AUTHORITARIAN BACKSLIDING

MONIKA NALEPA

GEORG VANBERG

CATERINA CHIOPRIS

Abstract. A prominent contemporary phenomenon is an apparent "backsliding" of democratic countries into (semi-)authoritarian practices. Importantly, such episodes unfold over time, and often involve uncertainty about the ultimate intentions of governments. Governments typically do not attempt to engage in authoritarian practices immediately, but rather initiate policies or institutional reforms that may in the future facilitate or enable actions that are inconsistent with liberal democratic practices. Building on recent work (Svolik 2017), we develop a formal model that explores both features (and their interaction) in a two-period game in which a government takes an action in period 1 that may allow for subsequent actions that are inconsistent with the rule of law in period 2. Citizens face uncertainty over the ultimate intentions of the government, and must decide whether to replace the first-period government before period 2. The model generates several insights. First, consistent with other work (e.g., Svolik 2018), it suggests that polarization among citizens and elites is an important factor in driving authoritarian backsliding. Extending this logic, the model demonstrates that the degree of polarization necessary to generate the potential for backsliding depends critically on the uncertainty facing citizens about the type of incumbent. Finally, in such a setting, citizens may support incumbent governments even if there is some risk that these governments are "closet autocrats" despite the fact that citizens are fundamentally opposed to authoritarianism. One consequence is that in our model, citizens may genuinely come to regret their electoral choices—a marked contrast from models in which citizens accept authoritarian outcomes because on balance these are preferable to the alternative. We illustrate the model's implications through an analytic narrative that focuses on the current Polish government's efforts to reform the Polish judiciary.

1. Introduction

A prominent development in recent years has been an apparent "authoritarian backsliding" in countries that appeared to be consolidated democracies, or on the way to becoming such. (Bermeo 2016; Lust and Waldner 2015; Serra 2012). Take as an example Poland and Hungary, two countries that until recently were seen as exemplifying a successful transition to democracy. Each emerged from behind the Iron Curtain after the Soviet Union's dissolution. By 1999, they had joined NATO, and five years later became full-fledged members of the European Union. By all appearances, both countries were well-ensconced among Western democracies. And yet, only a decade later, each is governed by a right-wing party that appears to be shaving away at democratic institutions and norms, raising concerns of creeping authoritarianism (Sedelmeier 2014; Jenne and Mudde 2012;

Cinar 2017). One prominent explanation for such developments, formalized by Svolik (2017), highlights the potential contribution of political polarization to authoritarian backsliding. The underlying logic is clear: Citizens may be willing to tolerate or support incumbents with authoritarian tendencies if the alternative offered by the opposition is sufficiently unattractive ideologically – a condition more easily met in a polarized environment.

We build on this explanation, but expand on it to highlight another typical feature of authoritarian backsliding: its dynamic nature. Even if incumbents have authoritarian aspirations, they usually do not pursue these openly and all at once. Instead, authoritarianism proceeds in incremental steps as current policy choices or institutional reforms are used to lay the groundwork for future authoritarian moves. For example, critics of the Polish government's reforms of the judiciary are concerned that by providing the government with more direct control over the constitutional court and the electoral commission, the government may be in a position after the reforms to rule with fewer constraints on its power, or to persecute political opponents. Similarly, a concern about the Hungarian government's changes to the country's electoral institutions has been that these reforms are designed to curtail political competition, and to preserve the party's hold on power. These dynamics matter, because, while they may be still consistent with the constitution, they constrain citizen resistance to potential authoritarian actions in the future. There typically is disagreement about the intentions behind specific policy choices or institutional reforms. Consider the Polish and Hungarian cases again. The PiS and Fidesz governments – and many of its supporters – adamantly deny that the goal of institutional reforms is to undermine democratic institutions or norms. On the contrary, both claim that they are motivated by a desire to enhance democratic accountability, and to remove the residual influence of holdovers from the former communist regime. Voters thus face a potential dilemma: On the one hand, the most direct defense against a "closet autocrat" is to remove the incumbent from power before authoritarianism has advanced too far. On the other hand, voters face genuine uncertainty about whether the incumbent is, in fact, a closet autocrat or is pursuing a sincere policy with no intention of undermining democracy.

We place this dynamic nature of authoritarian backsliding and the resulting uncertainty confronting voters at the heart of our argument. Consistent with other arguments, our model suggests that polarization is a critical factor in creating circumstances that allow for "closet autocrats" to retain a hold on power. Additionally, the model demonstrates that the degree of polarization required depends critically on the uncertainty facing voters, and the degree to which voters are concerned about authoritarianism. Finally, and perhaps most importantly, our model shows that the dynamic nature of authoritarian backsliding, and the uncertainty confronting voters, can allow for "closet autocrats" to establish a hold on power even if citizens are fundamentally opposed to authoritarianism – something that is not possible in existing models in which voters knowingly accept an autocrat. Thus, in our model, voters may experience sincere regret as it becomes apparent that they have backed an incumbent who reveals himself to be an autocrat over time.

The paper proceeds as follows. In the next section, we elaborate the argument, and present a simple model that formalizes the theoretical logic. We then derive some empirical conjectures from the model that form the basis of an empirical application to the current Polish situation. A final section concludes.

2. A theory of authoritarian backsliding

Authoritarian backsliding is a complex phenomenon, and there are, of course, multiple potential causal factors at work. According to an increasingly popular argument, recently formalized by Milan Svolik (2017), polarization in the electorate may create a situation in which citizens are willing to vote for autocratic parties despite the fact that these parties have authoritarian tendencies because citizens prefer the ideology of the autocrat to the ideology of the available alternative. Although this argument is compatible with elite polarization, it is driven by polarization of the electorate. In other words, without sufficiently high polarization in society, voting for the autocrat would not take place.

An alternative argument for voting in autocrats focuses on the supply side of authoritarian politicians and identifies elite polarization as the root cause. For example, the inability of opposition groups, including civil society groups, to coordinate on challenging authoritarian developments may enable an authoritarian power grab. An argument in the spirit of Weingast (1997) would blame for authoritarian backsliding the polarization of elites, which coupled with majoritarian institutions replaces a consociational model of democracy with a winner-takes-all scenario. In a nutshell, where the first argument underscores societal polarization, the second accentuates elite polarization.

In this paper, we explore another potential mechanism that builds on both of these intuitions. In particular, we examine how polarization in the electorate *and* among political elites can generate opportunities for backsliding. The key contribution of our approach and difference between our approach and existing ones, such as Svolik's, is to demonstrate that backsliding can happen *even if* citizens are fundamentally opposed to authoritarianism, and would not choose to support a party they *knew* to be authoritarian. Such a phenomenon may occur even if polarization in society is not very large, as a long as polarization among elites is sufficiently large. A crucial assumption in our model is that, in practice, there are dynamic elements to authoritarian backsliding that often make it difficult for citizens, especially in the early stages of these developments, to know precisely what kind of government they are facing. Governments make initial policy choices that are condemned by some as laying the foundation for authoritarian moves, but are seen by others as ideologically more polarized, but sincere policy choices that do not portend autocratic intentions. In other words, citizens must make choices about whom to support while facing uncertainty about the ultimate intentions of incumbent governments.

The key insight of our model is to show that given such uncertainty, sufficient polarization in the electorate and among political elites can lead citizens to support incumbent governments even if there is some risk that these governments are "closet autocrats" despite the fact that citizens are fundamentally opposed to authoritarianism. One consequence is that in our model, citizens may genuinely come to regret their electoral choices – a marked contrast from models in which citizens accept authoritarian outcomes because on balance these are preferable to the alternatives (Svolik 2018). Our model is consistent with recently published findings about voters' regret. In 2014, Ahlquist et al. (2018) conducted an experiment around the Hungarian election. The authors randomly assigned respondents to three treatment groups, which remained constant before and after the election: a no information treatment, an information treatment and an information + partisanship treatment. Information in this manipulation consists of providing voters with information about the nature of the electoral reforms that enabled FiDeSz to win the 2014 election.[1] In the next step, all respondents were asked about the legitimacy of the election. The authors found significant information +treatment effects in the group of new FiDeSz voters (that is voters who did not vote for FiDeSz in 2010, but voted for FiDeSz in 2014). This group of voters lowered their belief in the legitimacy of FiDeSz in a way that is consistent with out theoretical finding of voter regret. (Ahlquist et al. 2018).

2.1. **The Formal Model.** We employ a simple model to formalize the intuition underlying our argument. The key features of authoritarian backsliding we wish to capture are the following:

(1) Backsliding is a dynamic process: Governments engage in policy choices or institutional reforms that do not immediately create an autocracy, but that may open up possibilities for authoritarian moves in the future.
(2) Citizens face uncertainty about the ultimate intentions behind these governments' choices, that is, they are not sure whether the incumbent is pursuing these policies or reforms in an effort to facilitate a slide towards authoritarianism or not.

To capture these features, we assume a two period model. In the first period, an incumbent government chooses a policy in a one-dimensional policy spectrum, $p \in [0, 1]$. At the end of this period, a representative citizen chooses whether to reelect the incumbent, or to replace him with the opposition. To capture the fact that period 1 policies have implications for potential authoritarian moves in the future, we assume that in period 2, the incumbent can choose whether to make an authoritarian policy move that depends on the policy choice in period 1. Specifically, in the second period, a new "authoritarian" dimension becomes available, and the incumbent can choose an authoritarian policy $a \in [0, p]$. That is, the more extreme the first period policy p, the greater the potential authoritarian move available in period 2. In summary, the underlying political space is two-dimensional, comprised of a policy and an authoritarian dimension, but the

[1] The reforms included reducing the number of seats in the legislature, changing the formula compensating for "wasted" SMD votes in favor of large parties and gerrymandering to FiDeSz's advantage.

authoritarian dimension only becomes available in the second period, and the authoritarian space available depends on the first period policy.

We assume that there are two types of incumbent governments. An "ideological" incumbent is purely ideologically motivated, and has no interest in pursuing an authoritarian policy. The ideal point of this government is given by $X_I = \{x_I, 0\}$, where $x_I \in (0, 1)$. In words, the ideologue has a policy ideal point of x_I but no authoritarian preferences. The second type of government is a "closet authoritarian" whose ideal point is given by $X_{CA} = \{1, 1\}$. The ex ante probability that the incumbent is a closet authoritarian is given by $Pr(\text{Closet Authoritarian}) = \alpha \in (0, 1)$. We assume that both types of government have standard quadratic preferences over the two dimensional policy space, i.e., the utility of government i with ideal point $X_i = \{p_i, a_i\}$ from outcome $\{x, y\}$ is given by:

$$U_{X_i}(\{x, y\}) = -(x - p_i)^2 - (y - a_i)^2$$

We assume that the representative citizen has an ideal point $X_C = \{x_c, 0\}$, where $x_c \in (0, 1)$, i.e., the citizen is purely ideologically motivated and prefers for the government not to make authoritarian moves. Moreover, we parameterize how much the citizen cares about the authoritarian dimension by weighing this dimension by $\beta \geq 1$.[2] Thus, the utility of outcome $\{x, y\}$ for the citizen is given by:

$$U_{X_C}(\{x, y\}) = -(x - x_c)^2 - \beta(y)^2$$

Without loss of generality, we assume that the ideal point of the opposition is given by $x_O = \{0, 0\}$, i.e. the opposition is more moderate than either of the two incumbent types, and also has no authoritarian tendencies. Figure 1 provides a graphical illustration of the basic model.

Summarizing, the sequence of play is as follows:

(1) Nature chooses the type of incumbent, with $Pr(\text{Closet Authoritarian}) = \alpha \in (0, 1)$.

(2) Period 1: The incumbent government chooses a policy $p \in [0, 1]$.

(3) Period 2: The citizen updates beliefs about the type of the incumbent, and then either reelects the incumbent, or replaces him with the opposition.

[2] As we discuss in more detail below, the restriction that $\beta \geq 1$ is a substantive restriction imposed to ensure that a citizen who is certain that the incumbent is a closet authoritarian will always vote to replace the incumbent.

