A WOLF IN SHEEP'S CLOTHING: CITIZEN UNCERTAINTY AND DEMOCRATIC BACKSLIDING

CATERINA CHIOPRIS

MONIKA NALEPA

GEORG VANBERG

Abstract. A prominent contemporary phenomenon is "backsliding" of democratic countries into (semi-)authoritarian practices. Importantly, such episodes unfold over time, and often involve uncertainty about the ultimate intentions of governments. Building on recent work (Svolik 2020), we present a model in which a government engages in a reform that may allow for subsequent actions that are inconsistent with the rule of law. Citizens must decide whether to replace the incumbent following the reform. Consistent with existing work, the model suggests that polarization is an important factor in democratic backsliding. More importantly, the model demonstrates that in a dynamic setting, citizens may support incumbent governments even if citizens are fundamentally opposed to authoritarianism. One consequence is that citizens may genuinely regret their electoral choices. We illustrate the model's implications using a survey experiment in contemporary Poland.

1. Introduction

Recent years have witnessed a wave of "democratic backsliding," including in countries that appeared to be well on their way to being consolidated democracies (Bermeo 2016; Lust and Waldner 2015; Serra 2012). Poland and Hungary provide two vivid examples. Following the dissolution of the Soviet empire, each quickly established a reasonably well-functioning democratic regime. By 1999, they had joined NATO, and five years later became full-fledged members of the European Union. By all appearances, both countries seemed well-ensconced among Western democracies. And yet, only a decade later, each is governed by a party that is arguably chipping away at democratic institutions and norms, raising concerns of creeping authoritarianism that have even been voiced by European Union institutions (Sedelmeier 2014; Jenne and Mudde 2012) [1].

A prominent explanation for such democratic backsliding, formalized by Svolik (2020), highlights the significance of polarization and political culture. The logic of this explanation is intuitive: If

[1] See also *Revised Statute of the European Commission for Democracy through Law* (2002); of Justice of the European Union (2019*a*,*b*); *Sedziowie pod Presja [Judges under Pressure]* (2019); *Lex Super Omnia Association of Prosecutors to the Prosecutor General* (2019)

the ideological alternative offered by the opposition becomes sufficiently unattractive, citizens may be willing to tolerate or even support incumbents with authoritarian tendencies. Polarization, which implies that citizens increasingly view "the other side" as ideologically distant, makes it more likely that this condition is met. Similarly, as citizens become less committed to democratic institutions or processes, they are more willing to trade-off perceived ideological gains for adherence to democratic norms and institutions. Experimental evidence from Venezuela (Svolik 2020) and the United States (Graham and Svolik 2020) provides powerful support for this logic.

The central theoretical mechanism in this account is that citizens value ideological gains more than an institutional commitment to democracy: A slide into authoritarianism is the price citizens willingly pay in order to secure policies they favor (or to avoid policy outcomes they dislike). Our contribution in this paper is to extend this account to consider whether democratic backsliding is also possible when citizens have strong commitments to democratic institutions, that is, when citizens would not *knowingly* support an autocrat. This question is relevant because there is at least some evidence to suggest that democratic backsliding can occur even when citizens – including those who support would-be autocrats – have strong commitments to democratic values. For example, Voeten (2016) has shown that in Poland and Hungary (where backsliding has occurred) citizens have displayed consistently high support for democracy, comparable to their western counterparts, while there has been a downward trend in democratic support in the Czech Republic, Croatia and Bulgaria (despite little evidence of democratic backsliding).[2]

We develop a theory of democratic backsliding that reconciles popular support for democracy with an ascent to power of "closet autocrats" through regular electoral channels. The key to our argument is that backsliding is a dynamic process. Incumbents with authoritarian aspirations usually do not pursue these openly or all at once. Instead, authoritarianism proceeds in incremental steps

[2] As we show in the appendix, in the survey evidence we employ for Poland, there is no difference in support for democracy across supporters of the ruling Law and Justice party (PiS) and those who cast their votes for parties in the opposition. Of course, as Svolik (2020) notes, survey responses may not be reliable indicators of true support for democracy. Nevertheless, these results suggest that backsliding does not necessarily go hand-in-hand with low attachments to democracy.

as current policy choices or institutional reforms are used to lay the groundwork for future authoritarian power grabs (Levitsky and Ziblatt 2018).[3] These dynamics are critical because there usually is disagreement among observers and citizens about the intentions that motivate specific institutional reforms. Consider the Polish and Hungarian cases again. The PiS and Fidesz governments adamantly deny that the goal of institutional reforms is to undermine democratic institutions or norms. Instead, both argue that they are designed to enhance democratic accountability by removing 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.

This uncertainty takes center stage in our model. Consistent with other arguments (Svolik 2020; Bermeo 2016), our model suggests that polarization is critical in creating circumstances that allow "closet autocrats" to gain power.[4] The degree of polarization required for backsliding depends critically on the uncertainty facing voters, and the extent to which voters are concerned about authoritarianism. Most importantly, our model shows that the dynamic nature of democratic backsliding combined with the uncertainty confronting voters can allow strategic "closet autocrats" to establish a hold on power by playing the "wolf in sheep's clothing" even if citizens are fundamentally opposed to authoritarianism – something that is not possible in existing models. One consequence is that in our model, voters may experience sincere regret when it becomes apparent that they have backed an incumbent who reveals himself to be an autocrat. Our empirical application examines this phenomenon by employing a survey experiment in the context of recent Polish elections. The results provide clear support for our argument.

The paper proceeds as follows. In the next section, we present a simple model that formalizes the theoretical logic. We then derive empirical conjectures from the model that inform our analytical

[3] Such backsliding through "stealth" has been described in a recent paper by (Luo and Przeworski 2019).
[4] But for an argument about the opposite effect of polarization see Grillo and Prato (2019)

narrative and the survey experiment we conducted around the 2019 Polish parliamentary election. A final section concludes.

2. A theory of democratic backsliding

Our model of democratic backsliding is designed to capture two key features. First, democratic backsliding is typically a dynamic process: It unfolds over time. In particular, the initial steps towards authoritarian power grabs are rarely explicitly autocratic. Rather, in most instances, governments intent on pursuing an authoritarian agenda engage in institutional reforms that will, subsequently, make it easier to act in a more authoritarian manner. For example, incumbents may reform judicial institutions to loosen judicial oversight, reform media law to reduce journalistic scrutiny, or alter electoral institutions in order to decrease the threat of effective political competition. Once such reforms have taken hold, governments may then be in a position to pursue an autocratic agenda more openly by prosecuting political opponents, manipulating the media, or raising the bar for being removed from power.

This dynamic aspect leads directly to the second key feature: Citizen uncertainty over incumbent motivations. Generally, governments do not acknowledge autocratic goals in initiating such institutional reforms; instead, they emphasize a commitment to democratic values, and claim that reforms are merely intended to facilitate "ordinary" political decisions. Loosening judicial oversight may be a way for a new democracy to extricate itself from the influence of judges with ties to a prior authoritarian regime; reform of media law may be directed at protecting privacy of citizens; electoral reforms may be guided by the goal of making politicians more accountable to voters.

