Opportunistic Authoritarians, Reference-Dependent Preferences and Democratic Backsliding

Edoardo Grillo* Carlo Prato†

October 28, 2019

Abstract

Recent attempts to weaken constraints of accountability on elected leaders are causing growing concerns about the state of liberal democracy. Yet, the evidence shows that electorates remain largely committed to democratic norms. This paper shows that democratic backsliding can occur even when most voters *and most incumbents* intrinsically value democratic institutions. Due to voters' reference-dependent preferences, *opportunistic authoritarians* emerge: against their own liberal tendencies, these leaders choose to challenge norms of democracy (and then partially back down) in order to lower the standards to which voters will hold them. In equilibrium, voters do sanction serious, sudden blows to democratic norms, but electoral incentives do not inoculate against gradual backsliding—in fact, under certain conditions they encourage it. We show that polarization, voter's access to information, and institutional checks and balances all have an ambiguous effect on the occurrence of democratic backsliding.

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

†Columbia University. Email: cp2928@columbia.edu

1. Introdution

In the summer of 2019, after withdrawing his party's support to the cabinet in which he was serving as deputy Prime Minister, Matteo Salvini announced his demands to Italian voters: "full powers, to carry out what we promised in full, without holdups or stumbling blocks." Throughout his tenure in the executive, Salvini accused the judges who ruled against his initiatives of left-wing and pro-migrant bias. From his Facebook page, he also threatened to remove police protection from a journalist who criticized him. While he eventually chose to respect the verdicts and re-authorize his critic's security detail, his very public attempts to weaken democratic checks and balances translated in substantial gains in the polls.

Salvini's actions are hardly exceptional. From the Prime Minister's suspension of Parliament in the U.K. to the attempts to stonewall congressional oversight by the U.S. President, from the forced retirement of constitutional judges in Poland to the purges of public employees by emergency decree in 2016-2018 Turkey, scholars and observers are becoming increasingly concerned about democratic backsliding (Waldner and Lust, 2018; Levitsky and Ziblatt, 2018; Przeworski, 2019): democratically elected leaders are trying to weaken institutional constraints of accountability that were generally respected by their predecessors. Crucially, these attempts are (i) very public, (ii) do not appear to hurt the popularity of these leaders and (iii) are frequently followed by a partial backing down.

These features are at odds with the available observational and experimental evidence that most voters all else equal dislike challenges to democratic norms (Voeten, 2016; Graham and Svolik, 2019; Carey et al., 2019). Even if electoral institutions had lost part of their sanctioning power over politicians' autocratic ambitions—for instance, due to polarization (Svolik, 2019; Nalepa, Vanberg and Chiopris, 2018)—we should still expect these attempts to take a less overt form.

In this paper, we show that challenges to democracy can arise even when (i) most incumbents do not have autocratic ambitions or intrinsic preferences for breaking these norms, and (ii) most voters would prefer incumbents not to engage in these actions. The key to our proposed mechanism is that the way in which voters form retrospective judgments of incumbents can

create an electoral reward for *incremental* forms of democratic backsliding. We uncover the possibility of *opportunistic authoritarians*, that is, incumbents who intrinsically value democratic norms and yet choose to challenge them because doing so enhances their electoral success.

In a departure from the standard rational-choice theoretic paradigm, we assume voters exhibit *reference-dependence*: they evaluate outcomes and politicians not solely based on external standards (e.g., the viability of challengers), but also based on context-dependent factors that can be manipulated by incumbents. We capture this idea by assuming that voters' expectations about an incumbent's behavior shape their restrospective evaluations of the incumbent's performance: the more pessimistic a voter becomes after observing the incumbent's early actions, the lower the standard she will employ in her evaluation of the incumbent's final performance. As in Lindstädt and Staton (2012), incumbents have an incentive to lower voters' expectations, and challenging democratic norm is particularly effective *precisely because* most voters oppose these actions. After initially challenging (and damaging) democratic institutions, opportunistic authoritarians partially back down (or refrain from fully escalating), which leads to beat voters' expectations.

Crucially, we show that opportunistic authoritarians can only arise when voters' are sufficiently uncertain about politicians' ideology. We relate this result to the documented disintermediation of political representation and communication: in our model, challenging democracy is a more viable strategy when voters' 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 formalizes and underscores the importance of intermediation by parties and media—and how their weakening in recent years is behind the rise of populist authoritarianism (Mair, 2002; Rosenblum, 2010).

Our theory does not suggest that electoral always promote democratic backsliding. In fact, elections do rein in the authoritarian impulses of truly autocratic incumbents: as in Svolik (2019) and Nalepa, Vanberg and Chiopris (2018), voters do sanction incumbent who challenge democratic norms and, as a result, produce "closet autocrats." However, the contemporaneous presence of opportunistic authoritarians and closet autocrats complicates the

relationship between the occurrence of democratic backsliding and several key factors identified by previous scholarship. In particular, electoral responsiveness (i.e., lower polarization), voter information, and the strength of checks and balances all reduce the disciplining effect of elections vis a' vis closet autocrats but also encourage opportunistic authoritarianism. As a result, stronger checks and balances and a more attentive electorate can actually increase the likelihood of democratic backsliding.

In addition to these novel empirical implications, our theory provides a mechanism that simultanously account for voters' instrinsic commitment to democracy described in Voeten (2016) and for the public, gradual and often uncompleted challenges to democratic norms observed in Turkey, Poland, Hungary, and—on a smaller scale—in United States and other west European countries: 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 fully disappointing them.

2. Related Literature

Our paper contributes to two main strands of literature: first a literature on the causes of democratic backsliding and, more generally, of expansion of executive authority. Second, a formal-theoretical literature on context-dependent preferences to political science.

Existing formal theoretical accounts of democratic backsliding emphasize polarization as a cause of disciplining role of the electoral mechanism (Svolik, 2019; Nalepa, Vanberg and Chiopris, 2018). These explanation necessarily rest on the premise that a substantial share of democratically elected incumbents have "authoritarian ambitions," and that backsliding is caused by the diminished willingness and ability of voters to monitor politicians. These explanation are at odds with the fact that in recent years (i) voter engagement and political participation grew relative to the previous decade and (ii) political polarization did not uniformly grow and (iii) political parties have selected leaders whose programmatic commitment are often largely dissimilar from their parties' traditional positions. Our theory accounts for all three elements.

In this paper, democratic backsliding refers to the violation (full or partial) of the constraints that in a liberal democracy limit the executive's ability to influence government action. The notion encompasses actions that clearly infringe the law, merely test its boundaries or try to circumvent its goals (e.g., term limits evasion—see Versteeg et al., 2019), or simply impact norms that have been traditionally respected. In this sense, it is close to what other authors refer to as "executive absolutism" (Howell, Shepsle and Wolton, 2019), "constitutional hardball" (Helmke, Kroeger and Paine, 2019), and brinkmanship (Schwarz and Sonin, 2007). What distinguishes our paper from previous explanations is the focus on how these actions affect voters' evaluations of incumbents *and* the standards to which they are held.

The non-formal literature on democratic backsliding, on the other hand, is heavily focused on the question "why now." While alleged culprits abound (e.g., the financial crisis, the demise of Communism), a key narrative focuses on the the rise of social media as the expense of traditional media in shaping political discourse: while citizens have a unprecedented access to politicians' daily activities, their ability to interpret and evaluate their actions seem to be diminished. Our approach resolves this tension using the idea that context matters for these evaluations. in that sense, we apply ideas articulated by Lindstädt and Staton (2012) to the context of democratic backsliding and provide a rigorously microfounded way to model context-dependent preferences: Lindstädt and Staton's reduced-form approach interprets expectation differentials as a source of excess return for an access-oriented donor, not as a psychological cost.