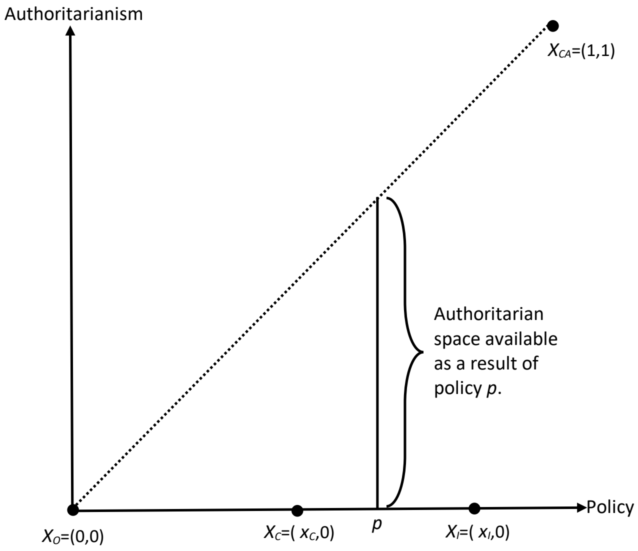


FIGURE 1.

(4) Period 2: If reelected, the incumbent can choose an authoritarian policy on the basis of its period 1 policy choice, i.e., $a \in [0, p]$.

(5) Period 2: If the incumbent is replaced, the opposition can revise the incumbent's period 1 policy choice.

(6) The game ends, and payoffs are collected.

We assume that the payoffs to all of the players are the sum of the payoffs from each of the two periods. In order to keep the analysis as simple as possible, we assume no discounting, though none of the results reported depend on this assumption.

For substantive reasons, we place a number of restrictions on the model parameters. Recall that we are interested in exploring the potential for authoritarian backsliding *even if* citizens themselves are opposed to authoritarianism. As a result, we assume that $\beta \geq 1$. Given $x_C \in [0, 1]$ this assumption ensures that no matter what a citizen's ideological preferences, her anti-authoritarian convictions are so strong that she prefers the opposition to the closet authoritarian.[3] Second, we also assume that $x_C < x_I$, i.e., that the citizen's ideal point lies somewhere between the alternatives offered by the opposition and the ideological incumbent. This assumption is designed to capture

[3] This differentiates our approach from Svolik's (2018) model in which citizens vote for the autocrat because the ideological profile of the opposition is not acceptable to them.

the fact that (decisive) citizens are typically choosing between alternatives that are on either side of citizen preferences.[4]

The appropriate solution concept is Perfect Bayesian Equilibrium, which requires that the players' strategies are sequentially rational, and that they update their beliefs (in our case, about the type of incumbent) in accordance with Bayes' rule along the equilibrium path.[5] We reserve a full statement of equilibria and proofs to the appendix, and focus on a presentation of the substantive implications and intuition underlying the equilibria here.

2.2. **Equilibria and Substantive Implications.** The model yields four types of equilibria, with a unique equilibrium for any combination of parameters. These types of equilibria are:

(1) Separating equilibria in which the ideologue and the closet authoritarian make different proposals in the first period, the citizen learns the type of incumbent she is facing, and she only reelects the ideologue and replaces the closet autocrat.
(2) Separating equilibria in which the ideologue and the closet authoritarian make different proposals in the first period, the citizen learns the type of incumbent she is facing, and replaces both with the opposition.
(3) Pooling equilibria in which the ideologue and the closet authoritarian make the same proposal in the first period, the citizen remains uncertain about the type of incumbent, and reelects the incumbent.
(4) Semi-pooling equilibria in which the closet autocrat sometimes makes the same first-period proposal as the ideologue, the citizen learns something about the type of incumbent but remains uncertain, and she reelects the incumbent with some probability.

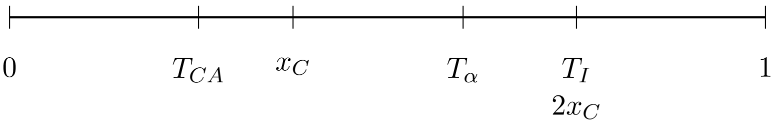
[4] Relaxing this assumption does not change the general character of the equilibrium results.

[5] A critical element in such equilibria are the off-equilibrium path beliefs of the players that support an equilibrium. Applied to our specific model, this means that how the citizen interprets "unexpected" policy proposals in the first period becomes key. Generally speaking, off equilibrium path beliefs can be specified in a large variety of ways, and this has potentially significant consequences for the equilibria that can be supported. A critical issue is therefore how to restrict the off-equilibrium path beliefs in ways that are substantively reasonable, i.e., capture plausible beliefs of the players. We (loosely) employ the logic of the intuitive criterion to restrict beliefs in two ways: First, we specify that if an off-equilibrium path proposal is only consistent with the interests of one type of incumbent, the citizen updates accordingly. If the proposal is consistent with the interests of both, the citizen's beliefs are given by the prior. If a proposal is not in the interests of either incumbent, the citizen believes that it comes from the type with the lower cost of making this "mistake." In addition, we impose the requirement that beliefs are (weakly) monotonic in the extremity of the proposal. See the appendix for details.

Before turning to the intuition behind these equilibria, and the conditions under which each occurs, consider their substantive interpretation and connection to "authoritarian backsliding." A natural interpretation of the separating equilibria is as political environments in which authoritarian backsliding does not occur. This is true for two reasons: Citizens are unwilling to elect candidates (or parties) known to be closet autocrats, and – given the political environment – closet autocrats cannot (or will not) disguise themselves as ideologues in order to gain office. As a result, closet autocrats are effectively isolated. No authoritarian moves occur in the second period. In contrast, the pooling (and semi-pooling) equilibria raise the specter of authoritarian backsliding. In these equilibria, the political environment allows closet autocrats to "mimmick" the behavior of purely ideologically motivated governments in period 1. As a proverbial "wolf in sheep's clothing," closet autocrats can be reelected *despite* the fact that the citizen is opposed to authoritarianism because the citizen is sufficiently uncertain about the ultimate intentions behind governmental policy choices. In these equilibria, it is possible for a closet autocrat to be returned to power – and then to exploit the room created for authoritarian moves in the second period. Authoritarian backsliding occurs.

The central issue of interest are the conditions that give rise to these different equilibria. What are the political environments that allow for authoritarian backsliding, and what are the environments that prevent it? To develop the intuition underlying the equilibria, consider the position of the citizen first. The citizen observes a first period policy choice by the incumbent, and must then choose whether to reelect the incumbent or to replace him with the opposition. How the citizen evaluates this choice depends on several factors: How does the citizen feel about the opposition relative to the ideologue (x_C)? How concerned is the citizen about preserving the rule of law (β), and what is her ex ante belief that the incumbent might be a closet autocrat (α)? Finally, does she believe that a closet autocrat might be willing to disguise himself in the first period in order to be reelected? Taken together, these factors result in a citizen strategy that is characterized by thresholds that define how the citizen will respond to the incumbent's first period policy. These thresholds – which are summarized in Table 1, and illustrated in Figure 2 – have a simple logic.

Figure 2. Illustration of Citizen Threshold Strategy



Suppose the citizen believes – having observed the first period policy choice – that the incumbent is a closet autocrat. This knowledge potentially puts the citizen in a quandary: On the one hand, she does not want to see an authoritarian move in the second period. On the other hand, the policy choice of the closet autocrat may be ideologically more palatable than the alternative offered by the moderate opposition. The citizen resolves this tension by adopting a strategy under which she will reelect a known closet autocrat *only if* the autocrat chooses a first period policy that (while preferable to the opposition) is so moderate as to significantly constrain future authoritarian moves. Put differently, by constraining autocratic possibilities in the next period, a moderate first period policy acts as insurance against authoritarianism. Therefore, the citizen is only willing to reelect a closet autocrat for whom this constraint is sufficiently strong. This threshold is given by T_{CA}. Note that – as illustrated in Figure 2 – this threshold is *below* the citizen's own ideal point: The citizen is trading ideological proximity for constraints on future authoritarianism. Note also that this threshold depends on β, the degree to which the citizen is concerned about the rule of law. As this concern looms larger (β increases), T_{CA} moves left, implying that the citizen demands more and more significant constraints on the closet autocrat's future behavior to reelect.

In contrast, suppose the citizen is sure that the incumbent is an ideologue, and will not make an authoritarian move in the future. In this case, the citizen's reelection decision simplifies to a straightforward ideological choice: She reelects the incumbent if the incumbent's first period proposal is closer to her ideal point than the opposition, and votes for the opposition otherwise (threshold T_I in Figure 2). Concerns about authoritarianism become irrelevant. As is intuitive, this threshold is above T_{CA}: Because the citizen knows that she is not facing a closet autocrat, she is not concerned about restricting the first period policy in order to constrain authoritarian moves in period 2.

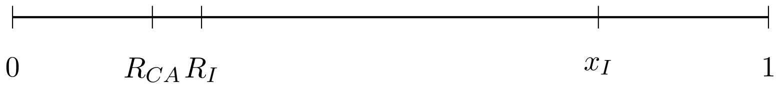
Table 1. Summary of thresholds

Threshold	Imposed by	Explanation
$T_{CA} = \frac{2x_C}{\beta+1}$	Citizen	Citizen will reelect known closet authoritarian who sets policy below this threshold.
$T_I = 2x_C$	Citizen	Citizen will reelect known ideologue who sets policy below this threshold.
$T_\alpha = \frac{2x_C}{\alpha\beta+1}$	Citizen	If the citizen is uncertain about the incumbent's type, she will reelect an incumbent who sets policy below this threshold.
$R_{CA} = 1 - \sqrt{\frac{2}{3}}$	Closet autocrat	Closet autocrat will only set policies at or above this threshold to get reelected.
$R_I = (1 - \frac{1}{\sqrt{2}})x_I$	Ideologue	Ideologue will only set policies at or above this threshold to get reelected.

Finally, suppose the citizen is uncertain. Given the observed policy choice by the incumbent, she believes that the incumbent could be purely ideologically motivated, and does not pose an authoritarian risk, but she also believes that there is some probability that the incumbent is a closet autocrat who would impose authoritarian policies in the future. Once again, there is s threshold such that the citizen will reelect the incumbent only if the first period proposal is below this threshold (T_α in Figure 2). As is intuitive, this threshold falls between the two others, and depends critically on the citizen's ex ante belief that the incumbent is a closet autocrat (α) and the degree to which she is concerned to maintain the rule of law (β). As she becomes more and more certain that the incumbent is likely to be a closet autocrat (α goes to 1), this threshold converges on T_{CA} and as she becomes more convinced that the incumbent is an ideologue (α goes to 0), this threshold converges on T_I. (Note that this implies that T_α may be to the right or left of x_C.)

Now, consider the situation confronting the two types of incumbents. The key choice each type faces is what policy to adopt in the first period, since this policy choice will determine the citizen's reelection decision. The incumbent has an incentive to be reelected: Being in power in the second period ensures that the ideologue can preserve the initial policy choice instead of

Figure 3. Illustration of Incumbent Threshold Strategy



having it overturned if he is replaced by the opposition. The same is true for the closet autocrat. Additionally, the closet autocrat wants to remain in power in order to pursue his autocratic agenda in the second period. Thus, there is some value in proposing policies that will result in reelection, even if these policies diverge from the incumbent's ideal point. At the same time, there is a limit to the incumbent's willingness to do so. If he gives up on the goal of reelection, he is free to impose his ideal policy in period 1, and this has some value. Taken together, these considerations imply that like the citizen, the two types of incumbents adopt a strategy that is characterized by a threshold. Specifically, for each incumbent, there is a threshold (below the incumbent's ideal point) such that the incumbent is not willing to make a proposal below this threshold in order to be reelected. These thresholds are summarized in Table 1 and Figure 3 provides a graphical illustration. The threshold R_{CA} denotes the constraint imposed by the closet autocrat: The autocrat is not willing to adopt policies to the left of this threshold in order to be reelected. R_I denotes the threshold imposed by the ideologue, who will not adopt policies to the left of this threshold to be reelected. While R_{CA} is fixed (since the relative distance of the closet autocrat and the opposition are fixed), R_I depends on the ideological preferences of the ideologue: As the ideologue's preferences become more extreme, R_I shifts to the right, while it approaches 0 as the ideologue becomes more moderate.

The model's equilibria derive from the interaction (that is, relative location) of these thresholds. We summarize these results in a series of propositions.