To capture these aspects, we employ a two period, incomplete information model with three players: A representative citizen, two types of incumbents, and an opposition. The two period set-up allows us to model, in the simplest way possible, the dynamic aspect of the process we are considering. Specifically, we assume that in the first period, an incumbent government chooses an institutional reform that will define the actions available to the government in the second period. To model citizen uncertainty, we assume that there are two types of incumbents: One who has autocratic intentions, and one who does not. A representative citizen is uncertain about the type

of incumbent in office. Having observed the first-period institutional reform, the citizen updates her belief about the nature of the incumbent's intentions, and chooses whether to reelect the incumbent, or to replace him with the opposition. In the second period, the government-in-office chooses actions in light of its own preferences, given the maneuver room afforded by the institutional reform undertaken in period 1.[5]

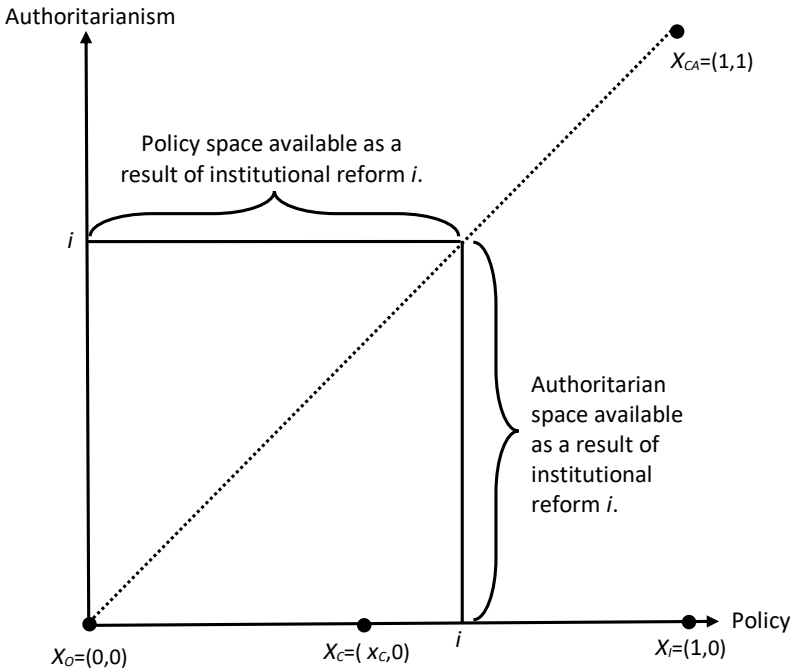


FIGURE 1. Determination of the policy space as a result of first

We model the connection between the first-period institutional reform and the actions available to second-period governments in a simple way that captures – at an intuitive level – the fact that institutional reforms can enable both "ordinary" policy goals as well as potential autocratic power grabs. Specifically, the first-period incumbent chooses an institutional reform $i \in [0, 1]$. This institutional reform defines a two-dimensional space for the second period, $[0, i] \times [0, i]$, as illustrated in Figure 1. One dimension of this space is an ordinary policy dimension. The other captures authoritarian power grabs. Put differently, institutional reform in the first period enables

[5] The idea of using a first period move to broaden executive authority in a second period is similar to the dynamic described inHowell and Wolton (2017), who describe unilateral interventions of executives into a policy domain, which a legislature would not support. These authors show that such interventions have the potential to generate new legal authority to support still more interventions into a policy domain. Another similarity between ours and Howell and Wolton's paper is that this new authority has the potential to change the electorate's evaluation of the incumbent and (potentially) his challenger

the second period government to engage in two different decisions: It can make use of the room created by institutional reform to pursue ordinary policy objectives. But it can also take advantage of the opportunities created by the institutional reform to engage in an authoritarian power grab. To use the Polish example, reforming the judiciary to reduce the influence of communist-era judges may allow a government to pursue "ordinary" policies more effectively (say, economic reform), but it may also enable it to make authoritarian moves (such as engaging in politically-motivated prosecutions of the opposition). We capture this feature by assuming that in the second period, the incumbent takes an action, $(p, a) \in [0, i] \times [0, i]$ that has (potentially) a policy as well as an authoritarian component.

To capture the representative citizen's uncertainty regarding the motivations of incumbent governments, we assume that there are two types of incumbents. An "ideological" incumbent is purely policy-motivated, and has no interest in authoritarian power grabs. The ideal point of this incumbent is given by $X_I = (1, 0)$. In words, the ideologue favors policy 1 in the ideological dimension, but has no preference for authoritarianism. The second type of incumbent is a "closet autocrat" whose ideal point is given by $X_{CA} = (1, 1)$. In words, the closet autocrat has the same preferences as the ideologue in the policy dimension, and also wants to undertake authoritarian power grabs. The citizen's ex ante belief that the incumbent is a closet autocrat is given by $Pr(\text{Closet Autocrat}) = \alpha \in (0, 1)$.

We assume that the citizen faces an electoral choice between the incumbent (over whose type she is uncertain) and an opposition challenger whose ideal point is given by $x_O = (0, 0)$, i.e. the opposition is ideologically to the left of the incumbent and is not interested in authoritarian power grabs. The citizen's ideal point is given by $X_C = (x_C, 0)$, where $x_C \in (0, 1)$. Thus, the citizen is purely policy-motivated: she favors institutional reforms in so far as they are instrumental to her preferred policies and prefers that the incumbent refrain from authoritarian power grabs.[6]

[6] The assumption that the opposition's ideal point is at $(0, 0)$ is without loss of generality; as we explain below, allowing the opposition to become less moderate, i.e., moving its policy ideal point below 0 (while maintaining its ideal point on the authoritarian dimension at 0) can only exacerbate the threat of democratic backsliding. Similarly, the assumption that the ideologue and the closet autocrat share the same ideal point in the policy dimension is not necessary; see the appendix for a more general version in which we allow the ideological preferences of the two types of incumbent to diverge.

Specifically, we assume that the citizen's preferences over the government's second-period actions are given by

$$U_C((p,a)) = -(p - x_C)^2 - \beta a^2 \quad (1)$$

where β captures the intensity of the citizen's opposition to authoritarian moves. Because we are interested in democratic backsliding when citizens are intrinsically opposed to authoritarianism, we assume that $\beta \geq 1$, which ensures that the citizen always prefers the opposition to the closet autocrat if both could implement their ideal points (even if the citizen's ideal point approaches $(1,0)$).[7]

The final element of the model concerns the preferences of politicians, i.e., the incumbent and the opposition. We assume that these actors are motivated by three concerns. The first is that they place a value $b \geq 0$ on being in office. Second, they have *instrumental* outcome concerns: They would prefer the second-period outcomes (both in terms of policy and authoritarianism) to be as close as possible to their ideal point. Finally, they have *expressive* concerns: They prefer that the policies they back publicly when in office are closer to their sincere preferences rather than further away. Substantively, one way to think about these expressive preferences is in terms of "branding:" Parties represent a set of policies they are committed to and that form the basis of their appeal. All things being equal, they prefer to announce or back policies that are consistent with these underlying commitments, and dislike those that require them to compromise the party's brand (Lupu 2014). Such expressive preferences may even lead a party to forego joining a government, if doing so requires backing policies that dilute its brand. The Polish PiS and Hungarian FiDesz parties' behavior during the early 2000's illustrates this well: Instead of compromising to make themselves politically viable, they stayed in opposition to preserve their anti-liberal brand.

The incumbent has no intrinsic preferences over the first-period institutional reform; rather, preferences over institutional reform are derived endogenously from the implications that follow

[7] 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.

from the reform for the second period.[8] Moreover, given the office and expressive concerns, the payoffs for the incumbent depend critically on whether or not she is reelected to a second term. Specifically, consider an incumbent who implements reform i, but is not re-elected. Letting $(\hat{p}, \hat{a})$ denote the second period policy chosen by the (new) government, we assume that the payoffs of the ideologue and closet autocrat are given by an additive function of standard quadratic loss across the ideological and authoritarian dimension:

$$U_I((\hat{p}, \hat{a})) = -((\hat{p} - 1)^2 + \hat{a}^2) \tag{2}$$

$$U_{CA}((\hat{p}, \hat{a})) = -((\hat{p} - 1)^2 + (\hat{a} - 1)^2) \tag{3}$$

In contrast, if the incumbent is reelected and implements policy (p, a), she secures the office benefit, but must also incur the expressive costs of implementing a policy that may not conform to her ideal point. These payoffs are given by:

$$U_I((p, a)) = -((p - 1)^2 + a^2) - \theta((p - 1)^2 + a^2) + b \tag{4}$$

$$U_{CA}((p, a)) = -((p - 1)^2 + (a - 1)^2) - \theta((p - 1)^2 + (a - 1)^2) + b \tag{5}$$

where $\theta \geq 0$ is the weight attached by the ideologue to expressive preferences. The first term in the expression captures the outcome concerns, the second the expressive concerns, and the final term the office motivations.

Summarizing, the model is defined by the following sequence of play:

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

(2) Period 1: The incumbent government chooses an institutional reform $i_j \in [0, 1]$, where $j \in \{I, CA\}$.

[8] The model's results are robust to incorporating direct preferences over the institutional reform.

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

(4) Period 2: The second-period government chooses an action $(p, a)_j \in [0, i] \times [0, i]$, where $j \in \{I, CA\}$, consisting of a choice in the policy dimension and a move on the authoritarian dimension.