Our model also contributes to the formal literature on context dependent preferences in political science—of which reference dependence is a special case. At its core, the literature on reference dependence assumes that the utility of individuals depend not only on the outcome experience, but also by how this outcome compares to some reference point. This idea has a long history in the social and behavioral sciences (Kahneman and Tversky, 1979; Bell, 1985; Tversky and Kahneman, 1991) that includes axiomatic foundations (Gul et al., 1991; Sugden, 2003; Ok, Ortoleva and Riella, 2015) and a substantial body of evidence in favor of it.[1]

[1]See, for instance, Farber (2008) on labor markets, Pope and Schweitzer (2011) on sports,Lien and Zheng (2015) on gambling behavior and Fehr, Hart and Zehnder (2011) on contractual settings.

Our theoretical formulation builds on these ideas. Specifically, we follow the theoretical work of Kőszegi and Rabin (2006, 2007, 2009) and the experimental evidence of Abeler et al. (2011) and assume that 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 these ideas to politics. The empirical and experimental evidence in Kimball and Patterson (1997); Waterman, Jenkins-Smith and Silva (1999); Corazzini et al. (2014) highlights how expectations shape political preferences and has been rationalized through reference dependence by recent theoretical work (Lockwood and Rockey, 2015; Grillo, 2016; Acharya and Grillo, 2019; Alesina and Passarelli, 2019). Unique to this paper, is the focus on reference dependence as a potential pathway for democratic backsliding.

3. Baseline Model

A polity is composed of a unit mass of voters indexed by i ("she"), and is ruled by an incumbent I ("he"). The interaction has three stages: challenge, policy, and election.

First, I chooses whether to challenge democratic norms ($c = 1$, for example announcing a *prima facie* unconstitutional measure, or that he will disregard judicial review of such measure) or whether to operate within the constraints of democracy ($c = 0$). Subsequently, he chooses a policy from the interval $\mathcal{Y}(c)$: challenging democracy 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, his subsequent policy choice will be constrained to $y = 1$.

Conversely, $\mathcal{Y}(1) = [1 + \delta, 2]$: if he challenges, he can then choose how much to double down on his initial move in order to move policy in a more extreme direction. The variable $d \in [\delta, 1]$ captures the additional room for policy escalation associated with a challenge. A pure strategy by an incumbent can be then described by a pair (c, d), with $y(c, d) = 1 + cd$. When $d = 1$, the incumbent chooses full escalation. When $d = \delta$, the incumbent chooses not to further escalate. We interpret the parameter $\delta \in (0, 1)$ as the strength of checks and

balances, or how much push-back the incumbent receives 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 the institutional conflict.[2]

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

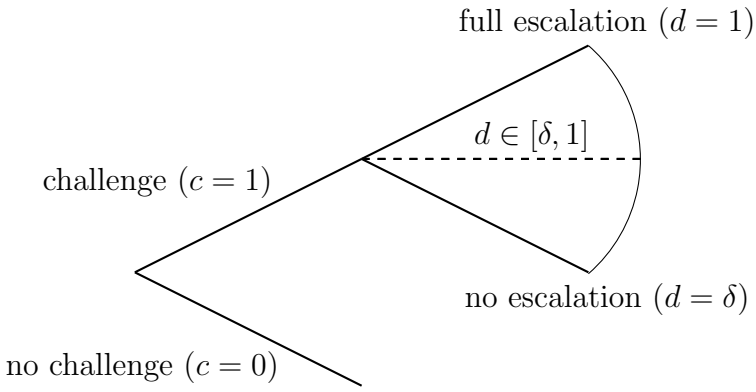


Figure 1: Incumbent's Sequential Decision Problem.

Voters 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 voter i evaluates policy outcomes $y(c, d)$ in light of her ideology θ_i, reflected in the term $\theta_i y(c, d)$. We assume that θ_i is distributed in the population according to a cumulative density function F. Voters with a positive (negative) ideology favor (oppose) a movement in policy in the direction of the Incumbent. 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, voters' *material utility* is given by:

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

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

[2]The assumption that the challenge to democratic institutions is associated to a discrete jump in policy plays a crucial role in our theory. It also captures the idea that authoritarian backsliding is a gradual process that often starts with institutional reforms (e.g., *de jure* or *de facto* weakening of independent authorities and separation of powers, impairment of the oppositions' ability to contrast the government) which if successful paves the way to more extreme policy measures.

The parameter ψ captures the degree of ideological homogeneity in the electorate: lowering ψ increases the mass of voters with extreme political preferences.[3] The upper bound on ψ guarantees that there is a small but positive minority of voters who support full escalation against democratic institutions.[4] Assumption 1 also implies that the voters who are willing to accept some weakening democratic of norms in exchange for a more radical policy shift (i.e., with $\theta_i > 1$) are a minority. A majority of the electorate is averse to democratic backsliding.

Like voters, the Incumbent cares about policy: we denote his ideology by θ_I. In addition, I also values reelection; her utility function is given by

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

where $\pi(\mathbf{q})$ is the Incumbent's vote share and $R \in \mathbb{R}_+$ measures the importance of office motivation, i.e., the Incumbent's electoral concern. The Incumbent knows his ideology θ_I, but the voters 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 ϕ measures voters' 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 electoral incentives (i.e., choose the policy maximizing their policy payoff).

Once the Incumbent has chosen the policy vector $\mathbf{q}$, elections are held. We assume that voting is retrospective. In particular, voters' electoral behavior depends on their *total utility*, which is the sum of her material utility, $u(\mathbf{q}; \theta_i)$, and an additional psychological component capturing reference dependence. The psychological component depends on how much the

[3]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.

[4]The assumption also simplifies the exposition by ensuring that the incumbent's vote share is always interior.

utility experienced by voter i exceeds or falls short of her reference point, $\underline{u}$. When this gap is positive, voter i experiences a psychological gain; when it is negative, she suffers a psychological loss. Parameter $\eta \in \mathbb{R}_+$ captures the relative importance of this psychological component relative to a voter'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 voter's expected utility given the behavior of the incumbent *following his 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 voter with ideology θ_i when she observes c is equal to:

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

A voter's reference point is determined upon observing the choice of c. As such, the challenge against democratic institutions has two consequences: (i) it changes the set of policy choices available to the Incumbent in the following period and (ii) it triggers a thought process about the ultimate consequences of the Incumbent's actions, which leads to the formation of the reference point.

An equilibrium is a profile $(\hat{\mathbf{q}}, \underline{u}(0; \hat{q}, \theta_i), \underline{u}(1; \hat{q}, \theta_i))$ 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 a voter's electoral behavior—and thus the behavior of the office-motivated incumbent—, on the other hand, the behavior of the incumbent feeds back in the reference point.

4. Analysis

Given retrospective voting, a voter with ideology θ_i votes for the Incumbent if and only if $v(\mathbf{q};\theta_i) \geq 0$, and votes for an un-modeled non-strategic challenger otherwise.[5] The Incumbent's vote share is thus equal to

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

In our model, incumbents' behavior is driven by two sets of concerns: (i) policy concerns, i.e., how their behavior affects their policy utility, and (ii) electoral concerns, i.e., how their behavior affects their electoral support. Electoral concerns, in turn, respond to two distinct mechanisms: (a) how the incumbent's behavior affects voters' material payoff and (b) how it affects voters' *psychological payoff*. The novelty of our contribution lies in the third channel. To clearly understand how these three channels operate, we introduce them sequentially. We begin with the benchmark case of no electoral concern ($R = 0$). We then introduce electoral concerns in the absence of reference dependence ($R > 0$ and $\eta = 0$), and we finally describe the novel incentives that reference dependence generates.