Proposition 1. *In a Downsian environment in which the ideologue and opposition have converged sufficiently ($x_I < T_{CA}$), the two types of incumbents separate. Each proposes its ideal point, and only the ideologue is reelected. No authoritarian backsliding occurs.*

We refer to this environment as "Downsian" in the sense that the opposition (with an ideal point at 0) and the ideological incumbent (with an ideal point of x_I) are relatively close together –

specifically, the ideologue's ideal point is so close to the opposition that the ideologue can pursue his preferred policy in period 1 and be reelected, and this policy is so moderate that the closet autocrat is not willing to emulate the ideologue in order to be reelected. In other words, this is a political environment in which the mainstream parties are both sufficiently centrist (in the sense of being close to the pivotal voter) that the closet autocrat is politically isolated, and not viable. Autocrats are unable to "infiltrate" the political system, and as a result, no authoritarian backsliding occurs.

Proposition 2. *In a polarized political environment in which the preferences of the ideological incumbent and the citizen are sufficiently far to the right of the opposition (specifically, $T_\alpha > R_{CA}$), the two types of incumbents adopt the same policy, and are reelected by the citizen. Authoritarian backsliding in period 2 is possible.*[6]

The intuition behind this proposition is straightforward: As the political environment becomes more polarized, i.e., as the citizen and the ideological incumbent move to the right, two dynamics begin to play out. First, replacing the incumbent with the opposition becomes less palatable for the citizen. Even if the citizen believes that there is some possibility that the incumbent is an autocrat, she is more willing to reelect the incumbent in order to obtain policies that are ideologically more proximate (that is, as x_C increases, T_α shifts to the right). This logic is closely related to the logic of Svolik's (2018) model. The second dynamic follows from this, and is concerned with the distinct element of our model: Citizen uncertainty about the type of incumbent. Because the ideological incumbent is now freer to adopt policies that are farther to the right, the cost to the closet authoritarian of playing the wolf in sheep's clothing has decreased sufficiently that he is willing to masquerade as an ideologue by making the same proposal in order to be reelected ($T_\alpha > R_{CA}$). As a result, the citizen continues to be unsure about the nature of the incumbent but – on balance – chooses to reelect. If the incumbent turns out to be the closet autocrat, he will exploit the room created in the first period to pursue authoritarian policies in period 2.

[6] Note that if elite polarization becomes extreme, i.e., the ideologue's ideal point moves sufficiently far to the right, it is possible to return to a separating equilibrium in which neither type of incumbent is reelected. See the appendix for details.

Figure 4. Equilibrium illustration

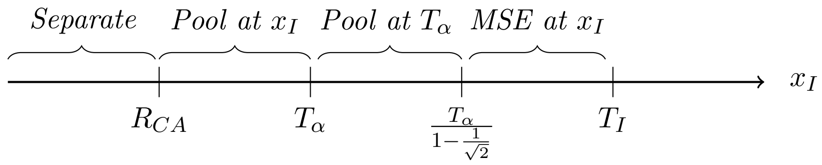


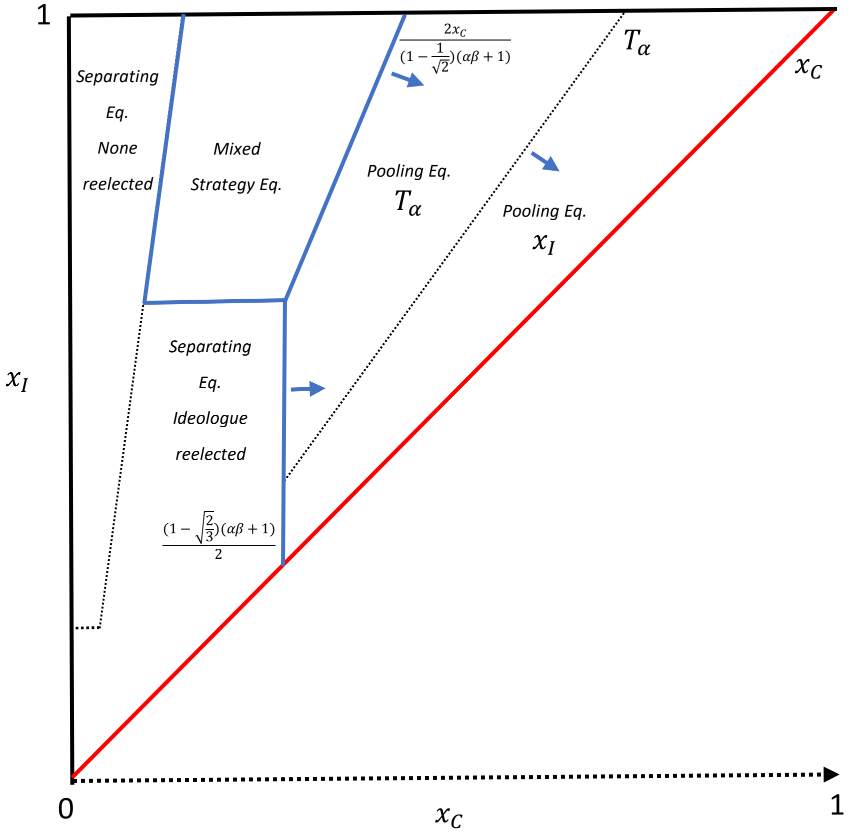
Figure 4 provides a graphical illustration. Along the x-axis, we plot the preferences of the ideological government. As the figure illustrates, as polarization increases, and the ideologue shifts to the right, we transition from a separating equilibrium in which the citizen does not reelect the autocrat to a pooling equilibrium in which this becomes possible.

Proposition 3. *As the citizen is either (a) less concerned about authoritarian backsliding (β decreases), or (b) believes that it is less likely that she is facing a closet autocrat (α approaches* 0*), the level of polarization required to allow for a pooling equilibrium decreases. Authoritarian backsliding is easer to achieve.*

The central implication of this proposition is that there is an interactive relationship between polarization (x_C and x_I move to the right) and the degree to which the citizen is concerned about the threat of authoritarianism – either because she does not believe that the incumbent is likely to be a closet autocrat (α is low) or because she does not care as deeply about preventing autocratic moves (β approaches 1). As concerns about the potential for authoritarian backsliding decrease, the citizen is increasingly focused on the ideological proximity of the first period proposal to her own ideal point, and as a result, T_α moves to the right. The implication of this is that the citizen is more likely to reelect incumbents even if they make more extreme proposals – and this opens the door for closet autocrats, who are now willing to masquerade as ideologues by adopting proposals that will allow them to be reelected.

The equilibria are summarized in the figure below as a function of elite polarization, x_I and polarization in the electorate, x_C, with blue arrows representing change in the constraint as a α or β changes.

Figure 5. Summary of Equilibria


1
Separating Eq. None reelected
Mixed Strategy Eq.
$\frac{2x_C}{(1-\frac{1}{\sqrt{2}})(\alpha\beta+1)}$
T_α
x_C
Pooling Eq. T_α
Pooling Eq. x_I
x_I
Separating Eq. Ideologue reelected
$\frac{(1-\sqrt{\frac{2}{3}})(\alpha\beta+1)}{2}$
0
x_C
1

Before specifying the empirical implications of our model that are the focus of the application in the next section, it is useful to summarize the central logic of the argument we have laid out. The aim of the model is to contribute to understanding how authoritarian backsliding can occur in contexts in which citizens (at least pivotal citizens) are, at bottom, *opposed* to creeping autocracy. The model is designed to capture two factors that we regard as potentially significant. The first is that there is a dynamic element to authoritarian backsliding: Governments take actions at one time (for example, institutional reforms) that may enable authoritarian moves in the future. The second is that there may be competing interpretations of these actions – that is, citizens face some uncertainty about whether the incumbent government is a closet autocrat intent on laying the groundwork for future authoritarian moves, or an ideologue who may have ideologically more extreme preferences, but who is otherwise committed to the rule of law. The central insight of the model comports well with intuition: Increasing polarization – both of the electorate, and of political elites – is a central driver of the potential for authoritarian backsliding. There are two reasons for this. The first – an insight also developed by Svolik (2018) – is that polarization makes electoral punishment of the incumbent more difficult by making the opposition less attractive to potentially pivotal citizens. The second feature is closely related to this, and derives from the uncertainty that voters face about the incumbent's intentions (an element not addressed in Svolik's model): By making more extreme policy choices electorally viable, polarization opens the door to closet autocrats who may enter the political system because they see an opportunity to begin laying the groundwork for authoritarian moves while retaining power. Being uncertain about the incumbent's ultimate intentions, citizens – given sufficient polarization – may be willing to reelect an incumbent only to discover afterwards (and with regret) that the incumbent turns out to be a closet autocrat. The second insight of the model is that how much polarization is required in order to give rise to these dynamics depends critically on voter beliefs about the nature of the incumbent they are facing, and the extent to which voters are concerned to prevent authoritarianism. If these concerns weigh less heavily, or they are more convinced that the incumbent is an ideologue, even more moderate levels of polarization can set in motion the potential for authoritarian backsliding.

2.3. **Empirical Implications.** The model we have presented is highly simplified, and focuses on a small number of actors and decisions in order to put one mechanism that may contribute to authoritarian backsliding into starker relief. This implies that some stretching is necessary to connect the logic of the model to "real world" situations. Nevertheless, we believe that the model suggests several empirical conjectures that are consistent with the underlying logic of the argument. Here, we focus on conjectures that are concerned with the behavior of individual voters. One way to do so is to conceptualize each voter as engaging in the kind of calculation that the pivotal citizen in our model undertakes. In a separating equilibrium, the voter is unwilling to vote for an incumbent perceived to be an authoritarian, and prefers the opposition. In a pooling equilibrium, the voter is willing to support the (potentially authoritarian) incumbent.

There are four types of equilibria in our model—separating, pooling on reelection, pooling on non-reelection and hybrid equilibria. The equilibrium of paramount interest to us is the pooling equilibrium that obtains when the voter is genuinely uncertain whether the incumbent he is reelecting into office is a closet autocrat or ideological incumbent, as this is the equilibrium that allows for the gradual backsliding into authoritarianism.

The conditions for this pooling equilibrium to obtain can be summarized as the following empirical implications: The reelection of the incumbent by citizens who are uncertain about whether they are about to elect a closet autocrat or an ideological incumbent is more likely, holding all else constant when:

- polarization between the incumbent and the opposition (x_i) is higher;
- the ideological distance between the representative citizen and the opposition is greater and the probability that the incumbent is in fact a closet authoritarian is lower;
- the importance of avoiding authoritarian rule is sufficiently high.

The second condition states that polarization in the electorate alone is insufficient for a pooling equilibrium. The additional assumption that is needed is that beliefs that the voter is facing a closet authoritarian be sufficiently low. The third condition is a general assumption of the model ($\beta > 1$) and ensures that citizens have lexicographic preferences over democracy vis a vis authoritarianism, an important point of departure for us from the model of Milan Svolik (2018).

2.4. **Survey Evidence.** In order to measure parameters defining these conditions we make use of a unique set of public opinion surveys conducted by the Center for Public Opinion Research in Poland (CBOS). These polls have been conducted on representative samples of Poles using the same sampling technique since 2001. 2001 is also when PiS and PO—the two parties vying for power over the last decade—appeared on the political scene. Both parties were created following the crisis of the Action Election Solidarity (AWS), the umbrella party organization uniting former anti-communist dissidents against the successor communist Democratic Left Alliance (SLD). The breakup of AWS and the crisis of SLD have been considered by scholars studying the region the end of the so-called "regime divide" in Poland (Grzymala-Busse 2001). Beginning with 2003, the main cleavage dividing Polish voters was no longer allegiance to the former communist autocrats or their opposition, but a more classical conservative-liberal cleavage (Carroll and Nalepa N.d.).

