voter's payoff from I and he probability of supporting him (standard retrospective channel).

This suggests that when incumbents expect voters not to observe r_2, decreasing their reference point is less electorally profitable.

We now compute the expected payoff associated with each of the three possible actions available to the incumbent.

$$\begin{aligned}
u_I(0,0;\theta_I) &= \theta_I + \frac{R}{2} - R\chi\frac{a^2}{2\alpha} - R\chi(1-p_0)\underline{d}_1(\rho_\emptyset - \eta\rho_2) \\
u_I(1,\delta;\theta_I) &= u_I(0,0;\theta_I) + \delta(\theta_I - 1) + R\chi[(\rho_{12}\eta - \rho_1)\underline{d}_1 - (\rho_{12}+\rho_2)\delta(1+\eta)] \\
u_I(1,1;\theta_I) &= u_I(0,0;\theta_I) + (\theta_I - 1) + R\chi[(\rho_{12}\eta - \rho_1)\underline{d}_1 - (\rho_{12}+\rho_2)(1+\eta)]
\end{aligned}$$

From this, it is immediate to see that there are two thresholds

$$\begin{aligned}
\underline{\theta} &\equiv 1 + R\chi\left[(\rho_{12}+\rho_2)(1+\eta) - (\rho_{12}\eta - \rho_1)\frac{\underline{d}_1}{\delta}\right] \\
\overline{\theta} &\equiv 1 + R\chi\left[(\rho_{12}+\rho_2)(1+\eta)\right],
\end{aligned}$$

such that an incumbent's individually rational strategy must satisfy

$$c^*(\theta), d^*(\theta) = \begin{cases} 0,0 & \theta \leq \underline{\theta} \\ 1,\delta & \theta \in (\underline{\theta}, \overline{\theta}] \\ 1,1 & \theta > \overline{\theta} \end{cases}$$

Compared to the baseline model, one can see that the two thresholds converge to each other as voter attention approaches zero (i.e., as $\rho_\emptyset$ approaches one): without voter attention *both* disciplined authoritarians and opportunistic authoritarians disappear. The reason is that voter attention governs the size of the electoral response to an incumbent's actions. Moreover:

- η, R, χ all increase $\overline{\theta}$, thereby strengthening the disciplining effect. $\overline{\theta}$ does not depend on δ: checks and balances decrease the policy gain and increase the electoral cost of full escalation in the same way, and thus do not affect the comparison between the two;
- the effect of the parameters (η, R, χ, δ) on $\underline{\theta}$ depends on the endogenous quantity $\underline{d}_1$;
- The effect of attention depends on what type of learning it favors: when $\underline{\theta} < 1$, it decreases in R, χ. The effect of η depends on the sign of $\rho_2 - \rho_{12}\frac{\underline{d}_1 - \delta}{\delta}$. When ρ_{12} is large enough relative to ρ_2 (i.e., voter attention is high enough) it decreases $\underline{\theta}$.
- ρ_2 and ρ_1, the probabilities of partial learning, increase $\underline{\theta}$: when the voter only observes the incumbent's first or second choices, challenging democratic institutions can only lower her expected payoff—but there is no gap between reference point and final payoff. It is only when the voter learns both c and cd that the incumbent can obtain an electoral benefit by lowering her reference point and then generating a positive surprise with his choice of not doubling down ($d = \delta$).

Proposition B.3 (No information avoidance)**.** *Suppose that voter attention increases ρ_{12} and that there exists $\underline{\rho}' < 0$ such that $\min\left\{\frac{\partial \rho_1}{\partial a}, \frac{\partial \rho_2}{\partial a}\right\} \geq \underline{\rho}'$. Then the marginal value of attention at $a = 0$ is strictly positive.*

Proof. Let p_0 and p_1 be the probabilities with which the Incumbent chooses $(c, d) = (0, 0)$ and $(c, d) = (1, 1)$, respectively. Then, the voter's expected payoff as a function of her attention is

$$W(a) = \Bigg\{ p_0 E_{r_1, r_2, \xi}\Big[\max\{\hat{v}(r_1, r_2; 0, 0), \xi\}\Big] + p_1 E_{r_1, r_2, \xi}\Big[\max\{\hat{v}(r_1, r_2; 1, 1), \xi\}\Big] + \\ (1 - p_0 - p_1) E_{r_1, r_2, \xi}\Big[\max\{\hat{v}(r_1, r_2; 1, \delta), \xi\}\Big] \Bigg\} =$$