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

A suitable solution concept is Perfect Bayesian Equilibrium, which requires that players' strategies are sequentially rational, and that they update their beliefs in accordance with Bayes' rule along the equilibrium path. We reserve a full statement of equilibria and proofs to the Formal Appendix. Here, we focus on presenting the intuition and substantive implications underlying the equilibria.

2.1. **Strategies.** To understand the intuition underlying the model's equilibria, and its implications for democratic backsliding, it is useful to begin by considering the core logic confronting the key players: the two types of incumbents and the representative citizen. We begin with the citizen's electoral decision. From the citizen's point of view, the incumbent's first-period institutional reform is significant for two different reasons. Most obviously, it defines the maneuver room for the second period government. As such, there are potentially two countervailing considerations. On the one hand, the reform can open up room for second-period governments to implement policies the citizen favors. On the other hand, the reform can also enable authoritarian moves, which the citizen opposes. Put differently, the citizen may be sympathetic to reforms to the extent that she is convinced that these reforms will allow for policy choices she agrees with. But she may also be concerned about reforms that loosen constraints on potential autocrats. The second significant aspect of the institutional reform from the citizen's point of view is that reform can act as a signal of the incumbent's type; that is, the nature of the reform may provide the citizen with information about whether the incumbent has autocratic goals or not.

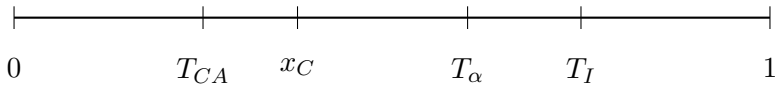
The representative citizen's strategy is characterized by a threshold that makes the electoral decision contingent on the citizen's (updated) beliefs about the nature of the incumbent, and the

extent of the institutional reform that has been adopted. Suppose, for example, that the citizen – having observed the first period institutional reform – has become convinced that the incumbent is an ideologue, and will not exploit the reforms in order to make an authoritarian move. In this case, the citizen's reelection decision simplifies to a straightforward ideological choice: She reelects the incumbent if the incumbent's first period reform implies a second period policy that is closer to her ideal point than the opposition, and votes for the opposition otherwise. As is intuitive, this threshold is given by $T_I = 2x_C$. We illustrate this threshold in Figure 2, which maps out the citizen's strategy.

In contrast, now suppose that – having observed the first period institutional choice – the citizen believes that the incumbent is a closet autocrat. This knowledge puts the citizen in a quandary: On the one hand, the policy choice of the closet autocrat may be ideologically more palatable than the alternative offered by the opposition. On the other hand, the citizen is opposed to the fact that the closet autocrat will also exploit the maneuver room created by the institutional reform to engage in authoritarian power grabs. 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 institutional reform that is so moderate as to constrain future authoritarian moves sufficiently. This threshold is given by $T_{CA} = \frac{2x_C}{\beta+1}$. 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 future behavior before she is willing to reelect a known closet autocrat.[9]

[9] This aspect of the model raises an issue that deserves future exploration: In our model, a closet autocrat is constrained by the extent of institutional reform, i.e., constitutional constraints are credible in the sense that the second period government cannot go beyond the reforms initiated in period 1. We believe this is – at least in the short to medium run – a reasonable assumption; having engaged in some reform does not mean that governments are then able to ignore all constraints. On the other hand, having once engaged in institutional reforms that loosen institutional constraints, citizens might obviously be concerned that they are embarking on a "slippery slope" and that a closet autocrat will not be constrained to act within the institutional reforms in the future. We reserve an explicit consideration of this issue for future work.

FIGURE 2. Illustration of Citizen Threshold Strategy



Finally, suppose that the citizen remains uncertain about the incumbent's type, i.e., the incumbent's authoritarian leanings, after the institutional reform. Once again, there is a threshold ($T_\alpha = \frac{2x_C}{\alpha\beta+1}$ in Figure 2) such that the citizen reelects the incumbent only if the first period reform falls below it.[10] As is intuitive, T_α falls between T_I and T_{CA}, 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}: The citizen requires more stringent constraints on the incumbent to limit potential authoritarian moves in the second period. In contrast, 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.)

Having worked through the logic of the citizen's reelection strategy, consider the two types of incumbents as they decide on institutional reform in the first period. The key issue confronting the incumbent is that the institutional reform he adopts has two separate effects: It determines the citizen's reelection decision, and it defines the maneuver room the incumbent has in the second period (if reelected). There are clear incentives to be re-elected, since office provides intrinsic benefits as well as the opportunity to make policy (including the possibility of pursuing an autocratic agenda). This implies that there is value in proposing an institutional reform that will result in reelection, even if this reform is not (from the incumbent's point-of-view) ideal. At the same time, for the incumbent, there are expressive costs associated with governing in ways that are not in line with the incumbent's preferences, that is, to have to adopt a policy package in the second period that diverges from the incumbent's ideal. As a result, there is a limit to the incumbent's willingness to adopt an institutional reform that will result in his reelection. If the institutional

[10] Note that in a more general formulation, the citizen's updated belief is given by γ, and the threshold is $T_\gamma = \frac{2x_C}{\gamma\beta+1}$. Since – as we show below – all equilibria in which the citizen remains uncertain over the incumbent's type are fully pooling, $\gamma = \alpha$, which is why we write the threshold in terms of the ex ante belief.

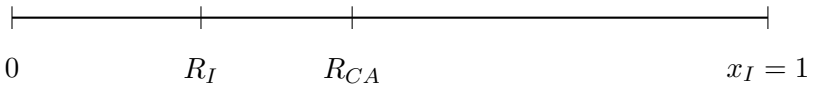
TABLE 1. Summary of thresholds

Threshold	Imposed by	Explanation
$T_{CA} = \frac{2x_C}{\beta+1}$	Citizen	Citizen will reelect known closet autocrat who sets reforms below this threshold.
$T_I = 2x_C$	Citizen	Citizen will reelect known ideologue who sets reforms 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 reforms below this threshold.
$R_{CA} = 1 - \frac{\sqrt{2+b}}{\sqrt{2(1+\theta)}}$	Closet autocrat	Closet autocrat will only set reforms at or above this threshold to get reelected.
$R_I = 1 - \frac{\sqrt{b+1}}{\sqrt{1+\theta}}$	Ideologue	Ideologue will only set reforms at or above this threshold to get reelected.

reform that ensures reelection is so modest as to constrain the second period policy choice too much, the incumbent prefers to leave office and tolerate the opposition's ascent to power.

These considerations imply that the two types of incumbent adopt a threshold strategy. For each incumbent, this threshold marks the most modest institutional reform the incumbent is willing to adopt in order to be reelected. Reforms to the left of these thresholds are so constraining as to make returning to power sufficiently unattractive. These thresholds are summarized in Table 1 (along with the thresholds imposed by the citizen), and Figure 2.1 provides a graphical illustration, where R_{CA} denotes the constraint imposed by the closet autocrat and R_I denotes the threshold imposed by the ideologue. Note that $R_{CA} > R_I$. This is a critical feature that plays a key role in the equilibria. The intuition behind this ordering should be clear: For the closet autocrat, institutional constraints on power are more costly than for the ideologue, since the closet autocrat is constrained in two rather than just one dimension. As a result, the closet autocrat is quicker to bristle at an institutional constraint and the institutional commitments incumbents make in the first period can potentially serve as a (costly) signal of their type.

FIGURE 3. Illustration of Incumbent Threshold Strategy



2.2. **Equilibria.** The model's five equilibria, which are unique for any combination of parameters, depend on the relative position of the thresholds that characterize the citizen's and the incumbents' strategies. The equilibria are described in the formal appendix.[11] Here, we focus on the substantive implications, which are most readily explored by grouping the equilibria into three categories. Two are separating equilibria in which the ideologue and the closet autocrat adopt different institutional reforms in the first period, thus allowing the citizen to learn the type of incumbent she is facing. Three are pooling equilibria in which both types of incumbents adopt the same institutional reform, which implies that the citizen continues to face uncertainty about the nature of the incumbent's intentions. In two of these, the citizen nevertheless reelects the incumbent. In the third, the citizen replaces the incumbent with the opposition. The three categories of equilibria are as follows:

(1) Given the institutional reforms adopted, the citizen replaces both types of incumbent with the opposition.
(2) Given the institutional reforms adopted, the citizen learns which type of incumbent she is facing, and only reelects the ideologue. She replaces the closet autocrat with the opposition.
(3) Given the institutional reforms adopted, the citizen chooses to reelect both the closet autocrat and the ideological incumbent.