In January 2001, CBOS started including among its feeling thermometer questions, items about about sympathy towards Kaczynski and Tusk, the leaders of PiS and PO, respectively.[7] Figure 3 shows the mean differentials between sympathy to each of these leaders.To create this measure, we first calculate the difference in sympathy scores for each respondent. Next, we take its absolute value. Finally, for each survey, we record the mean and the standard deviation of this absolute value. [8]

Figure 2.4 can be interpreted as the trend in elite polarization, a variable that we call *PiSmPO*. The trend starts with the 92^{nd} survey conducted by CBOS (around 2001) and ends in 2011. The figure shows that around 2005 (or the 140^{th} survey), the year in which PiS won a plurality in the parliamentary elections for the first time (and entered into a cabinet coalition with two other parties: Samoobrona and the League for Polish Families) there is a dramatic increase in elite polarization. Both the differential in sympathy scores and the standard deviation of our elite polarization measure increase. This result suggests that one of the conditions for the pooling equilibrium is satisfied in the Polish case. However, this evidence is also consistent with the separating equilibrium in our model and it is consistent with other models of authoritarian backsliding, such as that of Svolik (2018). Thus to provide stronger support for our theory, we need to look beyond increases in elite polarization over time and establish conditions for obtaining the equilibrium in which voters reelect incumbents because they are *uncertain* if they are dealing with an ideologue or with a closet autocrat. For this further evidence, we turn to a contemporary CBOS survey from 2017.

2.5. **The August 2017 survey.** In the subsection below, we use a contemporary survey from CBOS in order to illustrate the empirical implications of our model. The survey was conducted directly following Presidential veto of the bills proposing the politicization of the National Council

[7] Specifically, the question asked was “Persons in public life—in their behavior, in what they say, and what they try to achieve—can arouse more or less trust. We will show you now a list of persons active in public life and ask you to what extent this person is trustworthy. -5 means that you associate great distrust towards this person, 0 means you are indifferent and 5 means that you trust them completely. And of course, you can make use of any numbers between -5 and 5 to express your trust towards this person. Please let us know if you do not know this person.”

[8] Note that the absolutization of the differential is necessary, because if one respondent expressed their highest sympathy to Kaczynski and least towards Tusk, while another respondent expressed highest sympathy to Tusk and least to Kaczynski, the mean of their ‘differentials would be zero. Yet such preferences are consistent with the highest elite polarization admitted by the survey questions.

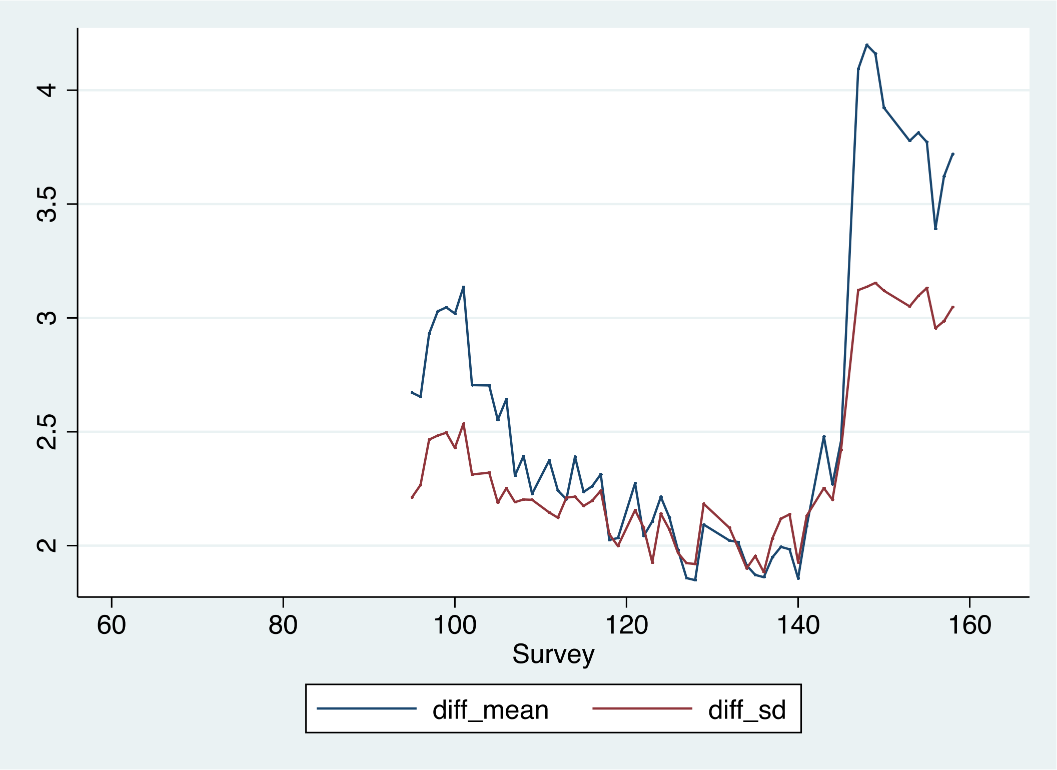


FIGURE 6. Mean and Standard Deviation of Difference in trust towards Lech Kaczynski and Donald Tusk, leaders of the two largest parties in Poland 2001-2011

of the Judiciary and the Supreme Court. We elaborate on the nature of these judicial bodies in the final section of our paper. Here, it suffices to note that the Supreme Court is the highest court of appeal and for all practical purposes overlaps in its tasks with the Supreme Court of the US, except for verifying the constitutionality of legislation. [9] The National Council of the Judiciary, on the other hand, is the sole advisory body nominating candidates for judges and initiating disciplinary action against judges. In short, the survey was conducted within days of intense and nationwide protests that prompted the President's decision to veto two bills that were to bring the judiciary under complete political control.

It is worth pointing out that the Elite Polarization (measured with the same PiSmPO variable as in the historic surveys in the Figure 2.4 is at 4.707 with a standard deviation of 3.21. Thus, elite polarization currently surpasses the highest value of *PiSmPO* in the 2001-2011 period.

Although observational data from surveys does not come close to the experimental data used by Ahlquist et al. (2018), a contemporary survey conducted on representative samples of Poles allows us to corroborate the conditions of the pooling equilibrium by exploiting the variation in the parameters across individual respondents. Building on what we established in the section on empirical implications earlier, we propose to do operationalize the parameters of interest as follows:

(1) In order to measure the elite polarization between (x_I in the model) we use the absolute value of the trust differential between PiS leader Kaczynski and current PO leader, Schetyna: $|TrustPiS_i - TrustPO_i|$ at the level of each respondent;

(2) In order to measure the polarization between the citizen and the opposition candidate (x_C in the model, we will use sympathy towards Grzegorz Schetyna, the PO leader: *Polarization in the Electorate*;

(3) In order to measure the citizen's beliefs as to whether he is facing a closet autocrat or an ideological candidate (α in the model), we use respondents' expressed attitudes to the

[9] The empirical appendix provides an organization chart of the Polish court system for reference

protests in defense of rule of law in Poland and against PiS rule that led the President to veto the two controversial bills, *Support Anti-PiS Protests*;[10]

(4) In order to measure how much the citizen cares for the authoritarian dimension (parameter β in the model), we use use answers to three questions gaging respondents' sensitivity to authoritarianism.

The four questions that measure the respondent's sensitivity to the authoritarian dimension asked the respondent to what extent he or she agrees with the following four separate statements.

- Democracy is superior to any other form of rule ($Authoritarian1$)
- For people like me, it does not matter whether the regime is authoritarian or democratic ($Indifferent$)
- Sometimes Non-democratic rule is better than democratic rule ($Authoritarian2$)
- Government by a strong leader is decidedly better than democratic rule ($StrongLeader$)

Respondents could "agree strongly", "agree somewhat", "rather disagree", "strongly disagree" with the above statements. Higher values of these variables represent stronger disagreement.

It is important to verify that that what these questions are picking up is not simply sympathy to the opposition. The table below shows the correlation matrix of all four measures with the variable *Polarization in the Electorate* included. What is clear from the correlation matrix is that the association is low. Thus sensistivity to authoritarian rule—our operationalization of β with

Table 2. Correlation matrix

	Polarization in Electorate	Indifferent	Authoritarian 2	Strong Leader	Authoritarian
Polarization in Electorate	1.00				
Indifferent	0.02	1.00			
	(0.50)				
Authoritarian2	0.11	0.42	1.00		
	(0.00)	(0.00)			
Strong Leader	0.19	0.38	0.48	1.00	
	(0.00)	(0.00)	(0.00)		
Authoritarian	-0.03	-0.18	-0.27	-0.17	1.00
	(0.35)	(0.00)	(0.00)	(0.00)	

Note: P-values in parentheses

[10] During these protests, the argument that was made is that the KRS and supreme court reforms violated the constitution. Thus, the variable "support of protests" can serve as an indicator of belief that the incumbent is a closet authoritarian (α in the model).

the four survey questions is tapping into something different than sympathy to the opposition candidate.

Our dependent variable is the dummy *PiSvoter* coding as 1 a voter who in response to "Were the parliamentary elections to take place this Sunday, which party would you vote for?" indicated "PiS." We used the question "Would you vote in the parliamentary elections were they to take place this Sunday" to filter out the non-voters. Thus the regressions are run only on voters.

All variables have rescaled so as to match the parameters of the model for ease of interpretation. In out regressions, we will use the interaction term between *Support for Anti-PiS Protests* and *Polarization in the Electorate* to reflect the fact that the pooling equilibrium requires both sufficiently high polarization in the electorate and sufficiently low beliefs that the voter is dealing with the closet autocrat.

In addition to the variables operationalizing the parameters determining the likelihood of the pooling equilibrium, our survey includes basic demographic information on all respondents, such as age, gender and education. We also include in our regressions information about employment, whether the respondent lives in a large city of village (with town being the default category) and whether the respondent is religious. Employment is measured with an ordinal scale between 0 and 1 with 0 and 1 representing unemployment and full time employment, respectively. Religiousness is measured with a dummy variable which take on the value of 1 when the respondent goes to church at lease once a week or more on average. For Poland, a highly Catholic country made up of systematic churchgoers, setting the dummy at this threshold captures most variation.

We note, that our expectations are that, after controlling for these demographic variables, the effect of Elite Polarization and Polarization in the Electorate will be positive and the effect of beliefs that the voter is dealing with a closet autocrat will be negative, though more negative for higher values of Polarization in the Electorate.

Because of the dichotomous nature of our dependent variable, we chose a nonlinear probability model and specifically, a logit in the following format:

$$Pr(PiSvoter_i = 1) = \frac{1}{1 + exp(-\boldsymbol{x_i\beta})} \tag{1}$$

, where $x_i\beta = \beta_1 ElectPolarization + \beta_2 ElitePolarization_i + +\beta_3 Anti - PiS_{P}rotest_{S}upp$

$\beta_4 ElectPolarization * Anti - PiS + \beta_5 SuppIndifferent_i$

$\beta_6 Male_i + \beta_7 Employed_i + \beta_8 Education_i + \beta_9 Religious_i + \beta_1 0village_i + \beta_{11} city_i$

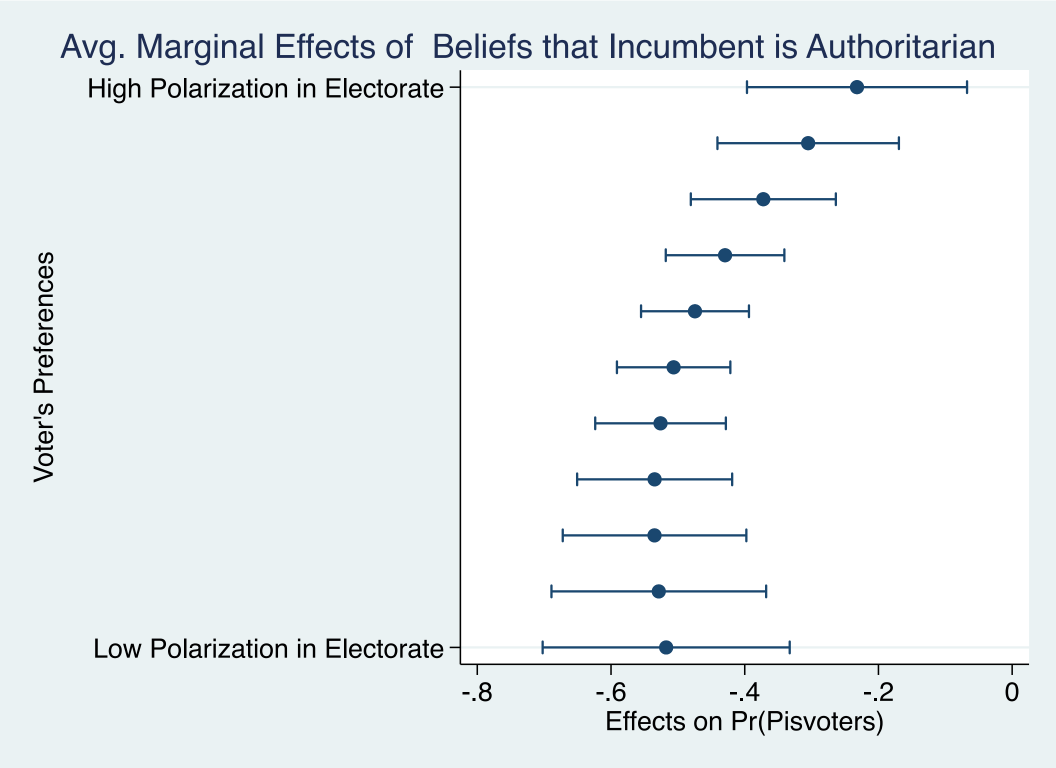
The results of four models ran using different operationalization of anti-authoritarian attitudes are presented in Table 1 below.