The Incumbent's Policy Concerns

When $R = 0$, the incumbent's behavior does not respond to the electoral consequences of his actions. In the absence of accountability to public opinion, 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

[5]The specific way in which voters 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 any constant $\underline{v}$.

incumbents with $\theta_I \leq 1$ as *liberals*: their intrinsic preferences lead them to respect democratic norms.

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

Electoral Incentives without Reference Dependence

Now, suppose that the incumbent is office motivated ($R > 0$), but voters do not exhibit reference dependence ($\eta = 0$). In this case, electoral concerns are entirely driven by voters' material payoffs. Only voters with $u(\mathbf{q}; \theta_i) \geq 0$ will support the incumbent. Since a majority of voters oppose authoritarian backsliding (i.e., for a majority of voters $u(\mathbf{q}; \theta_i) = \theta_i + (\theta_i - 1)cd$ is decreasing in both c and d), challenges to democratic norms are necessarily associated with an electoral loss. When the incumbent respects democratic norms, his vote share equals $\pi(0, 0) = F(\theta_1 \geq 0) = 1 - F(0) = \frac{1}{2}$. When he chooses to challenge them, more voters abandon him, and this electoral loss 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 an electoral cost, all liberal types must choose to respect democratic norms. Autocratic types, conversely, face a trade off: democratic backsliding increases their policy utility by $d(\theta_I - 1)$, but it is also associated to an electoral cost of $R\psi\frac{d}{1+d}$. Only autocratic types that are extreme enough will choose to violate norms. Because the electoral loss is concave on the level of escalation, conditional on challenging these norms, they will choose full escalation.

The electorate then acts as a check on leaders' autocratic ambitions. The electoral punishment generates a measure of *opportunistic democrats*—autocratic types that are induced to respect democratic norms by the threat of electoral punishment. These are essentially the same driving forces described in the existing formal literature on democratic backsliding (Svolik, 2019; Nalepa, Vanberg and Chiopris, 2018). The idea, however, has deeper roots: it directly links to a key argument for the centrality of electoral institutions in a democratic regime (Schumpeter, 1942; Popper, 1945).

Opportunistic democrats are those for which $u_I(1,1;\theta_I) \leq u_I(0,0;\theta_I)$, or

$$\theta_I + \frac{R}{2} \geq \theta_I + (\theta_I - 1) + \frac{R}{2} - \frac{R\psi}{2} \Leftrightarrow \theta \leq 1 + \frac{R\psi}{2} := \theta^\dagger$$

Proposition 2. *Suppose the Incumbent is office-motivated ($R > 0$), but the electorate does 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 electoral incentives. Autocrats with ideology in $(1, \theta^\dagger]$ will not challenge democratic institutions despite their primitive preference for doing so. Crucially, the power of this disciplining effect is increasing with office motivation (R) and decreasing with voters' ideological dispersion. In line with Nalepa, Vanberg and Chiopris (2018) and Svolik (2019), polarization limits voters' responsiveness and thus mitigates the electoral cost of 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 patters, (iii) creates incentives for liberal types to engage

in some form of democratic backsliding and (iii) significantly complicates how factors such as office motivation and electoral responsiveness affect the likelihood of democratic backsliding.

Reference Dependence and Opportunistic Authoritarians

We now consider the case in which an office-motivated incumbent ($R > 0$) faces voters who exhibit reference dependence ($\eta > 0$).

As discussed above, reference points are determined by voters expectations following the Incumbent's decision on whether to challenge or not, $\underline{u}(0; \hat{\mathbf{q}}, \theta_i) = E_{\hat{\mathbf{q}}}\left[u(\mathbf{q}; \theta_i) \mid c = 0\right]$ and $\underline{u}(1; \hat{\mathbf{q}}, \theta_i) = E_{\hat{\mathbf{q}}}\left[u(\mathbf{q}; \theta_i) \mid c = 1\right]$—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 an initial decision:

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

When no confusion arises, we omit specifying the dependence on $\hat{\mathbf{q}}$ and we define $E_{\hat{\mathbf{q}}}[d \mid c = 1] := \underline{d}_1 \in [\delta, 1]$.

If the incumbent chooses not to escalate (i.e., $c = 0$), voters face no uncertainty regarding the policy choice. Hence, the total utility of a voter is equal to her ideology: $v(0, d; \theta_i) = \theta_i$, the incumbent's vote share is equal to $1/2$ and his utility equals

$$u_I(0, 0; \theta_I) = \theta_I + \frac{R}{2}. \tag{7}$$

If instead the Incumbent challenges democratic institutions, voters' electoral behavior depends on the expected level of escalation, $\underline{d}_1$, which is determined in equilibrium. In particular, fixing an expected and actual level of escalation, $\underline{d}_1$ and d respectively, a voter with

ideology θ votes for the Incumbent if and only if

$$\begin{aligned} v(1,d;\theta) &= \theta + (\theta-1)d + \eta\Big[\theta + (\theta-1)d - \theta - (\theta-1)\underline{d}_1\Big] \\ &= \theta + (\theta-1)\Big[(1+\eta)d - \eta\underline{d}_1\Big] \geq 0. \end{aligned} \tag{8}$$

In the body of the paper, we assume that the strength of checks and balances is 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.

$$\delta > \frac{\eta - 1/2}{1+\eta} \tag{9}$$

Substantively, the assumption guarantees that in equilibrium a voter's propensity to reelect the Incumbent is increasing in her ideology.[6]

Assumption 3 implies that when $c = 1$ the Incumbent's vote share is interior and equals

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

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

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

Notice, however, that $\pi(1,d)$ is not necessarily lower than $\pi(0,0) = \frac{1}{2}$. The reason is that that voters' reference point might go down following the incumbent's decision to challenge democratic norms. Comparing (7) and (11), we can identify the potential trade-off faced by

[6]Since the Incumbent's average ideology is positive, the assumption ensures that more right-wing voters are more likely to vote for a right-wing Incumbent than less right-wing voters. In the Appendix, we show that when δ is too small, the relationship between ideology and voting can flip. While theoretically interesting, for expositional clarity we chose to confine the analysis of this case to the Appendix.

an Incumbent when she decides whether to challenge or not. If she challenges institutions, she might shift the policy (which she likes if $\theta_I > 1$), but this also entails an electoral cost in terms of a reduction in the Incumbent's vote share.

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

The policy drift implies that a challenge to democratic norms allows the Incumbent to set a more extreme policy. But it also changes voters' electoral response through an electoral feedback: it directly lowers the policy payoff of most voters, but it also lowers 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 voters' utility. Consider first the case in which η is very low.[7]

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

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

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

In this case, voters' equilibrium expectations are $\underline{d}_1 = 1$ (that is, full escalation conditional on observing a challenge) and the equilibrium utility of the incumbent is then equal to:

$$u_I^\star(\theta_I) = \begin{cases} \theta_I + \frac{R}{2} & \text{if } \theta_I < \theta^\dagger \\ 2\theta_I - 1 + \frac{R}{2}\left[1 - \psi\right] & \text{if } \theta_I \geq \theta^\dagger \end{cases} \tag{13}$$

In the settings covered by Proposition 3, all incumbents who challenge democratic institutions then decide to fully escalate. Because voters' reference point is determined in equilibrium, $\underline{d}_1 = 1$. Hence, voters are not positively, nor negatively surprised and the cutoff type $\overline{\theta}_I$ who is indifferent between challenging and not challenging is still $\theta^\dagger$.

