

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. Such episodes often unfold gradually over time in contexts where the ultimate intentions of governments are not clear. We present a model that focuses on the role of such uncertainty in backsliding. In the model, 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 increases 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. The less concerned citizens are about facing a potential autocrat, as in relatively well-established democracies, the more likely it is that even moderate levels of polarization can set off the process of democratic backsliding. 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 vivid examples. Following the collapse of the Soviet bloc, each quickly established a 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 (Sedelmeier 2014; Jenne and Mudde 2012).¹

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

A prominent explanation for such democratic backsliding, formalized by Svoblik (2020), highlights the significance of polarization and political culture. The logic of this explanation is intuitive: If 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. In short, 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 those they dislike. Experimental evidence from Venezuela (Svoblik 2020) the United States (Graham and Svoblik 2020), and eight other countries (Graham et al. 2022) provides powerful support for this logic.

In this paper, we extend this account to show that democratic backsliding is also possible when citizens have strong commitments to democratic institutions, that is, when citizens would not *knowingly* support an autocrat. Demonstrating this possibility is relevant since there is evidence to suggest that backsliding may occur even when citizens – including those who support would-be autocrats – have strong commitments to democratic values. Voeten (2016), for example, shows that despite apparent backsliding, citizens in Poland and Hungary display consistently high support for democracy comparable to their western counterparts.²

We develop a theory of democratic backsliding that highlights how closet autocrats can come to power through regular electoral channels despite popular support for democracy. The key to our argument is that backsliding is a dynamic process. Incumbents with authoritarian aspirations rarely pursue it openly or all at once. Instead, authoritarianism proceeds in incremental steps. Piecemeal institutional reforms, typically defended with justifications that appear consistent with democratic

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 Svoblik (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.

commitments, are used to lay the groundwork for future authoritarian power grabs (Levitsky and Ziblatt 2018).³ These dynamics are critical because they generate uncertainty among observers and citizens about the ultimate intentions of incumbents: Do they pose a threat to democracy (as “a wolf in sheep’s clothing”), or are they committed to the democratic order? Consider the Polish and Hungarian cases again. The PiS and Fidesz governments adamantly deny that judicial and electoral reforms they have undertaken are designed to undermine democracy. Instead, both argue that they are motivated by a desire to enhance democratic accountability by reducing the residual influence of the former communist regime. Such uncertainty leaves voters with a dilemma: The most direct defense against a closet autocrat is to remove the incumbent from power before authoritarianism has advanced too far. But removing an incumbent who is ideologically appealing is costly if she is not, in fact, a closet autocrat. This uncertainty takes center stage in our model.

Consistent with other arguments (Svolik 2020; Bermeo 2016), our model suggests that polarization is central in creating circumstances that allow closet autocrats to gain power.⁴ What our approach highlights is that the degree of polarization required for backsliding depends critically on the uncertainty facing voters about the ultimate intentions that motivate incumbent behavior and the depth of voters’ democratic attachments. Combining these elements, our model shows that the dynamic nature of democratic backsliding 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. A key implication is that voters may experience sincere regret when it becomes apparent that they have backed an incumbent who reveals himself to be an autocrat.

The paper proceeds as follows. In the next section, we present a simple model that formalizes the theoretical logic. We then derive a number of empirical implications. In the second half of the paper we turn to an original survey experiment, conducted in Poland, that examines the key empirical conjecture of the model: The central role of uncertainty surrounding the intentions of an

³ Such backsliding through “stealth” has been described in a recent paper by (Luo and Przeworski 2023).

⁴ But for an argument about the opposite effect of polarization see Grillo and Prato (2020)