$$
\begin{aligned}
= \Bigg\{ & \rho_{12}\Big[p_0 E_\xi\Big[\max\{\hat{v}(0,0;0,0),\xi\}\Big] + p_1 E_\xi\Big[\max\{\hat{v}(1,1;1,1),\xi\}\Big] + \\
& \quad (1-p_0-p_1)E_\xi\Big[\max\{\hat{v}(1,\delta;1,\delta),\xi\}\Big]\Big] + \\
& \rho_1\Big[p_0 E_\xi\Big[\max\{\hat{v}(0,\emptyset;0,0),\xi\}\Big] + p_1 E_\xi\Big[\max\{\hat{v}(1,\emptyset;1,1),\xi\}\Big] + \\
& \quad (1-p_0-p_1)E_\xi\Big[\max\{\hat{v}(1,\emptyset;1,\delta),\xi\}\Big]\Big] + \\
& \rho_2\Big[p_0 E_\xi\Big[\max\{\hat{v}(\emptyset,0;0,0),\xi\}\Big] + p_1 E_\xi\Big[\max\{\hat{v}(\emptyset,1;1,1),\xi\}\Big] + \\
& \quad (1-p_0-p_1)E_\xi\Big[\max\{\hat{v}(\emptyset,\delta;1,\delta),\xi\}\Big]\Big] + \\
& (1-\rho_{12}-\rho_1-\rho_2)E_\xi\Big[\max\{\hat{v}(\emptyset,\emptyset;\mathbf{q}),\xi\}\Big]\Bigg\}
\end{aligned}
$$

where we used the fact that $\hat{v}(\emptyset,\emptyset;\mathbf{q}) = -(1-p_0)\underline{d}_1$ for all $\mathbf{q}$. Now, let $w(r_1,r_2;\mathbf{q}) = E_\xi\Big[\max\{\hat{v}(r_1,r_2;\mathbf{q}),\xi\}\Big]$. We can thus rewrite the previous expression as

$$
\begin{aligned}
W(a) = \Bigg\{ & \rho_{12}\Big[p_0 w(0,0;0,0) + p_1 w(1,1;1,1) + (1-p_0-p_1)w(1,\delta;1,\delta)\Big] + \\
& \rho_1\Big[p_0 w(0,\emptyset;0,0) + p_1 w(1,\emptyset;1,1) + (1-p_0-p_1)w(1,\emptyset;1,\delta)\Big] + \\
& \rho_2\Big[p_0 w(\emptyset,0;0,0) + p_1 w(\emptyset,1;1,1) + (1-p_0-p_1)w(\emptyset,\delta;1,\delta)\Big] + \\
& (1-\rho_{12}-\rho_1-\rho_2)w(\emptyset,\emptyset;\mathbf{q})\Bigg\}
\end{aligned}
$$

Observe that $w(\emptyset,\emptyset;\mathbf{q})$ does not depend on $\mathbf{q}$. Consider the following three expressions.

$$
p_0 w(0,0;0,0) + (1-p_0-p_1)w(1,\delta;1,\delta) + p_1 w(1,1;1,1) - w(\emptyset,\emptyset;\mathbf{q}) \tag{B-9}
$$

$$
p_0 w(0,\emptyset;0,0) + (1-p_0-p_1)w(1,\emptyset;1,\delta) + p_1 w(1,\emptyset;1,1) - w(\emptyset,\emptyset;\mathbf{q}) \tag{B-10}
$$

$$
p_0 w(\emptyset,0;0,0) + (1-p_0-p_1)w(\emptyset,\delta;1,\delta) + p_1 w(\emptyset,1;1,1) - w(\emptyset,\emptyset;\mathbf{q}). \tag{B-11}
$$

We will show that they are all positive. To this goal, we first show that for no matter what the voter learns, her payoff is a random function of $\mathbf{q}$ with mean $\hat{v}(\emptyset,\emptyset;\mathbf{q}) = -(1-p_0)\underline{d}_1 = -(1-p_0-p_1)\delta - p_1$.