[7]Notice that if the condition on η stated in proposition 3 holds, Assumption (3) holds as well.

To understand why this equilibrium requires reference dependence not to be too important for voters' utility, note that conditional on challenging, the incumbent has an electoral benefit from choosing $d = \delta$. If voters are expecting full escalation, the choice not to escalate comes as a positive surprise for the majority of voters and this yields an electoral benefit. This behavior is particularly tempting for autocratic incumbents with less extreme ideologies, namely Incumbents with θ_I close to $\theta^\dagger$. Inequality $\eta \leq \delta/(2-\delta)$ guarantees exactly that type $\theta^\dagger$ strictly prefers to play according to the equilibrium strategy rather than to reap the electoral benefits associated to positively surprising voters. Indeed, when reference dependence has a small impact on voters' utility (i.e., when η is low), the extent of voters' relief is limited and incumbents do not engage in this strategic behavior. Note that the cutoff for η is increasing in δ. This is intuitive: as the strength of 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 $(d = \delta)$. Moreover, because the Incumbent's utility satisfies the single crossing condition, the level of escalation chosen by the incumbent must be weakly increasing in her ideology. Hence, the following proposition holds.

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

$$\eta \geq \frac{\delta}{2-\delta}. \tag{14}$$

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 voters' reference point following a challenge is given by

$$\underline{d}_1 = 1 - (1-\delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} \tag{15}$$

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 (23) and (24) in the appendix). In particular, $\overline{\theta}$ is always greater than 1, while $\underline{\theta}$ can also be lower than 1.

Opportunistic authoritarians. The characterization of these cutoffs further implies that $\underline{\theta} < 1$ happens when there is sufficient uncertainty concerning the Incumbent's ideology. In this case, some liberal incumbents (i.e., incumbents with $\theta_I \leq 1$) challenges democratic institutions even though this goes against their policy preferences. We refer to these incumbents as to *opportunistic authoritarians*. To understand why opportunistic authoritarians exist, consider the electoral feedback identified by (12). When the ideology of the Incumbent is sufficiently uncertain (i.e., ϕ small), incumbents with extreme ideologies are substantially likely. Hence, a voter who observes a challenge will expect full escalation with high probability, $\underline{d}_1 \simeq 1$. Then, An incumbent who chooses to challenge and not to escalate, $(c, d) = (1, \delta)$, enjoys the electoral benefit associated with voters's relief. When this electoral benefit is sufficiently strong, the Incumbent may be willing to pay a cost in terms of the policy implemented: she may accept $d = \delta$, while $d = 0$ would be optimal from her perspective.

Importantly, and somehow paradoxically, when opportunistic authoritarians exist, stronger electoral responsiveness (measured either as an increase in the relative importance of office motivation, R, or in the responsiveness of voters' behavior to their realized payoff, ψ) may exacerbate this strategic behavior pushing for a decrease in $\underline{\theta}$. Thus, in our setting, electoral incentives may induce some liberal Incumbents to display authoritarian tendencies despite the fact that in a model with unaccountable leaders, they would not (cf. Proposition 1). This goes against not only their intrinsic preferences, but also the interests of voters.

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 sufficiently strong, there are opportunistic authoritarians.*

Proposition 5 implies that challenging democratic norms becomes an electorally appealing strategy colorblue for a liberal incumbent only when (i) reference dependence is sufficiently strong and (ii) voters 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 voters' utility through the reference-dependent component, namely as a rise of η. Proposition 5 then says that this may increase the likelihood of opportunistic authoritarians. In Appendix 7.3, we confirm this intuition in a more rigorous way by allowing voters to rationally choose their level of attention (or inattention) toward the incumbent's behavior.

Figure 2 below summarizes the incumbent's equilibrium behavior under the assumptions.[8] 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 ($\theta_I \in (1, \theta^\dagger]$) behave as liberal incumbents. In short, electoral accountability discipline some incumbents and generate closet autocrats.

However, if the importance of reference dependence is sufficiently large (i.e., if $\eta > \frac{\delta}{2-\delta}$), the incumbent uses it strategically: a subset of relatively moderate autocrats with ideology in the interval $(\underline{\theta}, \overline{\theta}]$ finds it optimal to challenge democratic norms without escalating to cash in the electoral gain associated with voters' relief (cf. Proposition 4). When compared to the case of low reference dependence, this choice has two implications. On the one hand, previously authoritarian autocrats are partially disciplined: incumbents with ideology between $\theta^\dagger$ and $\overline{\theta}$ (highlighted in blue in Figure 2) choose not to escalate ($d = \delta$) instead of full escalation ($d = 1$). On the other hand, previously disciplined (i.e., closet) autocrats are encouraged behave in a more authoritarian way: incumbents with ideology in $(\theta^\dagger, \underline{\theta}]$

[8]Recall that Assumption 3 puts an upper bound on η. See Section 7.1 for a characterization of the equilibrium when Assumption 3 fails.