incumbent in driving voter choice. The results offer clear support for our argument. 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 a dynamic process: It unfolds over time. The initial steps towards authoritarian power grabs are rarely explicitly autocratic. Governments with authoritarian aspirations typically first pursue institutional reforms that will, subsequently, make it easier to act in an authoritarian manner. They may reform judicial institutions to loosen judicial oversight. They might amend media law to reduce journalistic scrutiny. They may alter electoral institutions to reduce effective political competition. Having secured such institutional changes, governments are subsequently in position to more easily advance an authoritarian agenda, for example by sidelining political opponents. 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 claim a commitment to democratic values. They defend the institutional reforms as making “ordinary” political decision-making easier. Such arguments appear plausible: For example, re-organization of the judiciary may be a way for a new democracy to extricate itself from the influence of judges with ties to a prior authoritarian regime. Reforms of media law may be directed at protecting privacy of citizens. Changes to the electoral system 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. In the first period, an incumbent government chooses an institutional reform that affects the actions available to it in the second period. The key feature of backsliding we want to examine is that citizens confront uncertainty regarding the intentions behind an institutional reform. The reform may be motivated by ordinary (and acceptable) ideological goals, or by a desire to open the door to authoritarian actions. This uncertainty creates a potential dilemma: Citizens who are

ideologically sympathetic to the “ordinary” policy goals of an incumbent (say, restricting abortion access or limiting the influence of ‘holdover’ communist judges) may nonetheless be concerned about constraining a closet autocrat (say, preventing incumbents from using the judiciary to restrict political competition, prosecute political rivals, etc.).

We capture this uncertainty through two modeling choices. First, we assume that in the second period, governments can take actions in two separate domains. The first are “ordinary” policy decisions, that is, policy choices that are acknowledged by all players as falling within the scope of acceptable democratic decision-making (though of course citizens can have ideological conflict over these). The second is an “authoritarian” dimension, that is, actions that we assume are deemed undesirable by all citizens, regardless of ideological position. In other words, this authoritarian dimension is a “valence” dimension. This modeling choice allows us to capture the possibility of backsliding in the face of firm democratic commitments by citizens.

Critically, the first-period institutional reform affects the maneuver room for the government in *both* domains: Loosening judicial scrutiny, for example, may make it easier to pursue ideological policy goals (say, restricting abortion access), but it may also make it easier to act in an autocratic fashion (e.g., through politically motivated prosecutions of political competitors). Specifically, the first-period incumbent chooses an institutional reform $i \in [0, 1]$. In the second period, the government chooses a policy $p \in [0, i]$ in the ordinary policy dimension. In addition, it takes a position on the authoritarian dimension $a \in [0, 1]$.⁵ In other words, the second period is defined by a two-dimensional “action space,” $A \equiv [0, i] \times [0, 1]$ in which the government chooses $(p, a) \in A$.

The second key modeling choice that allows us to capture citizen uncertainty is the assumption that there are two types of incumbents. The ideological incumbent is purely policy-motivated, and has no interest in authoritarian power actions. The ideal point of this incumbent is given by $X_I = (1, 0)$. In words, the ideologue favors policy 1, but has no desire to take in authoritarian actions.

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

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 *but also wants to move in an autocratic direction*. Putting it differently, the citizen understands the ideological position of the incumbent, but faces uncertainty over his authoritarian intentions, which is captured by the second dimension. The citizen's ex ante belief that the incumbent is a closet autocrat is given by $Pr(\text{Closet Autocrat}) = \alpha \in (0, 1)$. 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 $x_O = (0, 0)$. That is, the opposition is ideologically to the left of the incumbent and is not interested in authoritarian actions. Figure 1 provides an illustration.