To discuss the substantive interpretation of these equilibria, and the implications for democratic backsliding, it is useful to graph the location of these equilibria in the parameter space. We do so in Figure 52.2, which shows where each equilibrium type occurs as a function of the citizen's ideal point, x_C, (plotted on the x-axis) and the citizen's uncertainty about the type of incumbent, α, (plotted on the y-axis). The solid lines indicate the partition of the three types of equilibria. The light dashed lines indicate the partition of the equilibria that fall under each type. In addition, the

[11] As detailed in the Formal Appendix, we impose several restrictions on off-equilibrium path beliefs that serve to eliminate "nuisance equilibria" that are substantively uninteresting.

figure displays, for each equilibrium, the institutional proposal made by the ideological incumbent (i_I) and the closet autocrat (i_{CA}) in the first period.

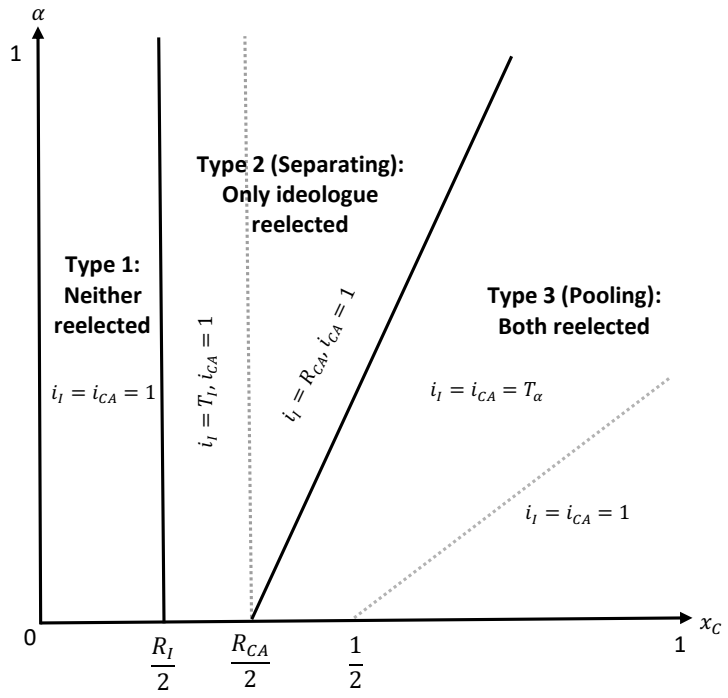


FIGURE 4.

The partition reflected in the figure is intuitive. In the southeastern, triangular, area, the citizen is ex ante relatively confident (as indicated by α close to 0) that the incumbent is an ideologue, and the citizen has ideological preferences that are closer to the ideologue than to the opposition. As a result, the concern of allowing an incumbent to "get away" with an extreme institutional reform ($i = 1$) is outweighed by the desire to give the ideological incumbent sufficient leeway: The citizen is willing to reelect the incumbent even though the initial reform is extreme; consequently both types of incumbents adopt that reform. Now consider what happens as the citizen becomes more moderate, or the ex ante belief that the incumbent may be a closet autocrat increases, i.e., as we move up or to the left in the figure (while still remaining in the pooling equilibrium of category 3). Because the citizen is now concerned that the risk of a closet autocrat is higher, she is no longer willing to reelect following an extreme institutional reform: In order to be reelected, the incumbent must adopt a "constrained" institutional reform ($i = T_\alpha$) that provides some insurance in case

the incumbent is a closet autocrat. With citizen's preferences sufficiently far to the right, this "institutional assurance" that the citizen requires is one that even the closet autocrat can provide: Once again, the two incumbent types pool, and the citizen reelects.

Now consider the separating equilibria of category 2 that occur as the citizen's ideological preferences become centrist (i.e. as x_C moves left, it is neither too close to the opposition, nor too close to the ideologue). Given that the citizen's preferences are now moderate, she is no longer willing to provide significant leeway to the incumbent because satisfying her policy preferences no longer requires far-reaching institutional reforms and only pose a risk of authoritarianism. As a result, the citizen will not reelect unless the first period reforms are sufficiently modest. The ideological incumbent is willing to respect this constraint and adopts a moderate reform. But for the closet autocrat, the costs of playing the "wolf in sheep's clothing" are now too high: He is not willing to display the moderation required to be reelected. The result is that the citizen learns the type of incumbent and only returns the ideologue to office.

Finally, consider the pooling equilibria of category 1 in the left-hand side of the figure. Again, the intuition behind these equilibria is not hard to see. Given that the citizen's preferences closely approximate those of the opposition, voting for the opposition is attractive. To convince the citizen to choose the incumbent instead, the incumbent would have to offer an institutional reform that is highly constraining, and close to the status quo (0). Given the expressive costs of governing within such constraints, neither the ideologue nor the closet autocrat are willing to constrain themselves in this way, and accept the fact that they will lose office.

2.3. **Substantive Implications.** We can also use Figure 2.2 to guide our discussion of the substantive implications of the model for the phenomenon of democratic backsliding. We begin by highlighting the implications of extending the logic established by Svolik (2020) to a dynamic setting in which citizens face uncertainty over the intentions of incumbents.

Implication 1. *As citizen preferences become sufficiently extreme and aligned with the position of the ideological incumbent, the two types of incumbent adopt the same reform in period 1, and*

are reelected by the citizen. Democratic backsliding in period 2 is possible. When it occurs, the citizen—ex post—regrets her electoral choice.

This implication captures the pooling equilibria in the right part of the parameter space (category 3). The closer the citizen's ideology aligns with the incumbent, two dynamics begin to play out. The first overlaps with the logic of Svolik (2020)'s model: Replacing the incumbent with the opposition becomes less palatable for the citizen. Even if there is some possibility that the incumbent is an autocrat, the citizen is more willing to reelect the incumbent in order to obtain policies that are ideologically proximate. The second dynamic concerns the distinct element of our model: Citizen uncertainty about the type of incumbent. As the citizen's preferences shift to the right, so does the citizen's tolerance for radical institutional reform. The consequence is that the cost to the closet autocrat of playing the "wolf in sheep's clothing" declines. Once the citizen is sufficiently far to the right, a strategic closet autocrat is willing to adopt the same institutional reform as the ideologue in order to be reelected—in effect "masquerading" as an ideologue. Given her right-leaning preferences, the citizen—despite her uncertainty about the nature of the incumbent—takes a gamble and chooses to reelect. If the incumbent is in fact a closet autocrat, he will exploit the space created by the institutional reform to pursue authoritarian policies in period 2. Democratic backsliding occurs. Critically, in such a scenario, the citizen feels genuine regret: she would—given the ideological and autocratic policy choices being made by the closet autocrat in the second period—prefer to "swap out" the closet autocrat for the opposition. But it is now too late. Put differently, unlike in models in which citizens knowingly reelect incumbents because they value ideological payoffs over democratic commitments (Svolik 2020; Howell and Wolton 2017; Helmke, Kroeger and Paine 2019), in a dynamic setting characterized by uncertainty, backsliding is possible even when citizens are opposed to authoritarianism – leading to the potential for genuine regret on the part of citizens. This results is most similar to Grillo and Prato (2019), who predict the emergence of "opportunistic autocrats" who become more likely with the uncertainty of citizens about incumbent types and to Luo and Przeworski (2019), who identify in their model an

equilibrium where the incumbent has taken so many small steps towards consolidating his power that by the time the citizens realize they are dealing with an autocrats he is no longer removable.

Implication 2. *As the citizen is (a) less concerned about democratic backsliding (β decreases), or (b) believes that it is less likely that she is facing a closet autocrat (α approaches 0), the pooling equilibria become possible even for more moderate citizen preferences. As a result, democratic backsliding is easier to achieve.*

This implication highlights that there is an interactive relationship between the level of polarization (in terms of the location of x_C relative to x_O) necessary to open the door to democratic backsliding, and the degree to which the citizen is concerned about the threat of authoritarianism. As the citizen believes that it is less likely that she is confronting a closet autocrat (α is low), or as she cares less deeply about preventing autocratic moves (β approaches 1), the citizen is more willing to take a gamble on reelecting an incumbent about whose intentions she is uncertain (T_α moves to the right). And as the citizen becomes more willing to reelect incumbents even if they adopt more extreme reforms, strategic closet autocrats perceive an opportunity to masquerade as ideologues by adopting institutional reforms that will allow them to be reelected. In other words, political environments in which citizens believe the threat of authoritarianism is low, or in which citizens have shallow commitments to democratic institutions, are fertile ground for closet autocrats to begin infiltrating a democratic polity—they invite "wolves in sheep's clothing."