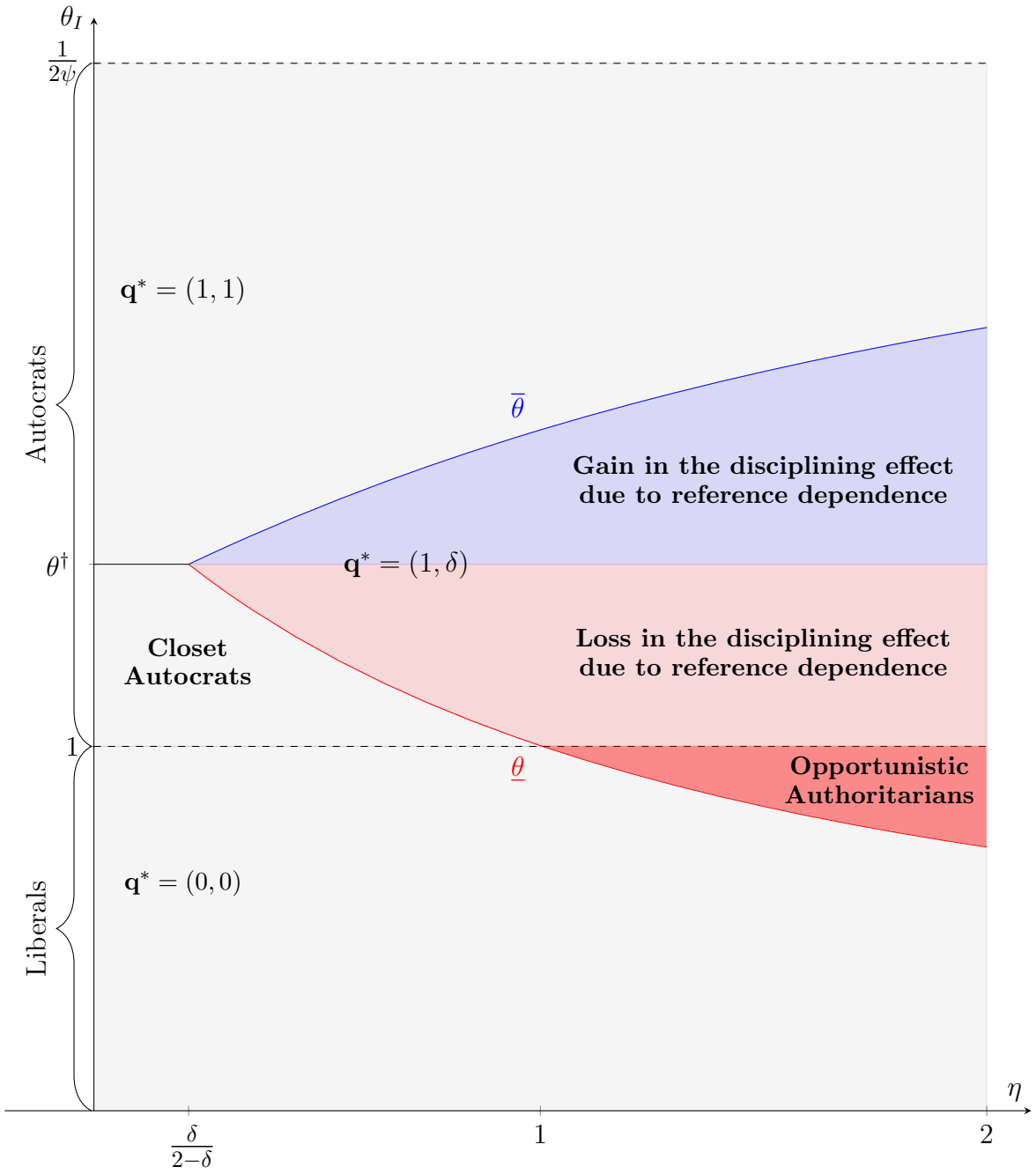


Figure 2: Incumbent's equilibrium Behavior as a function of his type θ and the importance of reference dependence η (parameter values $\psi = 0.2$, $\tau = 0.5$, $\phi = 0.25$, $R = 4$ and $\delta = 0.35$).

(highlighted in light red in Figure 2) begin to challenge democratic norms. As the importance of reference dependence keeps increasing, this latter effect can lead to the disappearance of closet autocrats and to the appearance of opportunistic authoritarians (highlighted in dark red in Figure 2): liberal incumbents with ideology in the interval $(\underline{\theta}, 1]$ challenge democracy

because they want to gain from voters' relief. Somehow paradoxically, these incumbents would not challenge democratic institutions in the absence of electoral accountability, but they do so order to enhance their reelection prospects. Hence, under the conditions of Proposition 5, electoral incentives can have a subtle and counterproductive effect.

5. Discussion

Challenges without Doubling Down

Even when the effect of reference dependence on the incumbent's electoral incentives does not yield opportunistic authoritarians, it still modifies the equilibrium of Proposition 3 in an important dimension: when a challenge occurs, it does not necessarily lead to full escalation. This is because the positive electoral incentive described above motivates autocrats on both sides of $\theta^\dagger$ to challenge and then to choose $d = \delta$. On the one hand, autocrats with ideology in the interval $(\underline{\theta}, \theta^\dagger]$ will challenge democratic institutions when they would have not done so in the absence of reference dependence. Reference dependence is thus weakening the disciplining effect exercised by electoral accountability.[9] On the other hand, autocrats with ideology in the interval $(\theta^\dagger, \overline{\theta}]$ do not escalate, while they would have chosen full escalation in the absence of reference dependence. In this case, reference dependence strengthen the disciplining effect played by electoral accountability.

The incentive to choose escalation level $d = \delta$ as opposed to $d = 1$ can become so strong to induce even the Incumbents with the highest possible ideology to choose this action. When this happens $\underline{d}_1 = \delta$. In such scenario the incumbent who is indifferent between choosing $c = 0$ or choosing $c = 1$ and then escalating at level δ has ideology

$$\theta_I = 1 + \frac{\psi R}{1+\delta} := \theta^\ddagger > \theta^\dagger \tag{16}$$

[9]The case of opportunistic authoritarians can be regarded as the extreme case in which reference dependence leads to a complete reversal of such disciplining effect.

In this case electoral concerns discipline the Incumbent and induce autocrats with ideology in the interval $(1, 1+\psi R/(1+\delta)]$ not to challenge. Moreover, even when a challenge occurs, the electoral incentive prevents full escalation. This type of equilibrium arises when even the Incumbent with the highest possible ideology prefers not to escalate, which occurs when

$$\left(\tau + \frac{1}{2\phi} - 1\right) \leq \frac{\psi R}{1+\delta} \cdot \frac{1+\eta}{1+1+\eta(1-\delta)}, \tag{17}$$

which is satisfied when δ is sufficiently low. Because its right-hand side is positive, (17) is always satisfied when $\phi \geq 1/(2(1-\tau))$. Instead if $\phi < 1/(2(1-\tau))$, (17) can hold only if η is sufficiently high and, $(\tau + \frac{1}{2\phi} - 1) \leq R/(2(1-\delta^2))$.

The effect of polarization. Previous scholarship has singled out political 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, polarization plays a more subtle role in our theory.

When either (i) reference dependence is sufficiently weak, or (ii) voters are not too uncertain about the incumbent's policy positions (i.e., ϕ is not too small), polarization does increases the likelihood of democratic backsliding. The reason is that higher polarization (i.e., lower ψ) reduces the electoral punishment associated with violating democratic norms, thereby reducing electoral accountability: fewer autocrats are deterred by the electoral punishment associated with democratic backsliding.

However, when reference dependence is strong enough and voters are sufficiently uncertain about the incumbent's policy positions, (i..e, the assumptions of Proposition 5 hold), higher polarization among voters reduces the likelihood opportunistic authoritarians.[10] The reason is that weakening voters' responses to incumbent behavior, polarization reduces politicians' incentives to try to lower voters' expectations (and then electorally benefit from their relief) by challenging democratic norms. Hence, when opportunistic authoritarians arise, polariza-

[10]Formally, lower ψ pushes both $\underline{\theta}$ and $\overline{\theta}$ closer to 1. An immediate corollary of this is that, if opportunistic authoritarians exhist, an increase in polarization reduces their likelihood.

tion decreases the overall likelihood of democratic backsliding *and* it increases its severity conditional on occurring (i.e., it increases the likelihood of escalation conditional on a challenge occurring).

Checks and Balances

Our model also illustrates how the strength of 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 implications, we focus on the more innovative part of our theory: the case of strong reference dependence (i.e., when Proposition 4 holds).[11] The first consequence of stronger checks and balances is that challenges to democracy are in expectation less damaging: conditional on incumbents not doubling down, voters are better off as δ goes down. As a consequence, when the incumbent is either an opportunistic authoritarian or a closet autocrat voters are going to be better off.

Proposition 7 in Appendix 7.2, however, shows that checks and balance also affect the likelihood and intensity of these challenges. Also in line with the conventional wisdom (however, the mechanism is novel), stronger checks and balances generally increase the disciplining effect of elections and increase the likelihood of closet autocrats: the reason is that the relief that voter experience when an incumbent backs down from a challenge is higher—and so is the electoral benefit from doing so.[12] Contrary to the conventional wisdom, and for the same reason, stronger checks and balances also encourage opportunistic authoritarians. As a result, stronger checks and balances always reduce the severity of democratic backsliding,

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

[12] 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.

but they also increase (strictly, when opportunistic authoritarians arise) the likelihood of democratic backsliding.

Rational Inattention

In Appendix 7.3, we analyze a simplified extension of the model with rationally inattentive voters. Specifically, we assume that voters can choose their level of attention to politics, which in turns increases the probability that they observe the incumbent's actions (we impose weak assumptions on how exactly it affects the likelihood of observing c and d).

Holding the behavior of the incumbent constant, attention is always valuable for the voter: it improves her ability to estimate the ex-post payoff from reelecting the incumbent, and thus improves her electoral choice. However, voter attention also feeds back into incumbents' incentives, and its effect is crucially mediated by reference dependence. Generally speaking, more attention increases the voter's responsiveness to the incumbent's action, similarly to a decrease in polarization. The importance of reference dependence governs how this increased responsiveness shape incumbent behavior, but its overall effect on the voter's *ex-ante* payoff is ambiguous. On the one hand, higher attention increases the likelihood that the voter 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 who challenges but does not escalate will manage to lower voters' expectations and benefit from the electoral boost idenfied in Proposition 5.[13]

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, a very attentive electorate genenerates 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.