	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Polarization in Electorate	1.276*	0.848	1.366*	1.015
	(1.97)	(1.25)	(2.07)	(1.61)
Support Anti PiS Protest	-4.734***	-5.719***	-4.275***	-4.826***
	(-6.22)	(-6.76)	(-5.41)	(-6.40)
Support Anti-PiS*Elect. Polarization	3.382**	4.570***	2.937**	3.582***
	(3.19)	(3.95)	(2.66)	(3.42)
Elite Polarization	2.344***	2.933***	2.351***	2.549***
	(4.85)	(5.71)	(4.75)	(5.44)
Authoritarian 1	0.242			
	(1.46)			
Authoritarian 2		0.264		
		(1.71)		
Strong Leader			-0.290*	
			(-1.98)	
Indifferent				-0.0629
				(-0.43)
Male	-0.292	-0.251	-0.366	-0.149
	(-1.16)	(-0.94)	(-1.39)	(-0.62)
Employed	-0.295	-0.366	-0.0517	-0.238
	(-0.99)	(-1.16)	(-0.17)	(-0.83)
Education	-0.128**	-0.118*	-0.105*	-0.106*
	(-2.69)	(-2.32)	(-2.15)	(-2.26)
Village	0.281	0.311	0.313	0.295
	(0.88)	(0.94)	(0.97)	(0.96)
City	-0.281	-0.231	-0.500	-0.336
	(-0.86)	(-0.68)	(-1.50)	(-1.06)
Age	0.0008	-0.0015	0.0077	-0.00011
	(0.09)	(-0.17)	(0.90)	(-0.01)
Religiousity	1.425*	0.819	1.067	1.179
	(2.19)	(1.31)	(1.72)	(1.95)
Constant	-1.898	-1.277	-1.177	-1.189
	(-1.92)	(-1.31)	(-1.24)	(-1.27)
N	577	545	560	603

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Without looking at marginal effects, since this is a logit, we can only interpret the sign and significant of the coefficients, but not their substantive meaning. The effects of Polarization in the electorate and Elite polarization are significant and in the expected direction, increasing the probability of voting for PiS. The individual effect of beliefs is negative, but recall that we are interested in uncovering the combined effect beliefs and polarization in the electorate. Since one

Figure 7. Average Marginal Effects of Belief that Incumbent is Closet Autocrat for different levels of polarization in the Electorate (Based on Model 4)



constituent term is negative, while the other is positive, the best way to see these effects is by graphing the average effect of beliefs for different levels of polarization.

The next figure, 2.5, graphs the marginal effect of citizen beliefs that they are dealing with a closet autocrat for different levels of polarization ii the electorate. Figure 2.5 clearly shows that the interaction is negative. An increase in the belief that the voter is facing a closet autocrat decreases the probability of reelection, for all levels of polarization in the electorate, but particularly for low levels of polarization. Concretely, when polarization is low, the average effect of the belief that the incumbent is closet authoritarian is to decrease the probability of reelection by about 52%, whereas when polarization is high, the effect is only about 20%.

Looking back to Table 3, we notice that the effect of anti-authoritarian attitudes is only significant in the case of one of the measures (*Strong Leader*). None of the other variables used to gage sensitivity to the authoritarian dimension (with the variables $NonDemocracySuperior$, $StrongLeader$, and $Indifferent$ and $DemocracySuperior$) turned out significant. However, the expectation for the effect of sensitivity is harder to interpret from the theoretical model, as the assumption there is that $\beta > 1$ or that voters have lexicographic preferences of democracy. Therefore, in order to illustrate the effect of β in the context of our survey, we proceed as follows. We split the sample into respondents who answered the sensitivity questions positively (yes and rather yes, $\beta < 1$) and those who answered the sensitivity questions negatively (no and rather no, $\beta \geq 1$). The results of these split sample tests are presented in Table 4 below.

	(Model 5)	(Model 6)	(Model 7)	(Model 8)	(Model 9)	(Model 10)
	$\beta \geq 1$	$\beta < 1$	$\beta \geq 1$	$\beta < 1$	$\beta \geq 1$	$\beta < 1$
Polarization in Electorate	0.900	0.977	2.925**	-1.181	0.0729	3.754**
	(0.94)	(0.83)	(3.27)	(-0.97)	(0.09)	(2.68)
Anti-PiS Protest Supporter	-5.870***	-5.548***	-3.155**	-5.275***	-5.969***	-1.675
	(-5.53)	(-3.44)	(-3.22)	(-3.38)	(-6.24)	(-1.24)
Anti-PiS Protest$*$Elect. Pol.	4.476**	4.869*	0.789	5.945**	5.253***	-0.881
	(3.00)	(2.27)	(0.56)	(2.63)	(3.98)	(-0.42)
Elite Polarization	2.439***	3.556***	1.781**	4.051***	2.991***	1.047
	(3.62)	(3.85)	(2.91)	(3.71)	(5.29)	(1.03)
Authoritarian 2	-0.245	0.464				
	(-0.70)	(0.73)				
Strong Leader			-0.805*	0.182		
			(-2.37)	(0.35)		
Indifferent					-0.610*	0.451
					(-2.06)	(0.67)
Male	0.196	-0.748	-0.326	0.0992	-0.250	0.120
	(0.59)	(-1.49)	(-1.00)	(0.19)	(-0.87)	(0.23)
Employed	-0.856*	0.403	-0.129	-0.0478	-0.404	0.598
	(-2.08)	(0.69)	(-0.33)	(-0.09)	(-1.20)	(0.93)
Education	-0.127*	-0.127	-0.146*	-0.0249	-0.0919	-0.118
	(-1.97)	(-1.31)	(-2.31)	(-0.29)	(-1.68)	(-1.16)
Village	0.621	0.192	0.455	-0.0225	0.246	0.862
	(1.49)	(0.30)	(1.12)	(-0.04)	(0.68)	(1.19)
City	0.320	-0.755	-0.348	-1.008	-0.346	-0.334
	(0.74)	(-1.19)	(-0.83)	(-1.63)	(-0.93)	(-0.48)
Age	-0.0215	0.0302	-0.00272	0.0196	-0.0109	0.0458*
	(-1.85)	(1.85)	(-0.25)	(1.27)	(-1.13)	(2.20)
Religiousity	1.039	1.582	1.292	0.595	1.307	-0.298
	(1.32)	(1.06)	(1.59)	(0.50)	(1.88)	(-0.20)
Constant	0.889	-4.172*	-0.0848	-1.522	1.201	-4.786*
	(0.61)	(-2.09)	(-0.06)	(-0.92)	(0.94)	(-2.11)
N	370	175	410	150	474	129

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

What we see if that, in line with our expectations, the effects of anti-authoritarian sensitivity (*Antiauthoritarian2, Strong Leader*) and *Indifferent* are significant and in the expected direction in two out of three of the $\beta > 1$ models (*Strong Leader* and *Indifferent*), but insignificant entirely and/or in the wrong direction in all $\beta \leq 1$ models. Furthermore, the substantive effects of these variables are higher than in the pooled sample. In addition, it is reassuring that the remaining effects of our parameters remain stable in the $\beta > 1$ models. This offers further support for our model's application to to authoritarian backsliding in Poland. Yet analyzing cross-sectional data at one point is a important limitation of this empirical section. Because experimental or panel data allowing us to measure the parameters of our model is not available, we supplement are Empirical Implications section with a qualitative analytic narratives from Hungary—briefly and—more extensively—from Poland.

2.6. **Narrative Evidence from Hungary and Poland.** Both Hungary and Poland—the two countries we use to motivate this paper are at best "sliding" autocracies. That is, neither pundits, scholars, nor politicians themselves are willing to call them actual autocrats.[11] For this reason, we argue that both Poland and Hungary match the conditions of the pooling equilibrium in which citizens reelect the incumbent *not knowing* whether or not he is a closet authoritarian. Analytical narratives from Hungary and then Poland will show that these two countries are consistent with the incumbent and the closet autocrat pooling policy proposals and getting reelected into office.

2.7. **Hungary.** Victor Orban was elected into office in 2010 with a majority allowing him to change the constitution. His party, Fidesz, had originated in the youth section of the Free Democrats' Party (SzDSz), the most prominent dissident organization that had negotiated the terms of democratic transition with the communists in 1989. When members of Fidesz became too old to call themselves a youth organization, they created a new party changing the spelling of the organization from an acronym FiDeSz (which stood for "Youth Organization of SzDSz) to Fidesz, which in Hungarian means "loyalty." At the same time, the new party also began experimenting with conservative values. The final push towards abandoning liberal ideology was the competition for power between Fidesz and JOBBIK, a radical anti-semitic and anti-Roma organization that was rapidly gaining traction in the polls, particularly in peripheries of the country. To capture JOBBIK's electorate, Fidesz moved even further to the right.

Upon winning the elections, Orban blamed liberal policies and failing to hold the communists accountable for the rise of chauvinistic parties like JOBBIK. The weakening of the Constitutional Court was the result of the first pieces of legislation that came out of the Fidesz-controled legislature.

[11] "Even his harshest critics concede that Mr. Orban has gone to nowhere near the lengths of Mr. Putin in silencing his opponents. No one has been tossed in prison for criticizing the government. There has been no overt censorship." (NYT Nov 7, 2014)

In the past, the Hungarian Constitutional Court had been among the most powerful in the Post-Communist region. Among other powers, it had the ability to conduct abstract review, which allows the court to issue decisions on the constitutionality of the law while bills were still in the legislative process. The court could also review any bill following its passage provided it had an impact on the country's budget. Through these two channels, the Court had repeatedly struck down any but the mildest transitional justice laws.

Upon winning a legislative majority, Fidesz's newly proposed legislation first increased the number of judges on the bench so that Orban could staff it with his own supporters. Successive pieces of legislation did away with the prerogatives of abstract review and budget impact review. Eventually, Fidesz severely restricted the rights of ordinary citizens to initiate the process of constitutional review. But even after passing ten amendments to the constitution, Fidesz was still not satisfied with the amount of checks on its power that remained in place. Shortly, Orban started work on drafting an entirely new constitution. The pretext for changing the constitution was that the current Hungarian basic law had been negotiated during the Roundtable Talks with the communist government in 1989 and thus was agreed to under duress, as the communist regime was still in power. Given that these reforms could be interpreted as "doing transitional justice right", many Fidesz supporters believed that Orban has remained "unchanged from the anti-communist rabble-rouser of the past and that charges of incipient dictatorship are left-wing fantasies" According to Zoltan Kovacs, the prime minister's international spokesman "He is the same guy he used to be 25 years ago [...] He wants to get rid of the attitudes, the remnants of the former system and get rid of the attitude that people live on social aid rather than work." (NYT Nov 224, 2014.)

In April 2018, Orban's FiDeSz yet again emerged victorious in the parliamentary elections. This time, with only a plurality of the vote, the party cleared the two thirds supermajority necessary to amend the constitution to bring the judiciary under the control of the executive. The effects are already visible. One month following the election, a flurry of judges from the National Judiciary Council started resigning just days before announcing a verdict against one of Orban's oldest cronies (NYT, After Orban's Victory, Hungary's Judges Start to Tumble, May 1, 2018). The intimidation of a large number of judges that are part of the body that is supposed to ensure judicial independence in Hungary is an ominous sign for the future of rule of law in Hungary and one that is consistent with a second period move by a closet autocrat who is reelected after the second period of our model and proceeds to implement the as far reaching policies as he can on the authoritarian dimension.