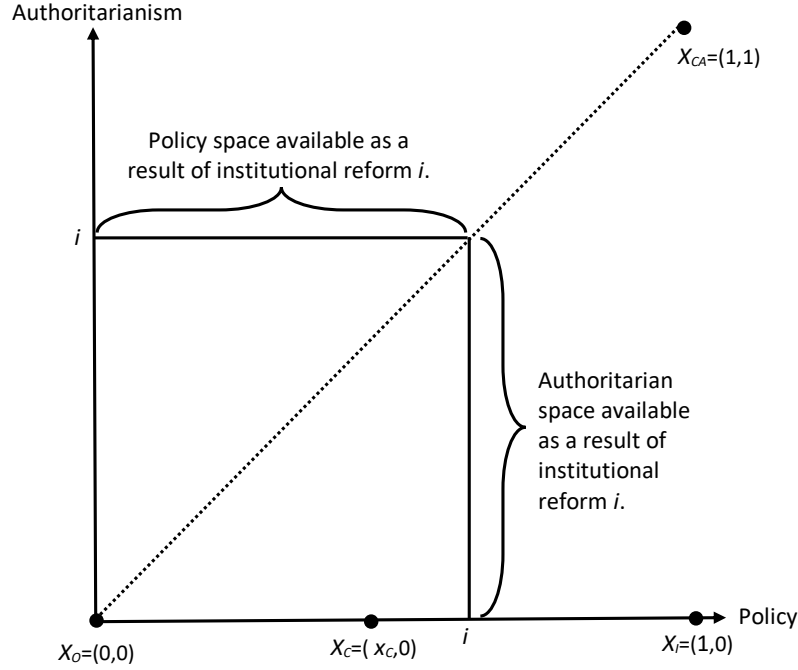


FIGURE 1. Determination of the policy space as a result of first

The citizen's ideal point is given by $X_C = (x_C, 0)$, where $x_C \in (0, 1)$. This implies that the citizen is purely policy-motivated: she favors institutional reforms in so far as they are instrumental to

her preferred policies *but she dislikes autocratic actions*.⁶ Specifically, we assume that the citizen's preferences over the government's second-period actions are given by

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

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* preferences: They prefer to publicly support policies that are close to their actual preferences 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 be sincere (politicians prefer to act in ways that are consistent with their convictions) or instrumental (building and maintaining a support base for the party requires some consistency between the party's rhetoric and actions). The Polish PiS and Hungarian FiDesz parties' behavior during the early 2000's provides an illustration: Instead

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.

⁷ This differentiates our approach from Svoboda's (2018) model in which citizens vote for the autocrat because the ideological profile of the opposition is not acceptable to them.

of compromising to make themselves politically viable, they stayed in opposition to preserve their anti-liberal brand.

Putting these elements together, the payoffs for the incumbent depend critically on whether he is reelected to a second term. 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 dimensions:

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

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

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:

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

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

where $\theta \geq 0$ is the weight the ideologue attaches 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 $j \in \{I, CA\}$.

- (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. Relation to other models of backsliding. Before presenting the equilibria of our model, it is useful to situate our approach in the context of other recent models of democratic backsliding. This rich, and still growing, literature has examined a number of distinct aspects of the phenomenon. An influential set of papers, including by Luo and Przeworski (2023), Helmke, Kroeger and Paine (2019), and Gratton and Lee (2020), focuses on incumbents' use of undemocratic reforms to "tilt the playing field" in an attempt to hold on to power. While this is clearly a plausible mechanism of backsliding, our model intentionally does not incorporate this feature: The first-period institutional reform does not manipulate the electoral process, and does not allow incumbents to boost their odds of holding on to power directly. By eliminating this channel in backsliding, this modeling choice allows us to isolate how citizen uncertainty over the intended purpose of institutional reforms influences democratic backsliding.⁸

Perhaps most closely related to our approach is the work by Grillo and Prato (2020), which also examines the gradual process of backsliding in the face of citizen uncertainty over incumbent intentions. As in our model, a distinctive element of Grillo and Prato's approach is that it can explain backsliding even when citizens are opposed to the incumbent. In their model, citizen

⁸ Note however that in our model, authoritarian action in the second period can be interpreted as "tilting," which may affect the likelihood of holding on to power in (not modeled) future periods.

evaluations of the incumbent depend on a reference point – loosely speaking, citizen expectations of how incumbents will perform. This reference point is endogenous to incumbent actions, including challenges to democratic norms. This gives rise to a logic in which incumbents challenge democratic norms and then “back off,” paradoxically increasing the incumbents’ popularity with voters because they are not as bad as might have seemed. As we will see, this logic is quite distinct from that of our model. In Grillo and Prato, incumbents openly engage in undemocratic moves, only to backpedal later. In contrast, our argument highlights the logic of the “wolf in sheep’s clothing:” Would-be autocrats exploit the maneuver room offered by citizen uncertainty over incumbent intentions to masquerade as a “regular” incumbent, hiding their authoritarian tendencies.⁹

2.2. Analysis. 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 government’s maneuver room in the second period. As such, there are 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 actions, which the citizen opposes. Put differently, the citizen may be sympathetic to reforms to the extent that these reforms will allow for ordinary policy choices she agrees with. But she may also be concerned that the same reforms remove checks on incumbents who reveal themselves to be closet autocrats. But the implications of reforms for second-period policy-making are only one consideration. A second significant aspect 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.