[13]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).

6. Conclusion

This paper presents a theory of democratic backsliding where voters and most incumbents intrinsically dislike violations of democratic norms and yet, these violations do not always result in an electoral sanction.

When (i) voters 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 voter polarization, the strength of checks and balances, and voter 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 voters' attitudes: even if most voters 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 voters to focus on the fact that "it could have been worse," all the while things continue to get worse.

References

Abeler, Johannes, Armin Falk, Lorenz Goette and David Huffman. 2011. "Reference points and effort provision." *American Economic Review* 101(2):470–92.

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.

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.

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

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 et al. 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.

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* .

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.

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. 2015. Negative voters: Electoral competition with loss-aversion. Technical report.

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

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.

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.

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

Schwarz, Michael and Konstantin Sonin. 2007. "A theory of brinkmanship, conflicts, and commitments." *The Journal of Law, Economics, & Organization* 24(1):163–183.

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.

Tversky, Amos and Daniel Kahneman. 1991. "Loss aversion in riskless choice: A reference-dependent model." *The Quarterly Journal of Economics* 106(4):1039–1061.

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.

7. Appendix

7.1 General Characterization

In the main text we analyzed what happens when checks and balances are sufficiently strong. 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 $v(\mathbf{q};\theta)$ in (3) . If $d > (<)\ d^\circ(\underline{d}_1)$, (8) implies that $v(\mathbf{q};\theta)$ is increasing (decreasing) in θ. Instead, if $d = d^\circ(\underline{d}_1)$, $v(\mathbf{q};\theta) = 1$ and thus the vote share of the Incumbent is equal to 1.[14] By continuity, we can then define an interval around $d^\circ(\underline{d}_1)$ such that when d falls in this interval, then the vote share of the Incumbent is equal to 1. To characterize the vote share of the Incumbent, let

$$\theta^*(d, \underline{d}_1) = \min\left\{\max\left\{\frac{d+\eta(d-\underline{d}_1)}{1+d+\eta(d-\underline{d}_1)}, -\frac{1}{2\psi}\right\}, \frac{1}{2\psi}\right\}. \tag{18}$$

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

$$d^\circ_\ell(\underline{d}_1) = \frac{\eta \underline{d}_1 - (1-2\psi)^{-1}}{1+\eta}, \tag{19}$$

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

$$d^\circ_h(\underline{d}_1) = \frac{\eta \underline{d}_1 - (1+2\psi)^{-1}}{1+\eta}. \tag{20}$$

Then, the following proposition holds.

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

Proposition 6. *Let $\underline{d}_1$ be the reference point of the voters. Then, the Incumbent's vote share is equal to*

$$\pi(1, d \mid \underline{d}_1) = \begin{cases} \frac{1}{2} + \psi\theta^\star(d, \underline{d}_1) & d < d_\ell^\circ(\underline{d}_1); \\ 1 & d \in [d_\ell^\circ(\underline{d}_1), d_h^\circ(\underline{d}_1)]; \\ \frac{1}{2} - \psi\theta^\star(d, \underline{d}_1) & d > d_h^\circ(\underline{d}_1). \end{cases} \tag{21}$$

The vote share 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 (8) and (18). Instead, the properties of the vote share 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 6, Assumption 3 in the main text restricts attention to the case in which the vote share 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 vote share is increasing in the level of extremism. To understand why, observe that when the Incumbent chooses policies that are not too extreme, all voters with low ideology will support him. Instead, voters with high ideology will not because they would rather pick higher values of δ; hence, when δ increases, part of these latter voters will support the Incumbent yielding an increase in his vote share.

Because the level of escalation is bounded below by δ and Proposition 6 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^*.[15] Depending on the value of δ, D^* is one of three possible sets. Figure 3 depicts the set of voters 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: $\delta > d_h^\circ$. In this case, $D^* = \{(0,0), (1,\delta), (1,1)\}$.

Case 2: $\delta \in (d_\ell^\circ, d_h^\circ]$. In this case, $D^* = \{(0,0), (1,\delta), (1,d_h^\circ), (1,1)\}$.

Case 3: $\delta \leq d_\ell^\circ$. In this case, $D^* = \{(0,0), (1,\delta), (1,d_\ell^\circ), (1,d_h^\circ), (1,1)\}$.

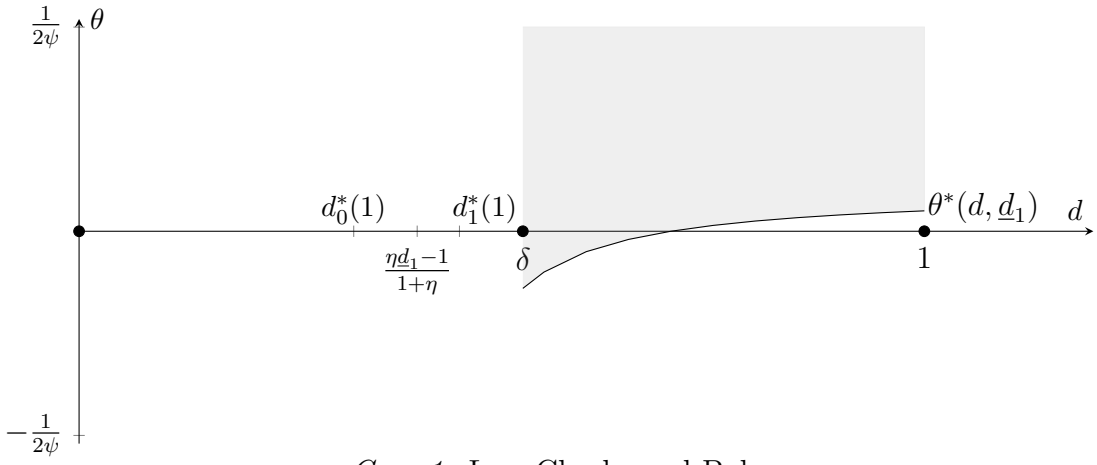
When δ is sufficiently large (i.e., $\delta \geq (2\eta - 1)/(2(1+\eta))$), (20) implies that $\delta > d_h^\circ(\underline{d}_1)$ independently of δ and of the voters' expectations. Hence, the relevant case is the third one, which is analyzed in the main text. We will now consider the other two possible cases.