Consider first information set $(\emptyset, \emptyset)$. Then:

$$\begin{aligned} E_{c,d}[\hat{v}(\emptyset,\emptyset;c,d)] =& p_0\hat{v}(\emptyset,\emptyset;0,0) + p_1\hat{v}(\emptyset,\emptyset;1,1) + (1-p_0-p_1)\hat{v}(\emptyset,\emptyset;1,\delta) \\ =& \hat{v}(\emptyset,\emptyset;0,0) = -(1-p_0)\underline{d}_1 = -(1-p_0-p_1)\delta - p_1 \end{aligned}$$

Now, consider information set $(\emptyset, cd)$. Then:

$$\begin{aligned} E_{c,d}[\hat{v}(\emptyset,cd;c,d)] =& p_0\hat{v}(\emptyset,0;0,0) + p_1\hat{v}(\emptyset,1;1,1) + (1-p_0-p_1)\hat{v}(\emptyset,\delta;1,\delta) \\ =& p_0\eta(1-p_0)\underline{d}_1 + p_1(-(1+\eta) + \eta(1-p_0)\underline{d}_1) + \\ & (1-p_0-p_1)(-\delta(1+\eta) + \eta(1-p_0)\underline{d}_1) \\ =& -(1+\eta)(p_1 + (1-p_0-p_1)\delta) + \eta(1-p_0)\underline{d}_1 \\ =& -(1+\eta)[p_1 + (1-p_0-p_1)\delta] + \eta(1-p_0)\frac{(p_1 + (1-p_0-p_1))\delta}{1-p_0} \\ =& -(1-p_0-p_1)\delta - p_1 \end{aligned}$$

Now consider information set $(c, \emptyset)$. Then

$$\begin{aligned} E_{c,d}[\hat{v}(c,\emptyset;c,d)] =& p_0\hat{v}(0,\emptyset;0,0) + p_1\hat{v}(1,\emptyset;1,1) + (1-p_0-p_1)\hat{v}(1,\emptyset;1,\delta) \\ =& (1-p_0)(-\underline{d}_1(1+\eta) + \eta\underline{d}_1) = -(1-p_0-p_1)\delta - p_1 \end{aligned}$$

Finally, consider information set (c, cd). Then

$$\begin{aligned} E_{c,d}[\hat{v}(c,cd;c,d)] =& p_0\hat{v}(0,0;0,0) + p_1\hat{v}(1,1;1,1) + (1-p_0-p_1)\hat{v}(1,\delta;1,\delta) \\ =& (1-p_0-p_1)(-\delta(1+\eta) + \eta\underline{d}_1) + p_1(-(1+\eta) + \eta\underline{d}_1) \\ =& -(1+\eta)(p_1 + (1-p_0-p_1)\delta) + \eta(1-p_0)\underline{d}_1 = -(1-p_0-p_1)\delta - p_1 \end{aligned}$$

Hence, we have shown that for each information realization $(r_1, r_2) \in \{c, \emptyset\} \times \{cd, \emptyset\}$

$$E_{c,d}[\hat{v}(r_1, r_2; c, d)] = \hat{v}(\emptyset, \emptyset; 0, 0)$$

Now, notice that for every $k \in supp(\xi)$

$$g(k) = E_\xi\Big[\max\{k, \xi\}\Big] = \frac{1}{8\chi} + \frac{k + \chi k^2}{2}$$

Given the convexity of $g(k)$ and the definition of $w(c, d; r_1, r_2)$, Jensen's inequality implies

$$E_{c,d}[w(r_1, r_2; c, d)] \geq E_\xi \left[\max\{E_{c,d}\left[\hat{v}(r_1, r_2; c, d)\right], \xi\}\right] =$$
$$= E_\xi \left[\max\{-(1 - p_0)\underline{d}_1, \xi\}\right] = w(0, 0, \emptyset, \emptyset).$$

Hence, equations (B-9)-(B-11) are all positive. Given the cognitive cost of information acquisition, $a^2/(2\alpha)$, we conclude that the voter will not fully avoid information. □

B.4 Extension: Challenges of Varying Severity

In the baseline model, the incumbent's initial choice is binary: she can only choose between challenging norms of democracy or not. In many situations, incumbents have a wider set of options available. In particular, they can choose between challenges of varying severity: restricting immigrants' legal right to apply for asylum is arguably less severe than ordering immigration enforcement agents to shoot at people crossing the border.

In our baseline model, the benefit associated with a challenge to democratic norms is proportional to its associated reduction in citizens' expected utility. Hence, one might conjecture that opportunistic authoritarians should prefer more severe challenges if they have the same ability to back down. In this section we show that, when the cost of backsliding is convex in the severity of the challenge (an assumption consistent with the baseline model) this conjecture does not hold.