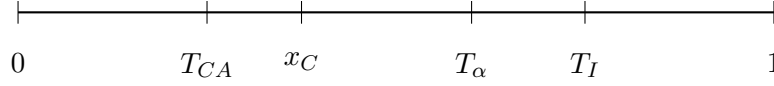
⁹ ? study a related but separate problem: The gradual expansion of executive power as a result of the interplay between executive overreach and the reluctance of the judiciary to engage in stringent enforcement of constitutional limits because judges are concerned to preserve executive discretion to react effectively in conditions of crisis.

As a result, the representative citizen's strategy is characterized by a threshold that makes the electoral decision contingent on the citizen's (updated) beliefs about the incumbent's type 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 to engage in 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. 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 constrains future authoritarian moves sufficiently. This threshold is given by $T_{CA} = \frac{2x_C}{\beta+1}$. Significantly – 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. Consequently, the 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.¹⁰

¹⁰ 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 on a “slippery slope” and that a closet autocrat will not be forced to act within the institutional constraints 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.¹¹ Intuitively, T_α is 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 actions 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 .)

Given the citizen's reelection strategy, we are now in a position to consider the incumbent's choice of institutional reform in the first period. The key issue confronting the incumbent is that the institutional reform he adopts has two separate effects: It defines the maneuver room the incumbent has in the second period if reelected, but it also affects the citizen's reelection decision. 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 ideal (from the incumbent's point-of-view). At the same time, there are expressive costs associated with governing in ways that are not in line with the incumbent's preferences. 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 reform that ensures reelection is so modest as to constrain

¹¹ 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.

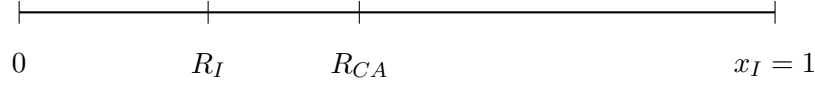
TABLE 1. Summary of thresholds

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.

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 also 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 for the incumbent to choose his ideal point instead. These thresholds are summarized in Table 1 (along with the thresholds imposed by the citizen). Figure 3 provides a graphical illustration, where R_{CA} denotes the threshold for the closet autocrat and R_I denotes the threshold for 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. This implies that 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



2.3. 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.¹² Here, it is useful to graph the location of these equilibria in the parameter space. We do so in Figure 4, 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 figure displays the institutional reform adopted by the ideological incumbent (i_I) and the closet autocrat (i_{CA}) in the first period.

After grouping the equilibria into three categories, we focus on the substantive implications. There are two separating equilibria (Area III and IV). 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. There are three pooling equilibria (Areas I, II and V). 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 equilibria (Areas I and II), the citizen nevertheless reelects the incumbent. In the third (Area V), the citizen replaces the incumbent with the opposition.

To discuss the substantive interpretation of these equilibria, and the implications for democratic backsliding, it is useful

Consider first the triangular area I in the southeastern part of the Figure. Here, the citizen is ex ante confident (as indicated by α close to 0) that the incumbent is likely to be an ideologue. Moreover, the citizen is ideologically distant from the opposition. As a result, the citizen's desire

¹² 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.