Suppose we are in case 2, thus $\delta \in (d_\ell^\circ, d_h^\circ]$. Abstracting from office motivation, Incumbents with ideology equal to 1 would be indifferent between all levels of escalation. Thus, their behavior would hinge on office motivation and would then be indifferent between any level of $\delta \in [\delta, d_h^\circ]$. Indeed all such levels of escalation would maximize the Incumbent's vote share. 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:

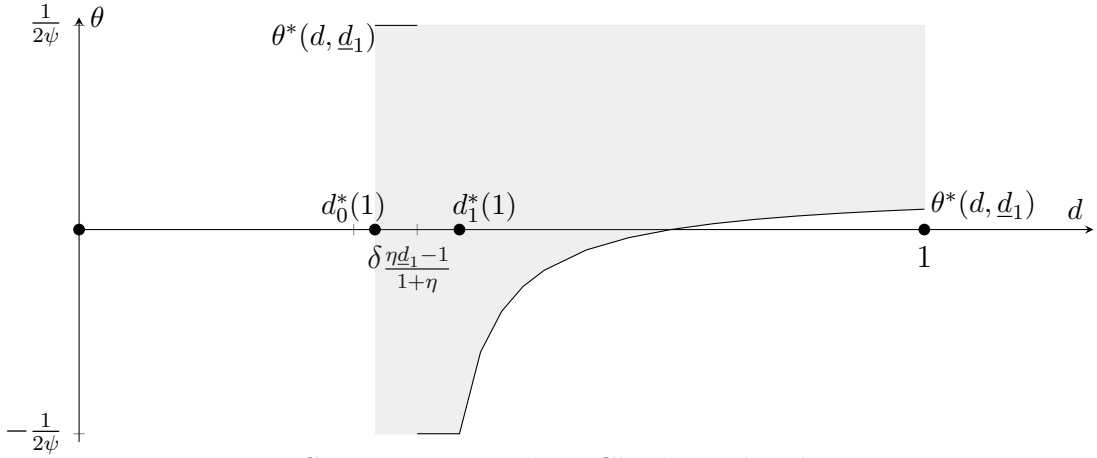
- if $\theta_I \leq \underline{\theta}$, then the Incumbent chooses $(0,0)$;
- if $\theta_I \in (\underline{\theta}, 1]$, then the Incumbent chooses $(1,\delta)$;
- if $\theta_I \in (1, \tilde{\theta}]$, then the Incumbent chooses $(1, d_h^\circ(\underline{d}_1))$;
- if $\theta_I > \tilde{\theta}$, then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

Obviously, this equilibrium exists as long as the implied $\delta \in (d_\ell^\circ(\underline{d}_1), d_h^\circ(\underline{d}_1)]$. Moreover, $\underline{\theta}$ is defined as the ideology of the Incumbents who are indifferent between choosing not to

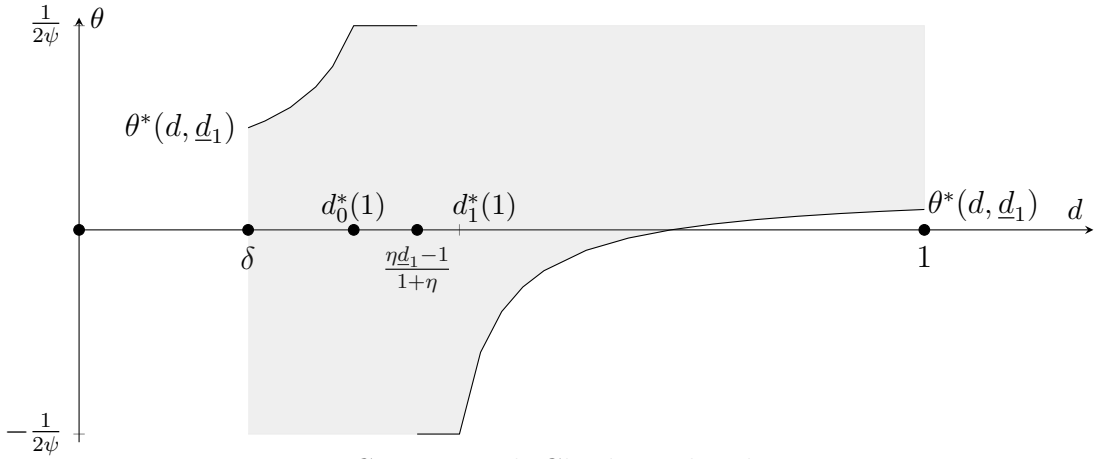
[15] An Incumbent with ideology equal to $\theta_I = 1$ may have a continuum of optimal strategies. However, because these types have measure, this is without consequence for the subsequent argument.



Case 1: Low Checks and Balances



Case 2: Intermediate Checks and Balances



Case 3: High Checks and Balances

Figure 3: 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)

challenge and those who challenge but then do not double down. Similarly, $\tilde{\theta}$ is defined as the ideology of the Incumbents who are indifferent between choosing a level of extremism equal to d_h° or equal to 1. Hence, 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, opportunistic authoritarians always exist.

Finally, suppose that checks and balances are sufficiently low to fall in the Case 3 of Figure 3. In this case, the same reasoning described above implies that Incumbents with ideology close but below (above) $\theta_I = 1$ will choose the lowest (highest) level of extremism that guarantees full vote share, $d_\ell^\circ(\underline{d}_1)$ ($d_h^\circ(\underline{d}_1)$). Differently from the case 2, however, Incumbents with ideology lower than 1 who decides to challenge democratic norms can now choose two possible levels of extremisms: δ or $d_\ell^\circ(\underline{d}_1) > \delta$. If the reference point is $\underline{d}_1$, then 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

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

Thus we can have two possible equilibrium configurations. In the first one, there is no mass of incumbents choosing $d = \delta$ after challenging. In this case, we can define two cutoffs, $\underset{\sim}{\theta} < \tilde{\theta}$, and describe the behavior of the Incumbent 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 cutoffs as well as the reference points are defined in equilibrium by the following system of three 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, instead, a positive mass of incumbents chooses δ. Thus, we have three cutoffs, $\underline{\theta} < \underset{\sim}{\theta} < \tilde{\theta}$ such that the behavior of the Incumbent can be summarized as follows:

- if $\theta_I \leq \underline{\theta}$, then the Incumbent chooses $(0, 0)$;
- if $\theta_I \in (\underline{\theta}, \underset{\sim}{\theta}]$, then the Incumbent chooses $(1, \delta)$;
- 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 $(1, 1)$.

In this case, the cutoffs and the reference point are characterized by the following system of 4 equations in 4 unknowns:

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

Observe that, independently of the actual equilibrium played, also in this third case there is always a range of Incumbents who, despite being liberal, challenge democratic institutions because of electoral concerns. In other words, also in this case opportunistic authoritarians always exist. The range is given by $(\underset{\sim}{\theta}, 1]$ in the first equilibrium and by $(\underline{\theta}, 1]$ in the second equilibrium.

7.2 Proofs

Proof of Proposition 1. Absent electoral 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:

$$\eta \leq \frac{\delta}{2 - \delta}, \tag{22}$$

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 (22) 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 11) 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 eacalate 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 (15). Furthermore, $\underline{\theta}$ satisfies

$$\delta(\underline{\theta} - 1) = \frac{R\psi[\delta(1+\eta) - \eta\underline{d}_1]}{1 + \delta(1+\eta) - \eta\underline{d}_1} \tag{23}$$

while $\overline{\theta}$ satisfies:

$$(\overline{\theta} - 1) = \frac{R\psi(1+\eta)}{[1+1+\eta(1-\underline{d}_1)][1+\delta+\eta(\delta-\underline{d}_1)]}. \tag{24}$$

In this case, we immediately get that

$$\underline{d}_1 = 1 - (1-\delta)\frac{2(\overline{\theta} - \underline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} = \delta + (1-\delta)\frac{1 + 2(\tau - \overline{\theta})\phi}{1 + 2(\tau - \underline{\theta})\phi} \tag{25}$$

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

$$\frac{R\psi}{1+\delta+\eta(\delta-\underline{d}_1)}\left[\eta\frac{\underline{d}_1}{\delta} - (1+\eta)\frac{1+\eta(1-\underline{d}_1)}{1+1+\eta(1-\underline{d}_1)}\right] \geq 0 \tag{26}$$

Observe that the right-hand side of (23) is decreasing in $\underline{d}_1$ and equals $\delta\frac{R\psi}{2}$ (in which case $\underline{\theta} = \theta^\dagger$) if and only if