A final implication results from the fact that institutional reforms in the first period serve competing purposes: On the one hand, they can enlarge the maneuver room for "regular" policy-making in the second period. On the other hand, they serve as a potential signal about the intentions of the incumbent. Finally, they have implications for the extent to which a closet autocrat is constrained in pursuing an authoritarian agenda. Consider the result: If the citizen believes that it is sufficiently likely that the incumbent may be a closet autocrat, and the citizen is sufficiently concerned about democratic backsliding, she has incentives to use the decision regarding institutional reforms to "weed out" incumbents she is not willing to support: She will only reelect the incumbent if the first period policy reform is modest – specifically, so modest that only the ideologue will be willing

to implement it. Significantly, this implies that the first period reform can be *below* the citizen's ideal point: The citizen knowingly supports an institutional environment that places constraints on policymaking that prevent the second-period policymaker from adopting the policy the citizen prefers. She does so because "sticking to" this institutional environment is necessary in order to sideline the closet autocrat:

Implication 3. *Under some conditions, the citizen enforces limits to institutional reform that will separate the closet autocrat from the ideologue, but which also prevent the policymaker from adopting the citizen's preferred policy: The citizen is trading away ideological proximity in order to prevent democratic backsliding.*

This last implication depends critically on the interaction between the citizen's (*ex ante*) belief that she is dealing with a closet autocrat, and her concern about preventing democratic backsliding. As the citizen becomes more accepting of authoritarianism (β declines), she is less and less willing to enforce strict limits on first-period institutional reforms – and this is especially the case when the citizen believes that it is unlikely that the incumbent is a closet autocrat. This combination of conditions suggests that "consolidated" democratic polities face a potential challenge precisely because of their established democratic traditions: In such circumstances, citizens may be less wary of the motivations of those who compete for political power, which potentially opens the door for closet autocrats to gain a foothold.

3. Analytical Narrative

In the remainder of the paper, we provide an analytical narrative to explore how well the intuition of the model comports with an important case: contemporary Poland. Analytic narratives are a common tool for illustrating the logic of formal models, and are particularly well-suited to examining the actors' beliefs about "off equilibrium path behavior" that are typically central to the explanation (Levi 2004; Bates 2007; Vanberg 2000; Nalepa 2010; Lorentzen 2014; Lorentzen, Fravel and Paine 2015; Goemans and Spaniel 2016). In our case, the focus of the analysis are the pooling equilibria, since it is in these equilibria that democratic backsliding occurs.

The logic captured in these equilibria is that citizen beliefs about the intentions of incumbents can be central to explaining backsliding: Voters who are ideologically proximate to the incumbent but committed to democracy may nonetheless choose to support an incumbent if there is sufficient uncertainty about the ultimate aims that an incumbent is pursuing. However, an increase in the belief that the incumbent is a closet autocrat may push these voters over their "belief threshold," and persuade them to support the opposition. Accordingly, our analytical narrative proceeds in two parts. First, we provide a brief overview of recent developments in Polish politics – specifically, controversial institutional reforms implemented by the incumbent government since 2015 – to establish that citizens plausibly faced substantial uncertainty over the nature of the incumbent. Second, we employ a survey experiment, conducted to coincide with the 2019 Sejm and Senate elections, in which we manipulate the beliefs of voters who are ideologically proximate to the incumbent about the likelihood that the incumbent is a closet autocrat. As predicted by the model, we find that such a change in beliefs substantially reduces the likelihood that these voters will continue to support the incumbent, and that this effect is particularly strong for those voters who are most likely to have deep democratic commitments.

3.1. **The context: Poland since 2015.** After a long period on the opposition benches, the Polish Law and Justice party (PiS) emerged from the October 2015 parliamentary elections with a clear victory, securing an absolute majority of seats in the legislature and forming the first single-party majority cabinet in Poland since 1989. Almost immediately, the government began to embark on a series of controversial institutional reforms, targeting in particular the Polish judiciary (Duncan and Macy 2020). The first of these efforts commenced shortly after the PiS government took office. In October 2015, just weeks before the parliamentary election, the incumbent legislative majority – led by the Civic Platform (PO) – had elected five justices to the Constitutional Tribunal to replace justices whose term was due to expire later that year, *after* the upcoming legislative elections.[12] The PiS President, Andrzej Duda, refused to accept their appointment. Instead, shortly after winning

[12] Poland's system of courts separates the function of the Supreme Court from the constitutional court, called the Constitutional Tribunal. In addition to being the court of appeal for lower-level courts, the Supreme Court is responsible for determining the validity of elections and nationwide referenda. The Constitutional Tribunal has jurisdiction over the constitutionality of legislation and executive action.

the parliamentary elections, the newly formed PiS government elected its own slate of five justices to replace those whose term was coming to an end – only to have the Constitutional Tribunal's chief justice refuse to include the newly-appointed justices in any judicial panels.

This dispute set the stage for the first major judicial reform initiated by the PiS government, adopted in December 2015. This legislation, aimed squarely at the Constitutional Tribunal, forced inclusion of the newly appointed justices by increasing the decision quorum of the court, raised the required majority for declarations of unconstitutionality to two-thirds, and introduced the possibility of removing justices of the Tribunal by a majority vote of the Sejm. Over the following years, the PiS government then engaged in systematic reform of the remaining judiciary. It changed the composition of the National Council of the Judiciary (KRS), the independent body that makes recommendations for judicial appointments (including to the Supreme Court) and initiates disciplinary action against members of the judiciary. Prior to the bill put forward by PiS, the KRS was made up exclusively of judges. The new PiS legislation dismissed the existing members, and replaced them with political appointees. Turning to the Supreme Court, legislation adopted in 2017 imposed a mandatory retirement age of 65 for Supreme Court judges, with exemptions requiring presidential approval, effectively forcing out roughly 40% of sitting Supreme Court judges. The prerequisites for holding a Supreme Court seat were lowered to a minimum of 12 years of experience in a regional court. Simultaneously, lower courts were reformed as well. Judges above the age of 65 were forced to retire unless they received an exemption from the Minister of Justice. Finally, in December 2019, the PiS government adopted legislation that empowered the disciplinary chamber of the reformed Supreme Court (now with a comfortable majority of PiS appointed judges) to investigate and punish lower court judges for "political activities" and other misconduct.[13]

3.2. **Ideologue or wolf in sheep's clothing?** Condemnation of PiS' reform efforts was swift and significant, both domestically and from abroad. Many of the reforms were greeted with wide-spread

[13] As a recent report makes clear, all judges facing disciplinary action under the new law had, in one way or another, objected to the PiS-led judicial reforms (*Sedziowie pod Presja [Judges under Pressure]* 2019). Similarly, prosecutors have been facing harassment. While prosecutors are not subject to disciplinary action by the Supreme Court, the Prosecutor General—a post also controlled by PiS—can take measures to make their lives difficult, for instance by moving them to geographically less appealing assignments.

public protest, motivated by a concern that PiS was attempting to ensure a compliant judiciary that would enable it to undermine democratic competition and effective opposition. Similar concerns about "creeping authoritarianism" were raised by international observers, some Western governments, and – most significantly – the European Union, which launched formal proceedings to investigate whether the reforms violate Poland's obligations to the rule of law under EU treaties.