2.8. **Poland.** In the October 2015 election, after sitting in the opposition benches for 8 years, PiS emerged victorious in the parliamentary elections. Even though it won only a minority of the vote, it took an absolute majority of seats in the legislature. The victory allowed PiS to establish Poland's first single-party majority cabinet since 1989.

It began its rule with undertaking several changes that could be perceived as heading away from democratic norms. A prominent example was a national security bill permitting wiretapping and granting government access to phone records and electronic data. The legislation was promptly sent to the Constitutional Tribunal (Poland's constitutional court) for an evaluation, but the authors of the bill defended it stating that its aim was to preempt criminal activity.

Two other bills—including one restricting public gatherings—were upon passages sent to the Constitutional Tribunal for review. But instead of waiting for what the Court decides, PiS started tinkering with the Constitutional Tribunal and courts at other levels as well.

In order to better understand the succession of judicial reforms that followed, a brief primer on the Polish court system is helpful. The system of courts is complex and made up of four levels: the regional, district, appellate, and the Supreme Court.[12].

Lower level courts are surprisingly influential due to peculiarity of the Polish constitution: Article 178 of the Basic Law allows lower-level courts to engage in interpreting the constitution when the Constitutional Tribunal is incapable of doing so. Consequently, Article 178 puts plans of any Polish closet authoritarian in jeopardy. Even after replacing key constitutional justices of the Constitutional Tribunal, a ruling party could only pass unconstitutional legislation and expect it to remain on the books if it were in a position to replace judges of every single court entitled to rule on the basis of article 178. Hard to do as this seems, there is a way around not being able to change the

[12] A graphic illustration of the court system is provided in the Empirical Appendix

judges themselves. One can influence their incentives once the Supreme Court—the court of appeal for lower-level decisions—has been brought under the control of the autocrat. Were that the case, any judge interpreting the constitution at odds with PiS would risk having the decision reversed. Since frequent reversals undermine judicial careers, few lower-court judges will likely choose this path.

The Supreme Court is not only the court of appeal for lower-level courts. The constitution makes it also responsible for determining the validity of all nationwide referenda, as well as elections to the Sejm and Senate. Hence, it is within the powers of the Supreme Court to rule an election invalid should an incumbent lose. Another Supreme Court task is to review reports from parties seeking reimbursement for electoral expenditures. Poland, in contrast to the United States, has a public system of financing electoral campaigns. Upon clearing the threshold of 3% of the national vote, parties seek reimbursements for campaign expenditures up to 4.04 PLN per vote from the state treasury. Yet to be reimbursed, the applicant's books must be deemed "in order" by Supreme Court judges. Denying compensation to opposition parties by an autocrat-controlled Supreme Court could thus be interpreted as a "slow-bleed" strategy to bankrupt the opposition and eliminate electoral competition.

The lateral institution to the Supreme court is the Constitutional Tribunal, the equivalent of the Supreme Court in the US and constitutional courts in other countries, except that it only deals with the constitutionality of legislation passed in the Sejm and the constitutionality of legal norms applied in decisions of lower level courts.

Finally, there is a National Council of the Judiciary. This is an independent body which makes recommendations of who should be appointed as a judge and also initiates disciplinary action against members of the judiciary. The final disciplinary decisions are then carried out by a special Ombudsman for Discipline.

A critical set of judicial reforms initiated in the beginning of 2016, eliminated the middle layers of the court system, the regional and appellate courts, forcing judges over 65 to retire, unless they received an exemption from the Minister of Justice. Permission would only be granted following vetting by a special commission. The reforms also made it easier for the government to impose disciplinary actions on all judges, by-passing the National Council of the. This has been accomplished giving parliament the authority to appoint members of the National Council of the Judiciary.

In July, 2017 the PiS controlled parliament approved legislation to drastically change the composition and functioning of the Polish Supreme Court. The Justice Minister was handed discretion over who among the judges of the Supreme Court remained in office and who was forced into retirement. Effectively, this reduced the number of judges from 87 to 31.[13] The prerequisites for holding a Supreme Court seat were lowered to a minimum of 12 years of experience in a regional court. Since the Minister of Justice already simultaneously holds the position of Prosecutor General, the reforms have allowed the ruling majority to choose both the prosecutor *and* the judge in every single court case. The bill was not implemented immediately as the speed with which PiS guided it through the legislative agenda invited public outcry. After tens of thousands of Poles protested the rapid and radical reform in over 100 cities, Andrzej Duda—PiS's President—vetoed the bill, ostensibly to protest the fact that he had not been consulted at the time of its preparation and because it transferred too much power into the hands of the Minister of Justice. Nevertheless, following some compromises which distributed control over the selection of judges between the Ministry of Justice and the Presidency, Duda conceded and the final set of judiciary reforms was passed in December 2017.

Experts concluded that this most recent bill alone potentially conflicts with at least two articles of the constitution (181 and 182), but the measure was not struck down because PiS sympathizers already occupied a majority of the Constitutional Tribunal bench.[14]. Hence, so far, nothing that PiS has done is unequivocally against the constitution. Yet at the same time, there are at least a few reasons why a closet autocrat would want to take control over Poland's Supreme Court, the National Council of the judiciary, and lower level courts in this way.

[13] technically, the number was 43, but 12 would sit in a newly created "disciplinary department", so there would be 31 judges doing the work of 87

[14] A section in the Empirical Appendix describes this first move by Kaczynski which took place in late fall of 2015

First, the Supreme Court could invalidate elections lost by PiS. Second, by using the "slow-bleed" strategy described above, it could deny reimbursements to opposition parties. Third, the most influential opposition party to date—Civic Platform– could take a blow if Donald Tusk, its former leader, and current Council of Europe president, were put on trial before the State Tribunal. The State Tribunal is a special judiciary body for assessing the constitutional liability of persons holding the highest state rank. This process could result in criminal punishment and a loss of civil rights. The chief justice of the Supreme Court serves, *ex oficio*, as the justice presiding over the state tribunal. Tusk's alleged crime is the murdering Jaroslaw Kaczynski's twin brother Lech, who was Poland's president at the time he perished in a plane crash over Smolensk, Russia. According to Jaroslaw Kaczynski, Tusk sabotaged the investigation into the catastrophe and allowed for declaring it an accident much sooner than it was warranted to do so. Finally, gaining steering control over the Supreme Court would allow Kaczynski to pardon a close ally and associate, Mariusz Kaminski. In a 2015 case, a regional court sentenced of Kaczynski and former head of the Central Anticorruption Bureau, to three years in prison for abuses of power. In November of 2015, within days of assuming office, President Andrzej Duda pardoned him. But the Supreme Court annulled the pardon in March 2017. Barring a reversal, Kaminski would go to prison.

While pundits are speculating about the true intentions behind judiciary reforms, the EU is evaluating the status of rule of law in Poland and deciding whether to invoke procedure number seven, which would strip Poland of voting rights in the EU. It is unlikely, however that much will happen before PiS stands for reelection in 2019. For one Orban has vowed to protect Poland from EU sanctions. Coming from a party that is part of an EU party holding a plurality in the European Parliament. PiS at this time is also unlikely to take any unequivocally authoritarian moves because if it wins a $\frac{2}{3}$ majority of seats in the upcoming election, it will be able to change the constitution and accomplish the same goals without drawing international attention. Given this, the moment of authoritarian regret, posited by our model and already illustrated in the Hungarian case is still to come in Poland, *if* Kaczynski is indeed the closet autocrat many believe him to be.

3. Conclusion

In this paper, we address the increasingly popular phenomenon of authoritarian backsliding in recently consolidated democracies. While some scholars attribute it to polarization of the electorate Svolik (2018), others focus on the polarization of elites stressing simultaneously the role of majoritarian political institutions that have the ability of shutting out opposition to the government. In this paper, we try to reconcile these two viewpoints. We also aspire to account for the possibility that voters support closet autocrats unknowingly as such closet autocrats reveal their authoritarian intentions gradually. By the time voters discover that whom they have elected into office is in fact a dictator, it is often too late. For an illustration, one need look no further than the Hungarian election of April 2018, followed by protests of tens of thousands in Budapest [15]. This most recent Hungarian election stands out in particular because for the first time, the incumbent has secured a $\frac{2}{3}$ majority in the legislature (allowing to amend the constitution) with only a plurality of the vote. In order to account for the gradual descent into authoritarianism observed in Hungary, but also Poland, Turkey, and Armenia, we propose a two-stage signaling model, in which the incumbent can be one of two types. The first type has no authoritarian tendencies, but is simply ideologically to the right. The second type is a closet autocrat. The incumbent picks a policy on an ideological dimension and then stands for reelection. Only upon being reelected does he get a chance to implement policy on the authoritarian-democratic dimension. This means that when the electorate is deciding whether or not to reelect the incumbent, there is uncertainty about his type.

What our model allows us to account for, is the behavior of voters in cases where it is hard to resolve whether FiDeSZ and PiS are parties of ideological conservatives or a party of closet autocrats.

We solve this signaling game for equilibria and identify a unique pooling equilibrium where both the ideological candidate and closet autocrat pool their first period policy choice and propose the same policy. This action makes it impossible for the voter to discriminate between the type of autocrat he or she is facing. After describing the conditions necessary for this equilibrium to hold,

[15] The Guardian “Thousands rally against Viktor Orban’s election victory in Budapest”, Sunday, April 2018

we use an analytical narrative from Poland as an illustration. The analytic narrative is developed in two parts. The first, quantitative part, uses historical survey data to establish that elite polarization has dramatically increased since 2001 and then uses a contemporary survey from Poland that allows us to operationalize other parameters of the model. This survey exploits individual level variation in respondent's preferences and beliefs to illustrate the equilibrium predictions. The second part offers analytic narratives from Poland and Hungary, discussing the last couple of years of PiS's policy choices and responses to it to illustrate how uncertainty about these governments' true motivations could lead voters to reelect closet autocrats.

Appendix A. Formal Appendix: Preliminaries

A.1. **The citizen's reelection decision.** Consider the reelection decision facing the citizens. Suppose the incumbent has chosen policy p in the first period, and let γ denote the citizen's updated belief that the incumbent is a closet autocrat. The expected utilities of the options confronting a citizen are given by:

$$EU_C(\text{Reelect}) = -(1-\gamma)(p-x_C)^2 - \gamma((p-x_C)^2 + \beta(y)^2) \tag{2}$$

$$EU_C(\text{Replace}) = -(x_C)^2 \tag{3}$$

The citizen (weakly) prefers to reelect the incumbent as long as the following condition holds:

$$p \leq \frac{2x_C}{1+\gamma\beta} \tag{4}$$

Consider what this threshold implies, depending on the beliefs that the citizen holds:

- If the citizen believes that the incumbent is a closet autocrat ($\gamma = 1$), she will only reelect if policy is below the following threshold:

$$T_{CA} = \frac{2x_C}{\beta+1}$$

- If the citizen believes that the incumbent is an ideologue ($\gamma = 0$), she will only reelect if policy is below the following threshold:

$$T_I = 2x_C \tag{5}$$

- If the citizen is uncertain ($\gamma = \alpha$), she will only reelect if policy is below the following threshold:

$$T_\alpha = \frac{2x_C}{\alpha\beta+1} \tag{6}$$

A.1.1. *Ideologue's first period policy choice.* Consider the ideological government's policy choice in the first period. The worst case scenario for the ideologue is that it sets its ideal point x_I in period 1, and is replaced by the moderate government. (Note, this need not happen: It's possible that the

Table 3. Summary of thresholds

Threshold	Imposed by	Explanation
$T_{CA} = \frac{2x_C}{\beta+1}$	Citizen	Citizen will reelect known closet authoritarian who sets policy below this threshold.
$T_I = 2x_C$	Citizen	Citizen will reelect known ideologue who sets policy below this threshold.
$T_\alpha = \frac{2x_C}{\alpha\beta+1}$	Citizen	If the citizen is uncertain about the incumbent's type, she will reelect an incumbent who sets policy below this threshold.
$R_{CA} = 1 - \sqrt{\frac{2}{3}}$	Closet autocrat	Closet autocrat will only set policies at or above this threshold to get reelected.
$R_I = (1 - \frac{1}{\sqrt{2}})x_I$	Ideologue	Ideologue will only set policies at or above this threshold to get reelected.

ideologue gets reelected even if it sets its ideal point.) This minimum payoff is given by $-(x_I)^2$. On the other hand, the ideologue could set a policy p that results in being reelected. The payoff of this proposal is given by $-2(x_I - p)^2$. These two payoffs define a threshold such that the ideologue is not willing to set a first period policy below this cut-off in order to get reelected. This cutoff is given by:

$$R_I = (1 - \frac{1}{\sqrt{2}})x_I \tag{7}$$

A.1.2. *Closet autocrat's first period policy choice.* Consider the closet autocrat's policy choice in the first period. The worst case scenario for the closet autocrat is that it sets policy $p = 1$ in period 1, and is replaced by the moderate government. The payoff of doing so is given by -3. On the other hand, the closet autocrat could set a policy p that results in being reelected. The payoff of this proposal is given by $-1 - 3(1 - p)^2$. These two payoffs define a threshold such that the closet autocrat is not willing to set a first period policy below this cut-off in order to get reelected. This cutoff is given by:

$$R_{CA} = 1 - \sqrt{\frac{2}{3}} \tag{8}$$

Summarizing, we have the following five thresholds, which are listed in Table 1 for reference. Note that the order of the thresholds imposed by the citizen is $T_I > T_\alpha > T_{CA}$.