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

Hence $\underline{\theta} \leq \theta^\dagger$ whenever $\eta\underline{d}_1/\delta \geq \eta + (1-\delta)/(2-\delta)$. Suppose this is indeed the case. Then, the squared bracket in (26) is bounded below by

$$\frac{1}{2-\delta}\frac{2\eta - \delta(1+\eta)}{2 + 2(1+\eta) - 3\delta(1+\eta) + \delta^2(1+\eta)},$$

which is decreasing in δ and equals 0 when $\delta = 2\eta/(1+\eta)$.[16] We conclude that whenever $\delta \leq 2\eta/(1+\eta)$ and $\underline{d}_1 \geq \delta + (1-\delta)\delta/(\eta(2-\delta))$, $\underline{\theta} \leq \theta^\dagger$. Because $\delta + (1-\delta)\delta/(\eta(2-\delta))$ is increasing in δ and equals 1 when $\delta = 2\eta/(1+\eta)$, the previous conditions are feasible and are satisfied whenever ϕ is sufficiently small. Indeed, our previous discussion implies that the equilibrium is well defined if $\underline{\theta} < \theta^\dagger$ which requires

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

This last inequality is satisfied as long as

$$[2\delta(\tau-\underline{\theta}) - 2\eta(2-\delta)(\tau-\overline{\theta})]\phi < 2\eta - (1+\eta)\delta$$

Under our parametric restrictions, the right-hand side of the previous inequality is positive. Hence, if the squared bracket on the left-hand side is negative, the condition is always satisfied. Instead, if the bracket is positive, it will be satisfied for sufficiently low values of ϕ. In particular, observe that the right-hand side of (23) and (24) are respectively decreasing and increasing in $\underline{d}_1$. Hence they are minimized and maximized when $\underline{d}_1 = 1$, in which case

$$\begin{aligned}\underline{\theta} &= 1 + R\psi\frac{\delta(1+\eta)-\eta}{1+\delta-\eta+\delta\eta}\\ \overline{\theta} &= 1 + R\psi\frac{1+\eta}{2(1+\delta)-2\eta(1-\delta)}\end{aligned}$$

This concludes the proof. □

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

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

[16] The fact that this quantity is decreasing in δ follows from (3).

In addition, opportunistic authoritarians arise when (iii) $\underline{\theta} < 1$, that is, using equation (23),

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

To prove the proposition, notice that as $\phi \to 0$, $\underline{d}_1 \simeq 1$. Then conditions (i) and (ii) can be comined 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.[17] □

Proposition 7. *When ϕ is small enough,*
(i) $\overline{\theta}$ is strictly decreasing in δ
(ii) when opportunistic authoritarians arise, $\underline{\theta}$ is strictly increasing in δ.

Proof of Proposition 7. 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 by inspection of Equation 24. To prove the second result, observe that differentiating $\underline{\theta}$ in(23), yields

$$\frac{\partial \underline{\theta}}{\partial \delta} \propto \frac{\partial}{\partial \delta}\left(\frac{1+\eta-\eta\frac{\underline{d}_1}{\delta}}{1+\delta(1+\eta)-\eta\underline{d}_1}\right) \propto -[\delta(1+\eta)-\eta\underline{d}_1]^2 + \eta\underline{d}_1 \approx -[\delta(1+\eta)-\eta]^2 + \eta \quad (27)$$

Suppose that the expression is negative. Notice that Assumption 3 requires that $\delta(1+\eta) - \eta > -1/2$ and the presence of opportunistic authoritarians requires that $\delta(1+\eta) - \eta < 0$. Together, they imply $-[\delta(1+\eta)-\eta]^2 > -\eta+\delta(1+\eta)$, which implies that (27) is positive. □

[17] Note that an excessively high η, however, may lead to the violation of condition (i) above. See Appendix 7.1 for details on what happens in this case.

7.3 Extension: Rational Inattention

Consider the following simplified extension of our baseline model:

1. There is a single 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

$$\hat{v}(r_1, r_2; \mathbf{q}) = E\left[u(\mathbf{q}) \mid r_1, r_2\right] + \eta\Big\{E\left[u(\mathbf{q}) \mid r_1, r_2\right] - E\left[u(\mathbf{q}) \mid r_1\right]\Big\} \tag{28}$$

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

$$\pi(\mathbf{q}; a) = \frac{1}{2} + \chi \hat{V}(\mathbf{q}, a) \tag{29}$$

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 $E[-cd] = -(1 - p_0)\underline{d}_1$, (30) π can be rewritten as

$$\pi(\mathbf{q}; a) = \frac{1}{2} - \chi\frac{a^2}{2\alpha} + \chi \begin{bmatrix} \rho_{12}\eta c\underline{d}_1 + (\rho_2\eta - \rho_\emptyset)(1 - p_0)\underline{d}_1 \\ -(\rho_{12} + \rho_2)cd(1 + \eta) - \rho_1 c\underline{d}_1 \end{bmatrix} \tag{30}$$

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.

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

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

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

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 $\rho_2 - \rho_{12}\frac{d_1-\delta}{\delta}$. 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 8 (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 is strictly positive.*

Proof. The voter's expected payoff as a function of her attention and the incumbent's strategy. 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

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

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

where we use the fact that $\hat{v}(\emptyset,\emptyset;\mathbf{q}) = -(1-p_0)\underline{d}_1$ for all $\mathbf{q}$

Let $w(r_1, r_2; \mathbf{q}) = E_\xi\Big[\max\{\hat{v}(r_1, r_2; \mathbf{q}), \xi\}\Big]$. We can then rewrite W as

$$W(a) = \left\{ \begin{array}{l} \rho_{12} \left\{ \begin{array}{l} p_0 w(0,0;0,0) \\ +(1-p_0-p_1) w(1,\delta;1,\delta) \\ +p_1 w(1,1;1,1) \end{array} \right\} + \rho_1 \left\{ \begin{array}{l} p_0 w(0,\emptyset;0,0) \\ +(1-p_0-p_1) w(1,\emptyset;1,\delta) \\ +p_1 w(1,\emptyset;1,1) \end{array} \right\} \\ +\rho_2 \left\{ \begin{array}{l} p_0 w(\emptyset,0;0,0) \\ +(1-p_0-p_1) w(\emptyset,\delta;1,\delta) \\ +p_1 w(\emptyset,1;1,1) \end{array} \right\} + (1-\rho_{12}-\rho_1-\rho_2) w(\emptyset,\emptyset;\mathbf{q}) \end{array} \right\}$$

where again $w(\emptyset, \emptyset; \mathbf{q})$ does not depend on $\mathbf{q}$.

To show the lemma, verify that the three expressions below are all positive.

$$p_0 w(0,0;0,0) + (1-p_0-p_1) w(1,\delta;1,\delta) + p_1 w(1,1;1,1) - w(\emptyset,\emptyset;\mathbf{q}) \tag{31}$$

$$p_0 w(0,\emptyset;0,0) + (1-p_0-p_1) w(1,\emptyset;1,\delta) + p_1 w(1,\emptyset;1,1) - w(\emptyset,\emptyset;\mathbf{q}) \tag{32}$$

$$p_0 w(\emptyset,0;0,0) + (1-p_0-p_1) w(\emptyset,\delta;1,\delta) + p_1 w(\emptyset,1;1,1) - w(\emptyset,\emptyset;\mathbf{q}). \tag{33}$$

We now show that for each type of voter learning, the voter's payoff is a random function of $\mathbf{q}$ with mean $\hat{v}(\emptyset, \emptyset; \mathbf{q}) = -(1-p_0)\underline{d}_1$.

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

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

where the last equality follows from the measurability of the electoral behavior with respect to the information set.

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

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

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

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

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

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

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

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

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

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

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

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

This completes the proof. □