It is not difficult to see why critics of the PiS government suspected – in the language of our model – that the reform efforts might be the work of a closet autocrat. Tinkering with the National Council of the Judiciary and the composition of the Supreme Court carries enormous payoffs for an incumbent who is, in fact, a closet autocrat. The reason is simple: Control of the Supreme Court provides an incumbent government with significant resources to erect barriers for political opponents. For one thing, such control might make it easier to engage in political prosecutions, a common tactic in post-Soviet states. For example, the Polish government has considered putting Donald Tusk, European Council president from 2014 to 2019 and former leader of the most influential opposition party (Civic Platform (PO)), on trial before the State Tribunal.[14] The chief justice of the Supreme Court serves, *ex officio*, as the justice presiding over this body. In addition, the Supreme Court plays a central role in Poland's public system of financing electoral campaigns. Upon clearing a threshold of 3% of the national vote, parties are eligible for (partial) reimbursements of campaign expenditures – but only *after* the party's finances have been declared "in order" by the Supreme Court. As a result, a PiS-controlled Supreme Court could potentially be used to reduce financial support for opposition parties, and stifle electoral competition. Finally, control of the Supreme Court and Constitutional Tribunal, in combination with reforms of the lower courts, reduces the risk of judicial interference in the government's decisions. The potential for reversals of lower court decisions by government-friendly high courts, coupled with the expanded disciplinary role of the Supreme Court, ensures that there are strong incentives for lower court judges not to challenge the government.

[14] Tusk's alleged crime is the murdering of 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.The State Tribunal is a special judiciary body for assessing the constitutional liability of persons holding the highest state rank.

While opponents of the reforms were alarmed by what they perceived to be the dangers of creeping authoritarianism, this was by no means the only potential interpretation of PiS's efforts. Indeed, the government offered several plausible explanations that aimed to assure its supporters that it had no designs to undermine the constitution. Perhaps the most salient of these took direct aim at the authoritarian past, and argued that far from threatening democratic rule, the reforms were needed to eliminate the (undemocratic) power of "hold over" judges of the communist dictatorship. During the anniversary celebrations of the Gdansk Agreements,[15] as crowds of protesters gathered and chanted "Constitution, Constitution!," President Duda defended the governments' reform efforts with an explicit reference to these communist-era judges: "You are allowed to protest because I respect the constitution. But rest assured that a majority of people living on the coast feel threatened by the fact that there are justices on the benches of courts who sentenced members of the opposition during Martial Law." In other words, the government argued that the fact that the judiciary had been exempted from the lustration and decommunization process following the end of communist rule necessitated reforms aimed at curbing pre-democratic influence within the judiciary (Nalepa 2010). Moreover, the government argued, there were practical reasons for the reform. Polish courts have been notoriously inefficient; commercial cases in particular often face delays of several years, contributing to lost foreign investment and slower economic growth. Increasing the accountability of judges, the government argued, was part of an effort to create a more favorable economic climate. In short, Polish citizens – particularly those sympathetic to the ideological positions of PiS – arguably faced genuine uncertainty about the ultimate motives behind the PiS-initiated reforms. Were these the machinations of a closet autocrat, or merely reforms needed to increase efficiency and to finally address the communist legacy of the judiciary?

4. Evidence from Survey Experiment

The aim of the previous section was to establish that there was plausible uncertainty regarding the ultimate intentions motivating the PiS' governments judicial reform efforts. As a result, voters

[15] These agreements were the result of negotiations that took place between Poland's communist government and the independent Solidarity trade union in 1980, and that are seen by many as the beginning of the end of communist rule in Europe.

who are ideologically aligned with PiS might find themselves precisely in the situation captured by the pooling equilibrium of our model, in which democratic backsliding can occur *even if* voters have deep commitments to democratic values. Central to this argument is the fact that in the face of uncertainty about the nature of the incumbent, voters who favor the ideological position of the incumbent have a "belief threshold" such that they are willing to vote for the incumbent if they believe that the likelihood that the incumbent has authoritarian ambitions is below this threshold, but will vote for the opposition (despite the fact that it is ideologically less appealing) if they believe that this likelihood is above the threshold. The deeper the commitment of a voter to democratic norms, the lower the voter's threshold.

Because beliefs about the nature of the incumbent are the key feature that distinguishes our account from those in which citizens knowingly trade-off policy gains for democratic erosion, we conducted a survey experiment designed to examine the implication of this "belief threshold" for voter behavior. Specifically, our experiment, conducted immediately after the 2019 Polish parliamentary elections, includes a manipulation designed to boost the belief of voters who are ideologically close to the incumbent (the treatment group) that the incumbent PiS government harbors authoritarian tendencies. In contrast, a second sample of voters who are ideologically proximate to the incumbent – the control group – received information that should not affect their beliefs regarding PiS's autocratic intentions. Because this manipulation may push some voters over the belief threshold, we expect that the treatment will significantly reduce the willingness of treated voters to support PiS, particularly in the case of voters with significant attachments to democratic norms.

Empirical Expectation. *The treatment will reduce the propensity of voters who are ideologically close to the incumbent to support PiS electorally, and this effect will be stronger for voters with deeper democratic attachments.*

To conduct our experiment, we commissioned a survey of a representative sample of eligible Polish voters. The survey was conducted by the Center for Public Opinion Research (CBOS) and imbedded in their monthly omnibus survey, providing us with a host of demographic variables

that we use to check covariate balance across the treated and control PiS voters (see Figure 5). The survey was conducted face-to-face from November 22 to November 25, 2019 with the 2019 parliamentary elections occurring on October 13, 2019. Our manipulation was inserted in the middle of the survey. At the end of the surveys voters were asked which party they would support if an election were held "next Sunday." Finally, respondents were asked to estimate the probability that they would vote for their chosen party if an election were held.[16]

The survey and experiment were designed to address three challenges:

(1) Identify voters who are ideologically close to the incumbent,

(2) Assess the depth of voters' democratic commitments,

(3) Provide a plausible treatment to manipulate voters' beliefs about the nature of the incumbent.

To address the first challenge, we focus our analysis on the sample of voters who reported voting for PiS in the 2019 parliamentary elections (which preceded the survey by a few weeks). These voters, who constitute roughly 50% of the survey respondents, comprise the group whose ideological preferences are most likely to be aligned with the incumbent.[17] Voters in this group were divided into a control group (whose beliefs about the nature of the incumbent would not be manipulated) and a treatment group (among whom we attempted to increase the estimation that the incumbent government is a closet autocrat).

Second, our analysis requires a measure of the depth of voters' democratic commitments, since these attitudes condition the magnitude of the treatment effect: Voters with stronger democratic attachments have a lower "belief threshold." As has been noted in the backsliding literature, assessing respondent's democratic commitments through surveys poses a significant measurement challenge (Svolik 2020; Graham and Svolik 2020). Luckily, we can draw on recent work in the literature on post-communist legacies to operationalize this concept. This work demonstrates that

[16] Respondents were offered to option to mark the party of their choosing on a tablet that prevented the interviewer from seeing their choice, and respondents were made aware that their choice would be private.

[17] Following the battery of demographic questions, respondents were asked if they had voted in the October 2019 elections. Those who answered affirmatively were then asked which party they had voted for. Roughly 50% of respondents who reported voting indicated a vote for PiS. Nearly half of the sample was comprised of non-voters, which are not included in our dataset

socialization under communism as opposed to under democratic regimes is a significant factor in shaping attitudes towards democracy (Pop-Eleches and Tucker 2017). As Pop-Eleches and Tucker (2017) argue, communist regimes inculcated citizens with a set of Marxist-Leninist values that are incompatible with constitutional democracy and representative government. This process involved several institutions and organizations, including schools, universities and work places. In contrast, younger cohorts, educated *after* the demise of communism and the transition to democracy, were socialized to democratic values, including the rule of law. Consistent with this argument, Pop-Eleches and Tucker (2017) show that democratic commitments are strongest among younger cohorts and weaker among those educated under communist rule, and advocate – in the context of post-communist polities – the use of age cohorts as a proxy measure for democratic commitments. We follow their advice. Specifically, we divide respondents into three age cohorts

- Those younger than 30 (baseline),
- Those between ages 30 and 55, and
- Those older than 55.

These three age cohorts correspond to the socialization periods identified by Pop-Eleches and Tucker (2017). Those younger than 30 were born after the transition to democracy, and thus were not exposed to communist values. Those between the ages for 30 and 55 spent their youth – including their formative educational period – under communism, but most of their adult life in a democratic society. Finally, those older than 55 spent both their formative years and the start of their professional careers under communism.