A.1.3. *Off-equilibrium path beliefs.* A critical issue is that we must designate the off-equilibrium path beliefs of the citizen as a result of observing out of equilibrium proposals. These beliefs are loosely based on the idea of the intuitive criterion, though we haven't formalized this. These beliefs are based on how the citizen should interpret proposals based on what he knows/thinks the incumbent types are willing to propose. Note that the closet autocrat is not willing to propose $p < R_{CA}$ and the ideologue is not willing to propose $p < R_I$. There are several possibilities we must consider:

Case 1: $R_{CA} < R_I$:

(1) For any proposal $p > x_I$, the belief is that the proposal comes from the closet autocrat, i.e., $\gamma = 1$. This is intuitive since the ideologue never has an incentive to make a proposal above his ideal point.

(2) For any proposal $p \in [R_I, x_I]$, the belief is that the proposal could come from either, i.e., $\gamma = \alpha$.

(3) For any proposal $p < R_I$, neither incumbent is expect to make this proposal, but the cost of this mistake is less for the ideologue. So set $\gamma = 0$.

Case 2: $R_{CA} \in [R_I, x_I]$:

(1) For any proposal $p > x_I$, the belief is that the proposal comes from the closet autocrat, i.e., $\gamma = 1$. This is intuitive since the ideologue never has an incentive to make a proposal above his ideal point.

(2) For any proposal $p \in [R_{CA}, x_I]$, the belief is that the proposal could come from either, i.e., $\gamma = \alpha$.

(3) For any proposal $p < R_{CA}$, neither incumbent is expect to make this proposal, but the cost of this mistake is less for the ideologue. So set $\gamma = 0$.

Case 3: $R_{CA} > x_I$:

(1) For any proposal $p > x_I$, the belief is that the proposal comes from the closet autocrat, i.e., $\gamma = 1$. This is intuitive since the ideologue never has an incentive to make a proposal above his ideal point. [Think about this: between x_I and R_{CA}, neither has an incentive to make this proposal. We could assume that $\gamma = \alpha$ in this region. But this raises a question because it suggests that we should always choose that belief when neither has an incentive to make a proposal.]

(2) For any proposal $p < x_I$, neither incumbent is expect to make this proposal, but the cost of this mistake is less for the ideologue. So set $\gamma = 0$.

(One could also make a different assumption, namely that in regions in which neither incumbent has an incentive to make a proposal, the citizen is uncertain and hence $\gamma = \alpha$. The downside is that such beliefs would give rise to non-monotonic beliefs in Case 2. So requiring monotonicity rules out these beliefs.)

A.2. **Equilibrium analysis.** Note that since we assume that $\beta \geq 1$ and $x_I > x_c$, it must be the case that $x_I > T_{CA}$. We can thus distinguish three cases that we need to analyze, depending on the location of x_I relative to the thresholds imposed by the citizen's ideal point.

(1) Case 1: $x_I \in [T_{CA}, T_\alpha]$. This is a case in which the ideologue is closer to the citizen than the moderate replacement, and the ideologue is not very extreme.

(2) Case 2: $x_I \in (T_\alpha, 2x_c]$. This is a case in which the ideologue is still closer to the citizen than the moderate replacement, but the ideologue is becoming more extreme.

(3) Case 3: $x_I > 2x_c$. This is a case in which the ideologue is further from the citizen than the moderate replacement. The ideologue is becoming extreme.

A.2.1. *Case 1: $x_I \in [T_{CA}, T_\alpha]$.* In this case, we must distinguish between two subcases. If $r_{CA} \leq x_I$, the closet authoritarian is willing to adopt the ideologue's ideal point in a pooling equilibrium, and the citizen is willing to reelect both since this ideal point is below T_α. On the other hand, if $r_{CA} > x_I$, then the closet authoritarian is not willing to adopt the ideal point of the ideologue. Note that the second case is one in which the ideologue is relatively close to the moderate replacement, while he is further from the moderate replacement in the first case.

Case 1.1: $x_I \in [T_{CA}, T_\alpha]$ *and* $r_{CA} \leq x_I$ [Moderate citizen and elite polarization]

This arrangement of ideal points can only occur if the ideologue is closer to the citizen than the moderate replacement, but at the same time, the citizen is sufficiently far to the right to move T_α to the right of x_I, given the levels of α and β, i.e., there needs to be some citizen polarization. This becomes easier as α and β go down, i.e., as the citizen is less worried about the rule of law, and as she thinks it is ex ante less likely that she is facing a closet autocrat.

In this case, in equilibrium, both types of incumbents will propose x_I and we have a pooling equilibrium at the ideologue's ideal point. The citizen will reelect the incumbent. The ideologue clearly has no incentive deviate. Neither does the closet autocrat. The only profitable deviation would be to the right. But given the off-equilibrium path beliefs, if the closet autocrat deviates to $p > x_I$, the citizen updates her beliefs to $\gamma = 1$, and will not reelect. Since $r_{CA} \leq x_I$, the closet authoritarian prefers to be reelected – thus we have a pooling equilibrium at x_I.

Importantly, note that given the off-equilibrium path beliefs we have specified, there can be no pooling equilibrium at any proposal other than x_I. The reason is immediate. Suppose that the incumbent types are pooling at $p > x_I$. Given the beliefs, the ideologue could deviate to x_I and continue to be reelected. Clearly that is preferable. Alternatively, suppose they are pooling at $p < x_I$. Again, given the beliefs, the ideologue can deviate to x_I and be reelected.

Moreover, there can be no separating equilibrium in this case. To see this, note that in any separating equilibrium, the ideologue must be proposing x_I. If he weren't, he could deviate to x_I and still be reelected. Moreover, note that given a proposal of x_I by the ideologue, the closet autocrat can either make a different proposal and not be reelected (since we are above T_{CA}, or he can pool at x_I and be reelected – which he prefers, given that $r_{CA} \leq x_I$. Thus, we have a unique pooling equilibrium with the following strategies (and the beliefs specified above):

Pooling Equilibrium 1.1:

- *Ideologue strategy: Set* $p_I = x_I$

- *Closet autocrat strategy: Set* $p_{CA} = x_I$
- *Citizen strategy: Reelect if* $p \leq g_I$*, replace if* $p > g_I$

Case 1.2: $x_I \in [T_{CA}, T_\alpha]$ *and* $r_{CA} > x_I$ [Low citizen and elite polarization]

What distinguishes this case from the previous one is that the closet autocrat is not willing to pool at the ideologue's ideal point. This is the case because the ideologue's ideal point is fairly close to the moderate replacement government. One way of interpreting this situation is that political competition between the mainstream parties has resulted in something close to convergence: the parties are close together, surrounding the pivotal citizen. As a result, the closet autocrat is sufficiently extreme relative to the mainstream parties that he is not willing to mimick the ideologue in order to get reelected. In equilibrium, we will observe separation: The ideologue will propose his ideal point, while the closet autocrat will propose $p = 1$. The citizen will reelect the ideologue, but will replace the closet autocrat.

To see that this must be the unique equilibrium, note that given the off equilibrium path beliefs, the ideologue is free to propose his ideal point and be reelected – it is a dominant strategy. But given that the ideologue's ideal point is below r_{CA}, the closet autocrat is not willing to pool at this proposal in order to be reelected. As a result, he separates, proposes $p = 1$, and is not reelected.

Separating Equilibrium 1.2:

- *Ideologue strategy: Set* $p_I = x_I$
- *Closet autocrat strategy: Set* $p_{CA} = 1$
- *Citizen strategy: Reelect if* $p \leq g_I$*, replace if* $p > g_I$

A.2.2. *Case 2:* $x_I \in (T_\alpha, 2x_C]$. In this case, we must distinguish between situations in which $R_I < R_{CA}$ and situations in which this is not the case. If $R_I \geq R_{CA}$, mixed strategy equilibria become possible. This requires that $X_I > .62$.

Subcase A: $R_I < R_{CA}$**, which requires** $x_I < .62$**.**

There are three subcases we need to distinguish.

Case 2.1: $x_I \in (T_\alpha, 2x_C]$ *and* $R_{CA} < T_\alpha$ [Moderate citizen and elite polarization]

This case requires that x_C is sufficiently to the right, given values of α and β to move T_α to the right of R_{CA}. This is easier to achieve as these two parameters decline. In other words, there has to be sufficient citizen polarization, and this is more likely when citizens are less concerned about the rule of law or believe that they are not dealing with a closet autocrat.

In this case, in equilibrium, both types of incumbents will propose T_α and we have a pooling equilibrium. The citizen will reelect the incumbent. Neither type of government has an incentive to deviate since making a higher proposal results in not being reelected given the off-equilibrium path beliefs.

Importantly, note that given the off-equilibrium path beliefs we have specified, there can be no pooling equilibrium at any proposal other than T_α. The citizen will not reelected for a proposal above T_α. And given the off-equilibrium path beliefs, for any proposal below T_α, the incumbent could deviate towards T_α and continue to be reelected.

Moreover, there can be no separating equilibrium in this case. To see this, note that in any separating equilibrium, the ideologue must be reelected, which immediately implies that he must be making a proposal below R_{CA} to keep the closet autocrat from making the same proposal. But given the off-equilibrium path beliefs, the ideologue could then deviate above R_{CA} and still be reelected. Thus, we have a unique pooling equilibrium with the following strategies (and the beliefs specified above):

Pooling Equilibrium 2.1:

- *Ideologue strategy: Set* $p_I = T_\alpha$
- *Closet autocrat strategy: Set* $p_{CA} = T_\alpha$
- *Citizen strategy: Reelect if* $p \leq T_\alpha$*, replace if* $p > T_\alpha$

Case 2.2: $x_I \in (T_\alpha, 2x_C]$ *and* $R_{CA} \in (T_\alpha, x_I]$ [Low citizen polarization, moderate elite polarization]

Compared to the previous case, for this case, either citizen polarization has decreased (i.e., x_C has moved to the left), or α or β have increased, thus driving T_α to the left (i.e., the citizen is more concerned about the rule of law or believes it is more likely he is facing a closet autocrat).