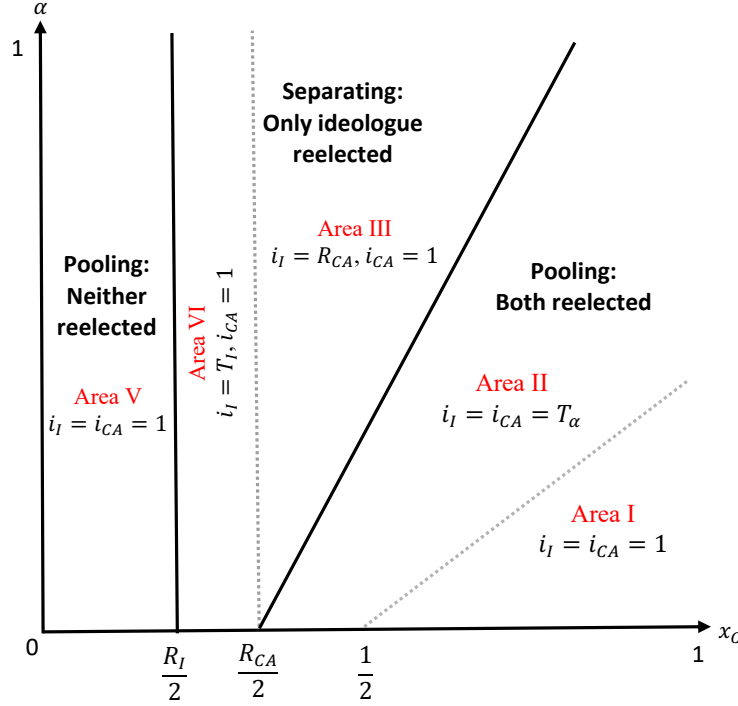


FIGURE 4. Separating and Pooling Equilibria as a function of Citizen's ideal point, x_c and Citizen's beliefs, α

to give the ideological incumbent sufficient policy leeway outweighs the concern to constrain a potential closet autocrat. The citizen is therefore willing to reelect the incumbent even if the first-period institutional reform is extreme (that is, $i_I = i_{CA} = 1$). Consequently both types of incumbents adopt that reform.

Now consider what happens if either the citizen is ideologically more moderate, or if the citizen's ex ante belief that the incumbent is a closet autocrat increases. This means that we move up or to the left in the figure into area II. 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. However, because the citizen's preferences are ideologically close to the ideologue, the “institutional assurance” that the citizen requires still provides sufficient maneuver room for the closet autocrat that he is willing to give

that assurance. Once again, the two incumbent types adopt the same institutional reform ($i = T_\alpha$), and the citizen reelects the incumbent.

As the citizen becomes ideologically moderate, we move into equilibrium areas III and IV. Because the citizen is ideologically centrist, satisfying her policy preferences no longer requires far-reaching institutional reforms. Instead, the risk of authoritarian backsliding enabled by far-reaching institutional reforms begins to loom large for the citizen. As a result, she will only re-elect the incumbent if 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 she only returns the ideologue to office. (Areas III and IV differ precisely in the precise first period reforms that the two types of incumbent adopt.)

Finally, consider equilibrium area V at the left side of the parameter space. Here, 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.

2.4. Substantive Implications. What are the substantive implications of the model for the phenomenon of democratic backsliding? We begin by highlighting the consequences of extending the logic established by Svobik (2020) to a dynamic setting in which citizens face uncertainty over the intentions of incumbents.

Implication 1. *As citizen preferences become sufficiently extreme, ideological incumbents and closet autocrats adopt the same institutional 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 (Areas I and II). As the citizen becomes ideologically more aligned with the incumbent, two dynamics begin to play out. The first overlaps with the logic of Svoboda (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 take authoritarian actions 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 (Svoboda 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 result is most similar to Grillo and Prato (2020), a model in which “opportunistic autocrats” can emerge as a result of citizens’ uncertainty about incumbent types and to Luo and Przeworski (2023), a model in which an incumbent can take successive small steps towards consolidating his power, leaving citizens unable to remove him when they realize that they are dealing with an autocrat. Critically, however, in these models incumbents stay in power in part because they have

the ability to “tilt the playing field” by directly rigging electoral processes. In contrast, in our model, the closet autocrat cannot directly help himself by tilting the playing field; he must still secure the support of the representative voter. Put differently, our model shows that the ability to manipulate electoral processes is not necessary for backsliding; the “wolf-in-sheep’s-clothing” mechanism we examine, which arises out of citizen uncertainty regarding an incumbent’s ultimate intentions, is sufficient.

Implication 2. *As the citizen is (a) less concerned about democratic backsliding (β approaches 1), 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. Such environments invite “wolves in sheep’s clothing.”