This brings us to the final challenge: Designing a manipulation to boost voters' estimation that the incumbent might be a closet autocrat. A central concern was to ensure that the manipulation would be perceived as non-partisan, plausible, and authoritative. To accomplish this, we take advantage of the fact that the European Union launched an investigation of the Polish judicial reforms to assess the compatibility of the reforms with EU member states' treaty obligations to democratic governance and the rule of law. The final opinion of the European Court of Justice's in this case was issued shortly before the survey, but had not received significant media attention

in Poland. We use a vignette surrounding this decision as our treatment. Specifically, respondents in the treatment group listened to the following vignette:

> "Last month, the Court of Justice of the European Union issued a judgment on Poland's law reforming the Supreme Court introduced by the government in the previous term. According to this judgment, the law is incompatible with Charter of the European Union law for the following three reasons:
>
> (1) it shortens terms of judges over 60 years old (in the case of women) or 65 years old (in the case of men);
>
> (2) It allows the Minister of Justice to extend the retirement age of judges on a case by case basis;
>
> (3) it treats men and women unequally."

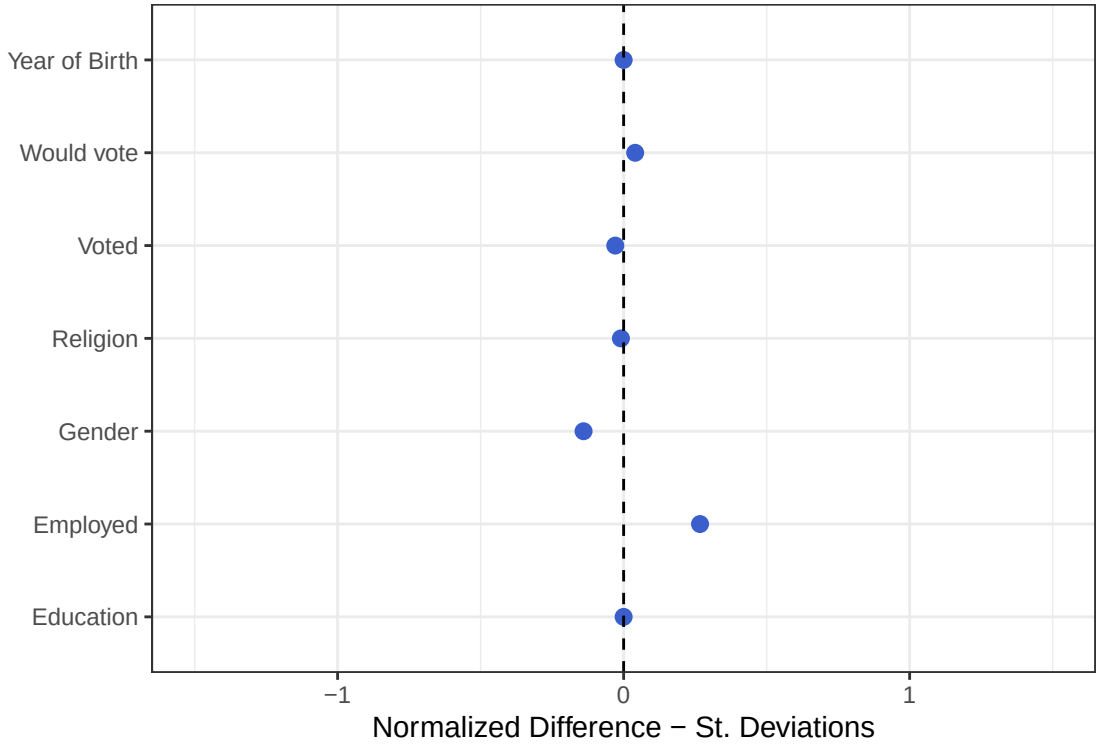
To force respondents to consciously reflect on the fact that an authoritative outside source had come to the conclusion that the government's reforms were incompatible with Poland's commitment to the rule of law as an EU member, respondents were asked to indicate which of the three aspects they thought was most important to the ECJ's decision.

Meanwhile, the control group of PiS voters received the following vignette:

> "On August 4 2019, the European Commissioner for Health, Vytenis Andriukatis, issued a comprehensive report on tobacco consumption in the European Union. The report indicated
>
> (1) that young people are initiated into the use various new products, such as heated tobacco products and e-cigarettes,
>
> (2) that the proportion of smoking Europeans aged 15-24 has increased (to 29%),
>
> (3) that the EU in an effort to combat the spread of nicotine addiction will introduce tobacco traceability and security systems."

As in the case of the treatment vignette, respondents were asked which of the three items were most important. Our expectation is that the treatment vignette will boost the belief of treated voters that the incumbent government may be a closet autocrat, while the control vignette (while

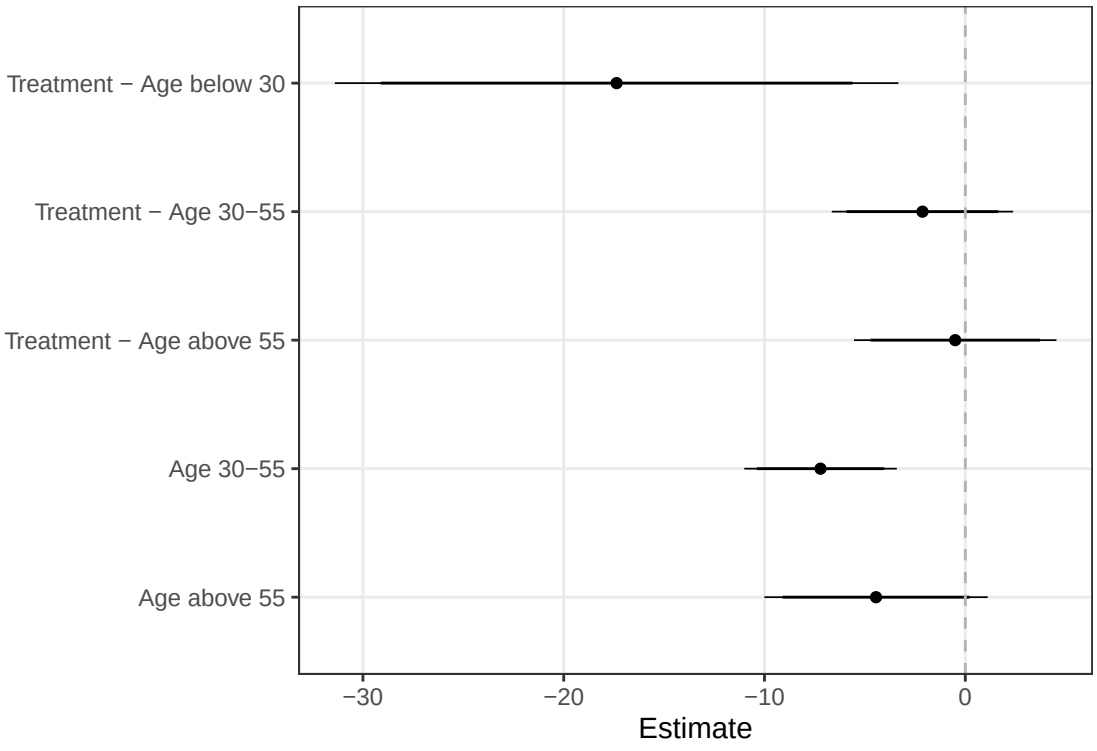
FIGURE 5. Covariate Balance across Treatment and Control Groups (Standardized Variables)



also focused on an EU report) would leave this belief unaffected. Before turning to a discussion of results, Figure 5 shows that the control and treatment groups were balanced along a number of basic demographic variables.