In this scenario, no pooling equilibrium is possible since for any pooling equilibrium in which the citizen is willing to reelect, the closet autocrat is not willing to make the required proposal. But we can get a separating equilibrium in which the ideologue proposes R_{CA}, is reelected, and the closet autocrat proposes his ideal point. Given the off-equilibrium path beliefs, no other separating equilibrium is possible. There can be no separating proposal for the ideologue above R_{CA} since then the closet autocrat would mimick the ideologue. And there can be no proposal for the ideologue below because given the off-equilibrium path beliefs, the ideologue could deviate to R_{CA} and still be reelected. So we have a unique separating equilibrium:

Separating Equilibrium 2.2:

- *Ideologue strategy: Set* $p_I = R_{CA}$
- *Closet autocrat strategy: Set* $p_{CA} = 1$
- *Citizen strategy: Reelect if* $p \leq R_{CA}$*, replace if* $p > R_{CA}$

Case 2.3: $x_I \in (T_\alpha, 2x_C]$ *and* $R_{CA} > x_I$ [Low citizen and elite polarization]

This is a case in which - compared to the previous one - elite polarization has gone down since x_I must be moving to the left. As a result, there is a unique separating equilibrium in which the ideologue can propose his ideal point, which the closet autocrat is not willing to do:

Separating Equilibrium 2.3:

- *Ideologue strategy: Set* $p_I = x_I$
- *Closet autocrat strategy: Set* $p_{CA} = 1$
- *Citizen strategy: Reelect if* $p \leq x_I$*, replace if* $p > x_I$

Subcase B: $R_I \geq R_{CA}$, which requires $x_I \geq .62$.

Case 2.1: $x_I \in (T_\alpha, 2x_C]$ *and* $R_I \in [T_\alpha, x_I]$ [Strong elite polarization and moderate citizen polarization, high α or β]

In this case, elite polarization must be high and citizen polarization must be considerable since $2x_C > x_I$. Moreover, to move T_α to the left of R_I, it must be the case that α or β are sufficiently high.

In this case, there can be no separating equilibrium. In any separating equilibrium, the closet autocrat would want to mimick the proposal by the ideologue. There can also be no pooling equilibrium, since the citizen is not willing to reelect in a pooling equilibrium for any proposal that the ideologue is willing to make. Thus, we have to look for a mixed-strategy, semi-pooling equilibrium.

Suppose the ideologue always proposes his ideal point x_I, and the closet autocrat mixes between his ideal point and proposing x_I, proposing x_I with probability t. This immediately implies that the citizen's updated beliefs on observing x_I are given by:

$$\gamma = \frac{\alpha t}{(1-\alpha) + \alpha t}$$

The citizen's expected paoffs from reelecting and replacing the incumbent are then given by:

$$EU_C(\text{Reelect}) = -(x_I - x_c)^2 - \gamma\beta x_I^2$$

$$EU_C(\text{Replace}) = -x_c^2$$

This implies that

$$\gamma = \frac{2x_c - x_I}{x_I \beta}$$

Substituting and solving for t implies that in equilibrium, the closet autocrat must mix with probability:

$$t^* = \frac{(2x_c - x_I)(1-\alpha)}{\alpha(x_I(1+\beta) - 2x_c)}$$

This is a proper probability and consistent with the ordering of cutoffs for this scenario.

We can now move to the strategies for the two types of incumbent. Consider the closet autocrat first. The expected payoffs of making the two proposals are given by:

$$EU_{CA}(x_I) = -(1-x_I)^2 - 1 + (1-r)(-2) + r(-2(x_I - 1)^2)$$

$$EU_{CA}(1) = -1 - 2$$

Solving for r implies that

$$r^* = \frac{(1-x_I)^2}{2x_I(2-x_I)}$$

A.2.3. *Case 3: $x_I > 2x_C$.* In this case, elite polarization is becoming strong: The ideologue is further from the citizen than the moderate replacement. Once again, we must distinguish between situations in which $R_I < R_{CA}$ and situations in which this is not the case. If $R_I \geq R_{CA}$, MSE become possible. This requires that $X_I > .62$. We ignore this case for the moment.

CASE 1: $R_I < R_{CA}$, which requires $x_I < .62$.

Case 3.1: $x_I > 2x_C$ and $R_{CA} < T_\alpha$ [Moderate citizen and elite polarization]

This case requires that x_C is sufficiently to the right, given values of α and β to move T_α to the right of R_{CA}. This is easier to achieve as these two parameters decline. In other words, there has to be sufficient citizen polarization, and this is more likely when citizens are less concerned about the rule of law or believe that they are not dealing with a closet autocrat.

In this case, in equilibrium, both types of incumbents will propose T_α and we have a pooling equilibrium. The citizen will reelect the incumbent. Neither type of government has an incentive to deviate since making a higher proposal results in not being reelected given the off-equilibrium path beliefs.

Importantly, note that given the off-equilibrium path beliefs we have specified, there can be no pooling equilibrium at any proposal other than T_α. The citizen will not reelected for a proposal above T_α. And given the off-equilibrium path beliefs, for any proposal below T_α, the incumbent could deviate towards T_α and continue to be reelected.

Moreover, there can be no separating equilibrium in this case. To see this, note that in any separating equilibrium in which he is not proposing his ideal point, the ideologue must be reelected, which immediately implies that he must be making a proposal below R_{CA} to keep the closet autocrat from making the same proposal. But given the off-equilibrium path beliefs, the ideologue could then deviate above R_{CA} and still be reelected. Thus, we have a unique pooling equilibrium with the following strategies (and the beliefs specified above):

Pooling Equilibrium 3.1:

- *Ideologue strategy: Set $p_I = T_\alpha$*
- *Closet autocrat strategy: Set $p_{CA} = T_\alpha$*

- *Citizen strategy: Reelect if $p \leq T_\alpha$, replace if $p > T_\alpha$*

Case 3.2: $x_I > 2x_C$ and $R_{CA} \in [T_\alpha, T_I]$ [Low citizen and moderate elite polarization]

Compared to the previous case, for this case, either citizen polarization has decreased (i.e., x_C has moved to the left), or α or β have increased, thus driving T_α to the left (i.e., the citizen is more concerned about the rule of law or believes it is more likely he is facing a closet autocrat). In this scenario, no pooling equilibrium is possible since for any pooling equilibrium in which the citizen is willing to reelect, the closet autocrat is not willing to make the required proposal. But we can get a separating equilibrium in which the ideologue proposes R_{CA}, is reelected, and the closet autocrat proposes his ideal point. Given the off-equilibrium path beliefs, no other separating equilibrium is possible. There can be no separating proposal for the ideologue above R_{CA} since then the closet autocrat would mimick the ideologue. And there can be no proposal for the ideologue below because given the off-equilibrium path beliefs, the ideologue could deviate to R_{CA} and still be reelected. So we have a unique separating equilibrium:

Separating Equilibrium 3.2:

- *Ideologue strategy: Set $p_I = R_{CA}$*
- *Closet autocrat strategy: Set $p_{CA} = 1$*
- *Citizen strategy: Reelect if $p \leq R_{CA}$, replace if $p > R_{CA}$*

Case 3.3: $x_I > 2x_C$ and $R_{CA} > T_I$ [Low citizen and moderate elite polarization]

This is a case in which - compared to the previous one - elite polarization has gone down since x_I must be moving to the left. We must distinguish between two cases: Those in which R_I is sufficiently far left to allow for separation, and those in which it is not.

Case 3.3a: $x_I > 2x_C$ and $R_{CA} > T_I$ and $R_I \leq T_I$ [Low citizen and moderate elite polarization]

In this case, we can get a separating equilibrium in which the ideologue proposes T_I. He prefers this to not being reelected, and the closet autocrat will not mimick the proposal. There can be no other equilibrium.

Separating Equilibrium 3.3a:

- *Ideologue strategy: Set $p_I = T_I$*

- *Closet autocrat strategy: Set* $p_{CA} = 1$
- *Citizen strategy: Reelect if* $p \leq T_I$*, replace if* $p > T_I$

Case 3.3b: $x_I > 2x_C$ *and* $R_{CA} > T_I$ *and* $R_I > T_I$ [Lower citizen and/or higher elite polarization compared to 3.3a]

Compared to the previous case, citizen polarization must have decreased, or elite polarization has increased since T_I has moved to the left of R_I. As a result, the citizen is no longer willing to reelect the ideologue for any proposal the ideologue is willing to make. In this case, we get a separating equilibrium in which both incumbents propose their ideal point, but neither one is reelected.

Separating Equilibrium 3.3b:

- *Ideologue strategy: Set* $p_I = x_I$
- *Closet autocrat strategy: Set* $p_{CA} = 1$
- *Citizen strategy: Reelect if* $p \leq T_I$*, replace if* $p > T_I$

Appendix B. Empirical Appendix

B.1. **Taking over Poland's Constitutional Court.** First, President Andrzej Duda refused to swear in three justices who had been elected by parliament in the previous term (so a parliament still controlled by PO) to replace judges whose terms were running out. That move was appealed to the Constitutional Tribunal, which upheld the constitutionality of the election of the three justices. The PiS cabinet's Chief of Staff refused to publish the Constitutional Tribunal's verdict, hoping to invalidate it. But after the Supreme Court ruled that the Constitutional Tribunal's verdicts have full force of the law from the moment they are handed down, regardless of publication, the three justices joined the bench.

PiS continued to deny the new judges legitimacy and elected its own three candidates. These judges (referred to as "extras") also joined the bench. But when the Constitutional Tribunal's Chief Justice, Andrzej Rzeplinski, refused to appoint them to sit on panels, PiS decided to take further action against the Court.

In the next step, PiS passed two bills that shortened the terms of the judges of the Constitutional Tribunal. This move effectively recalled from office the Chief justice Rzeplinski. A PiS loyalist, Julia Przelecka, was appointed to replace him. This was a controversial choice for at least two reasons. First, Poland's National Council of the Judiciary had evaluated—and rejected— Przelecka as lacking qualifications for a post in the Appellate Court. What is more, she began her judicial career in communist Poland, making her exactly the kind of judge PiS promised to get rid of.

Table 4. Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
PiS Voter	0.428	0.495	0	1	703
Polarization in Electorate	0.643	0.282	0	1	937
Anti-PiS Protest Supporter	0.445	0.497	0	1	926
Elite Polarization	0.471	0.321	0	1	930
Strong Leader	2	0.908	0	3	858
Male	0.461	0.499	0	1	1009
Employed	0.492	0.487	0	1	1009
Education	6.36	3.031	1	12	1009
Village	0.42	0.494	0	1	1009
City	0.335	0.472	0	1	1009
Age	51.734	17.539	19	87	1009
Religiousity	0.912	0.284	0	1	1009
Authoritarian 2	1.853	0.892	0	3	822
Indifferent	2.007	0.911	0	3	942

B.2. **Summary Statistics and additional correlations.**

Table 5. Correlation matrix

	Polarization in Electorate	PiS saving Poland	PiS Authoritarian	Anti-PiS protest supporter
Polarization in Electorate	1.00			
PiS saving Poland	0.48	1.00		
PiS Authoritarian	-0.41	-0.67	1.00	
Anti-PiS protest supporter	-0.36	-0.58	0.49	1.00

Note: P-values in parentheses

References

Ahlquist, John S, Nahomi Ichino, Jason Wittenberg and Daniel Ziblatt. 2018. "How Do Voters Perceive Changes to the Rules of the Game? Evidence from the 2014 Hungarian Elections." *Journal of Comparative Economics* .

Bermeo, Nancy. 2016. "On democratic backsliding." *Journal of Democracy* 27(1):5–19.

Carroll, Royce and Monika Nalepa. N.d. "Party Representation and the Organization of Eastern European Parliaments." . Forthcoming.

Cinar, Ipek. 2017. Democracy Dismantled: Strategic Choices of the Would-be Autocrats PhD thesis University of Chicago.

Grzymala-Busse, Anna. 2001. "Coalition formation and the regime divide in new democracies: East Central Europe." *Comparative Politics* pp. 85–104.

Jenne, Erin K and Cas Mudde. 2012. "Can outsiders help?" *Journal of Democracy* 23(3):147–155.

Lust, Ellen and David Waldner. 2015. "Unwelcome Change: Understanding, Evaluating, and Extending Theories of Democratic Backsliding." *US Agency for International Development* 11.

Sedelmeier, Ulrich. 2014. "Anchoring democracy from above? The European Union and democratic backsliding in Hungary and Romania after accession." *JCMS: Journal of Common Market Studies* 52(1):105–121.

Serra, Gilles. 2012. "The Risk of Partyarchy and Democratic Backsliding Mexico's 2007 Electoral Reform." *Taiwan Journal of Democracy* 8(1).

Svolik, Milan. 2018. "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents." *Unpublished Manuscript* .

Svolik, Milan W. 2017. "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents." *Unpublished Manuscript, Yale University* .

Weingast, Barry R. 1997. "The political foundations of democracy and the rule of the law." *American political science review* 91(2):245–263.