A final implication results from the fact that institutional reforms in the first period serve three distinct purposes: First, they can enlarge the maneuver room for “regular” policy-making in the second period. Second, 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. One consequence is that 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 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. In this sense, our model endogenously generates a kind of “constitutional culture.” The citizen knowingly supports an institutional environment that prevents policymakers from adopting policies 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 condition 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 (see implication 2).

3. AN APPLICATION: DEMOCRATIC BACKSLIDING IN POLAND

The key insight that animates our model is that democratic backsliding is a dynamic process in which closet autocrats – acting like a “wolf in sheep’s clothing” – use institutional reforms that appear plausibly democratic to lay the groundwork for future authoritarian actions. The resulting

uncertainty surrounding the intentions of incumbents puts citizens in a quandary that may lead to democratic backsliding even if citizens are fundamentally opposed to authoritarianism. Specifically, the model suggests an *interactive* effect between citizen preferences and their uncertainty about the intentions of incumbents: The degree of polarization required for backsliding depends on how certain citizens are about the incumbent's type. Moreover, with such uncertainty, backsliding can occur even if citizens would not knowingly vote for an authoritarian.

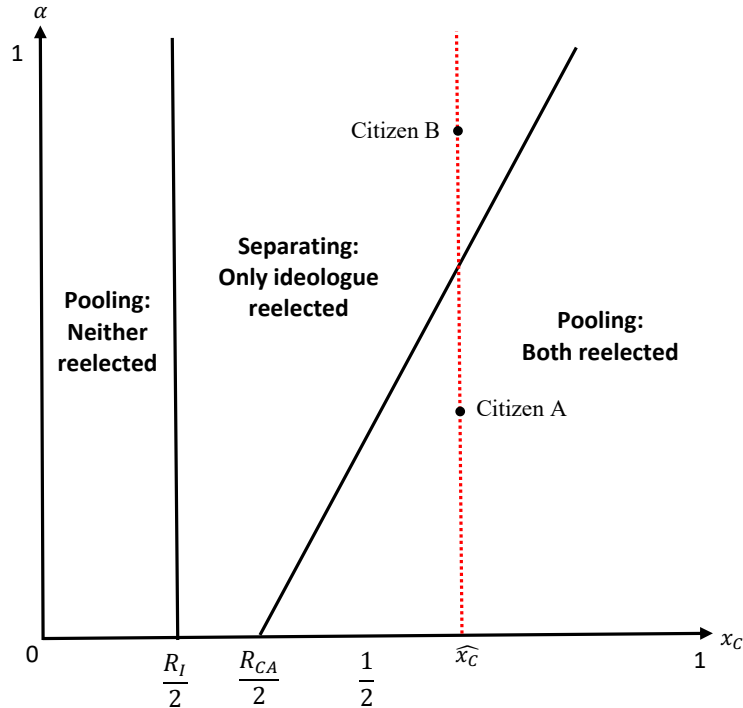


FIGURE 5. Interaction between ideology and uncertainty

We now turn to a survey experiment that explores this central feature of our model. To see the logic of our empirical approach, consider Figure 5. What differentiates the separating and pooling equilibria is how the citizen evaluates the institutional reform pursued by the incumbent: In the separating equilibria, the citizen believes that only a closet autocrat would pursue the reform in question. Given the belief that the reform is undemocratic, she refuses to re-elect. In contrast, in the pooling equilibria, she believes that the reform is not necessarily a sign that the incumbent is a closet autocrat: The citizen believes that the reform may well be democratic in the sense that the incumbent is an ideologue with no authoritarian intentions. As a result, she is willing to re-elect.

Thus, the difference between Citizen A (who re-elects following the reform in question) and Citizen B (who does not) is not ideology. Instead, they differ in how they evaluate the reform: Citizen B believes it is undemocratic and reveals that the incumbent is a closet autocrat while Citizen A is not so sure. In other words, the model suggests that citizen beliefs about the democratic or undemocratic nature of an incumbent’s institutional reform agenda are critical. Even citizens who are ideologically sympathetic to an incumbent are likely to reject the incumbent if they believe that institutional reforms the incumbent is pursuing are undemocratic:

Empirical Expectation. *Ceteris paribus, citizens are less likely to vote for a candidate as they become convinced that institutional reforms the candidate is pursuing are undemocratic. This is also the case for citizens who are ideologically proximate to the incumbent.*

We examine this expectation using an original survey experiment conducted in Poland in the spring of 2022 – a particularly relevant setting for studying the potential for democratic backsliding. Consistent with our model, we find that citizens who interpret institutional reforms undertaken by an incumbent as non-democratic are less likely to support the incumbent than citizens who are less certain of the motivations underlying the reforms. We show that this effect is robust to voters’ pre-treatment democratic commitments, as well as to their ideology. In other words, our experiment highlights that support for closet autocrats is not just driven by ideological proximity, but also by citizen uncertainty about the ultimate intentions of incumbents.