The reason is that citizens' reference point must be consistent with equilibrium behavior. Indeed, when a challenge is extremely severe, citizens will anticipate that the incumbent will almost certainly not escalate and their reference point will thus put low weight on this event. As a result, the electoral reward (again, in this extension we explicitly interpret the popularity concern as an electoral concern) associated with citizens' relief will be limited. More generally, the fact that the reference point is endogenous to equilibrium behavior implies that one cannot stretch the logic of the model to produce implausible outcomes.

To formalize this intuition, consider the following simplified extension of our baseline model:

1. There is a single citizen, with ideology $\theta_v = 0$;
2. c can take one of three possible values: $c \in \{0, 1, \kappa\}$, with $\kappa > 1$;
3. the choice of d is binary, but the outcomes under no escalation are independent of the severity of the challenge: $d \in \{\delta_c, 1\}$, with $\delta_1 \leq \delta_\kappa$;

4. the citizen v supports the incumbent if and only if his payoff exceeds the realization of a zero-mean uniform popularity shock $\xi \in \left[-\frac{1}{2\chi}\frac{1}{2\chi}\right]$—higher realizations implying a more charismatic/popular opponent and higher values of χ a less volatile electoral environment

5. the cost of backsliding is convex: $c^2 d$ (notice that this does not affect the baseline model)

Proposition B.4. *There exists κ^* such that for all $\kappa > \kappa^*$, there is no equilibrium in which a democratic type chooses $c = \kappa$.*

Proof. Fix the support of incumbent types to an arbitrarily large interval $\left[\tau - \frac{1}{2\phi}, \tau + \frac{1}{2\phi}\right]$ and suppose that in equilibrium a liberal type chooses $(c, d) = (\kappa, \delta_\kappa)$.

Notice that we have $u(\mathbf{q}; \theta) = \theta + \theta cd - c^2 d$ and thus

$$\pi(\mathbf{q}) = \frac{1}{2} - \chi c^2 d - \chi\eta c^2 \{d - E[d \mid c]\}$$

If $\theta_I < 1$ chooses $(c, d) = (\kappa, \delta_\kappa)$, it must be that he prefers it to $(c, d) = (0, 0)$ that is

$$u_I(0, 0; \theta_I) + \frac{R}{2} < u_I(\kappa, \delta_\kappa; \theta_I) + \frac{R}{2} - \chi(1+\eta)\kappa^2 \left\{\delta_\kappa - \frac{\eta}{1+\eta}\underline{d}_\kappa\right\}.$$

with $\underline{d}_\kappa = E[d \mid c = \kappa]$. Rearranging, the expression becomes

$$\begin{aligned} 0 &< \delta_\kappa \kappa(\theta_I - \kappa) - \chi(1+\eta)\kappa^2 \left\{\delta_\kappa - \frac{\eta}{1+\eta}\underline{d}_\kappa\right\} \\ \Leftrightarrow \frac{\kappa - \theta_I}{\chi(1+\eta)} &< \kappa \left\{\frac{\eta}{1+\eta}\frac{\underline{d}_\kappa}{\delta_\kappa} - 1\right\} \end{aligned} \tag{B-12}$$

Notice that expression (B-12) requires $\frac{\eta}{1+\eta}\frac{\underline{d}_\kappa}{\delta_\kappa} > 1$.

By single crossing, all types above θ_I must choose either (κ, δ_κ), $(1, 1)$, or $(\kappa, 1)$. The incumbent types choosing $(\kappa, 1)$ must be above $\overline{\theta}_\kappa$, which solves

$$u_I(\kappa, 1; x) + \frac{R}{2} - \chi(1+\eta)\kappa^2 \left\{1 - \frac{\eta}{1+\eta}\underline{d}_\kappa\right\} = u_I(\kappa, \delta_\kappa; x) + \frac{R}{2} - \chi(1+\eta)\kappa^2 \left\{\delta_\kappa - \frac{\eta}{1+\eta}\underline{d}_\kappa\right\}.$$

We obtain

$$\overline{\theta}_\kappa = \kappa[1 + \chi(1+\eta)]$$

As κ increases $\overline{\theta}_\kappa$ approaches $\tau + \frac{1}{2\phi}$, which implies that $\underline{d}_\kappa$ gets arbitrarily close to δ_κ. Hence, there exists κ' such that

$$\frac{\eta}{1+\eta}\frac{\underline{d}_\kappa}{\delta_\kappa} = 1$$

Since this leads to a certain failure of expression (B-12), we must have that $\kappa^* \in (1, \kappa')$. $\square$