The key question is how our treatment affects voters' support for PiS. Thus, the dependent variable we focus on is the reported probability of voters that they would support PiS in an upcoming election. We examine the effect of our treatment across the three age cohorts, which represent our proxy measure for the depth of voters democratic commitments. Our central results are presented in Figure 6. (The table with the corresponding regression results is provided in the Quantitative Appendix). The results are clearly consistent with our expectations. For the youngest cohort of voters – those who have been raised entirely in a democratic environment – exposure to the treatment decreases the probability of voting for PiS by about 17.37% compared to the control group. As expected, for the next cohort – individuals who experienced communist era education, but lived most of their adult lives under democracy – the treatment effect is significantly reduced: the treatment reduces the probability of voting for PiS by about 2.13%. Finally, for those above

FIGURE 6. Coefficient Plot of Treatment Effect on Propensity of PiS voting by Age Cohort



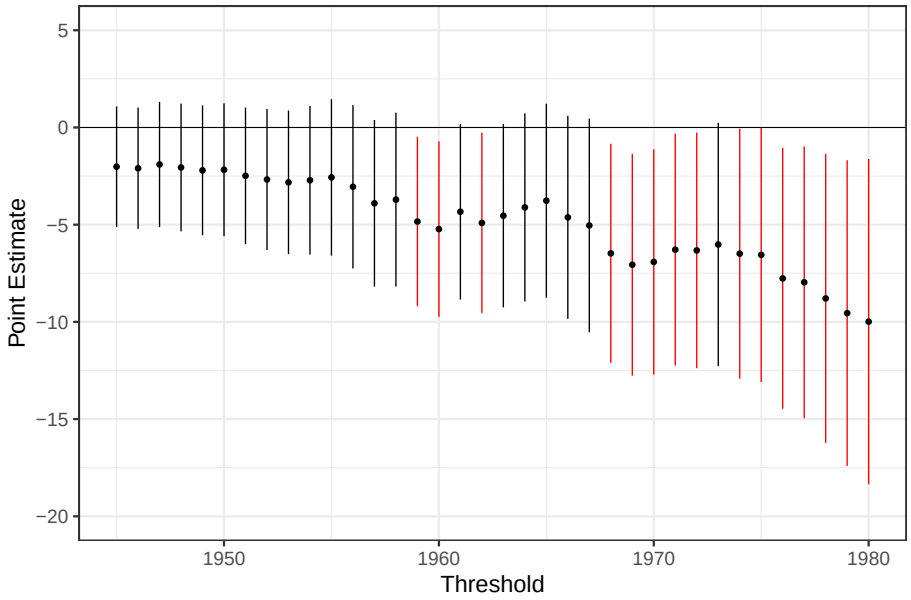
age 55, the treatment effect is insignificant – a finding that is consistent with the notion that among those who were socialized entirely under communism, democratic commitments are insufficiently deep for the treatment to affect their vote intention.

We present an alternative way of seeing the socialization/age effects in Figure 7. Here, we show the average treatment effect on the full sample (on the far left of the graph), and then successively restrict the sample by dropping respondents born before the relevant cut-off year. Because sample size decreases as we do so, the confidence intervals are getting wider and wider. As the plot reveals, for samples restricted to respondents younger than 45, the treatment effect is always negative and significant. These are respondents whose socialization under communism at most would have ended by middle school, and many of whom would not have been exposed to communist education at all.

5. CONCLUSION

Our aim in this paper has been to contribute to the emerging literature on democratic backsliding – an important phenomenon in a number of contemporary democracies. Existing literature has

FIGURE 7. Change in probability of PiS voting as result of treatment as a function of birth date



focused almost exclusively on the central role of polarization in the electorate in explaining why voters might be willing to support ideologically proximate parties and candidates, even if they have authoritarian tendencies (Svolik 2020; Graham and Svolik 2020). Others scholars have examined polarization of elites, and stress the importance of majoritarian political institutions that may allow governments to "shut out" opposition influence (Helmke, Kroeger and Paine 2019; Howell and Wolton 2017). Our contribution in this paper is to extend this explanation by focusing on the fact that democratic backsliding is a gradual process. Would-be-autocrats typically reveal their authoritarian intentions over time. This has an important consequence: How voters interpret the actions taken by incumbent governments – and in particular, whether those actions indicate authoritarian designs – becomes critical. Even voters who are opposed to authoritarianism may, unwittingly, promote backsliding. Moreover, voter uncertainty over the intentions of an incumbent can provide opportunities for would-be autocrats to "infiltrate" a democratic system as wolves in sheep's clothing. By the time voters discover that those whom they have backed are closet autocrats, it may be too late (Luo and Przeworski 2019).

Put differently, the argument we have presented is consistent with the central role of polarization that other scholars have stressed, but it demonstrates that increasing polarization – as captured by

more extreme voter preferences – matters for two distinct reasons. The first – an insight originally developed by Svolik (2020) – is that polarization makes electoral punishment of the incumbent more difficult by making the opposition less attractive to potentially pivotal citizens. The second derives from the uncertainty that voters face about the incumbent's intentions in a dynamic setting: By making more extreme institutional reforms electorally viable, polarization opens the door to strategic closet autocrats who may enter the political system because they see an opportunity to begin laying the groundwork for authoritarian power grabs while retaining power. Being uncertain about the incumbent's ultimate intentions, citizens – given sufficiently extreme preferences – may be willing to reelect an incumbent only to discover afterwards (and with regret) that the incumbent is in fact a closet autocrat.

As the model demonstrates, how much polarization is required in order to give rise to these dynamics depends critically on voter beliefs about the nature of the incumbent, and the extent to which voters are concerned to prevent authoritarianism. The less heavily concerns over autocracy weigh on citizens, or the more citizens are convinced that the incumbent is an ideologue, the more likely it is that even moderate levels of polarization can set off the process of democratic backsliding. This insight has clear implications for established democracies. In such settings, citizens have little experience with autocrats and may be less attuned to the possibility that incumbent actions are motivated by authoritarian intentions. This may be particularly true when closet autocrats do not present themselves in the guise of newly created parties, but instead emerge as a faction within an existing party with long-established democratic credentials. In such scenarios, citizens may – mistakenly – be too confident that they are dealing with an ideological incumbent. In short, it is exactly in established democracies, where citizens believe that incumbents are ideologues rather than closet autocrats that a wolf in sheep's clothing may have an easier time coming to power.

References

Bates, Robert H. 2007. From case studies to social science: A strategy for political research. In *The Oxford Handbook of Comparative Politics*.

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

Duncan, ALlyson and John Macy. 2020. "Judges on the march. the collapse of judicial independence in Poland." *Judicature* 104:40–50.

Goemans, Hein and William Spaniel. 2016. "Multimethod research: A case for formal theory." *Security Studies* 25(1):25–33.

Graham, Matthew H and Milan W Svolik. 2020. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *American Political Science Review* 114(2):392–409.

Grillo, Edoardo and Carlo Prato. 2019. "Opportunistic Authoritarians, Reference-Dependent Preferences, and Democratic Backsliding." *Reference-Dependent Preferences, and Democratic Backsliding (October 25, 2019)* .

Helmke, Gretchen, Mary Kroeger and Jack Paine. 2019. "Exploiting Asymmetries: A Theory of Democratic Constitutional Hardball.".

Howell, William G and Stephane Wolton. 2017. "The politician's province." *Available at SSRN 2545083* .

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

Levi, Margaret. 2004. "10 An analytic narrative approach to puzzles and problems." *Problems and Methods in the Study of Politics* p. 201.

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

Lex Super Omnia Association of Prosecutors to the Prosecutor General. 2019.

Lorentzen, Peter. 2014. "China's strategic censorship." *American Journal of Political Science* 58(2):402–414.

Lorentzen, Peter L, M Taylor Fravel and Jack Paine. 2015. "Using process tracing to evaluate formal models.".

Luo, Zhaotian and Adam Przeworski. 2019. "Subversion by Stealth: Dynamics of Democratic Backsliding." *Available at SSRN 3469373* .

Lupu, Noam. 2014. "Brand dilution and the breakdown of political parties in Latin America." *World Politics* 66(4):561–602.

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

Nalepa, Monika. 2010. *Skeletons in the closet: Transitional justice in post-communist Europe*. Cambridge University Press.

of Justice of the European Union, Court. 2019*a*. "The Polish legislation concerning the lowering of the retirement age of judges of the Supreme Court is contrary to EU law." Resolution(2002)3 adopted by the Committee of Ministers.

of Justice of the European Union, Court. 2019*b*. "The Polish legislation concerning the lowering of the retirement age of judges of the Supreme Court is contrary to EU law.".

Pop-Eleches, Grigore and Joshua A Tucker. 2017. *Communism's shadow: Historical legacies and contemporary political attitudes*. Vol. 3 Princeton University Press.

Revised Statute of the European Commission for Democracy through Law. 2002. Resolution(2002)3 adopted by the Committee of Ministers.

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.

Sedziowie pod Presja [Judges under Pressure]. 2019.

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

Svolik, Milan. 2020. "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents." *Quarterly Journal of Political Science* 15:3–31.

Vanberg, Georg. 2000. "Establishing judicial independence in West Germany: The impact of opinion leadership and the separation of powers." *Comparative Politics* pp. 333–353.

Voeten, Erik. 2016. "Are people really turning away from democracy?" *Available at SSRN 2882878* .