3.1. The context: Poland since 2015. After a long period in opposition, the Polish Law and Justice party (PiS) won an absolute majority of legislative seats in the October 2015 parliamentary elections, forming the first single-party majority cabinet in Poland since 1989. Almost immediately, the government undertook a series of controversial institutional reforms, targeting in particular the Polish judiciary (Duncan and Macy 2020). The first of these arose out of a dispute between the outgoing parliamentary majority and the newly-elected government regarding the appointment of justices to the Constitutional Tribunal. The PiS government passed legislation that forced the

inclusion of its newly appointed justices, 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 legislature. Over the following years, the PiS government embarked on systematic reform of the remaining parts of the judiciary. It changed the composition of the National Council of the Judiciary (*Krajowa Rada Sadownictwa*, 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 – a move that effectively forced out 40% of sitting Supreme Court judges. The qualifications 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. The only exception to mandatory retirement above 65 was receiving 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 – investigations that often targeted judges who objected to the PiS-led judicial reform (*Sedziowie pod Presja [Judges under Pressure]* 2019).

3.2. Ideologue or wolf in sheep’s clothing? Condemnation of PiS’ reform efforts was swift and significant. Many of the reforms were greeted with 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 violated 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 judiciary would carry enormous payoffs for a closet autocrat, providing the government with significant resources to erect barriers for political opponents. For example, such control makes it easier to engage in political prosecutions, a common tactic in post-Soviet states.¹³ The Supreme Court also plays a central role in Poland’s public system of financing electoral campaigns. Because parties are only eligible for reimbursements of campaign expenditures *after* the party’s finances have been declared “in order” by the Supreme Court, a PiS-controlled Supreme Court could potentially be used to reduce financial support for opposition parties. Finally, control of the courts more generally reduces the risk of judicial interference in the government’s decisions.

While opponents of the reforms were alarmed by the perceived danger of creeping authoritarianism, PiS 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,¹⁴ 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).

¹³ In fact, the PiS government has considered putting Donald Tusk, the leader of the most influential opposition party, Civic Platform (PO), on trial for treason. According to the Polish constitution, such a trial would take place before the State Tribunal (*Trybunał Stanu*), presided over by the chief justice of the Supreme Court, a position now controlled by PiS.

¹⁴ 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.

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. This uncertainty was compounded when in 2020, the PiS-packed Constitutional Court issued a decision significantly restricting access to abortion on constitutional grounds. The fact that this policy change – which had been a central plank of PiS’ electoral commitments – was accomplished at the hands of the Supreme Court appeared to rationalize PiS’s reform efforts in the eyes of some ideologically motivated voters.

4. SURVEY AND EXPERIMENTAL EVIDENCE

In short, contemporary Poland offers a natural context for exploring the logic of our argument. The Polish government has engaged in reform efforts that have given rise to fears about democratic backsliding – and yet, there remains significant uncertainty, especially among voters who are ideologically proximate to PiS, regarding the government’s ultimate intentions. The aim of our survey experiment, conducted using a representative sample of 3,000 Polish voters in the spring of 2022, is to examine the empirical expectation implied by our model: Holding constant ideological preferences, voters become less likely to vote for an incumbent engaged in institutional reforms as voters become more convinced that the reforms are motivated by undemocratic intentions.¹⁵

In constructing our survey, there is a significant complication we must confront. As is intuitive – and previous literature has established (Şaşmaz, Yagci and Ziblatt 2022) – partisan bias may influence how voters perceive particular reforms; those who are ideologically more sympathetic to an incumbent are more likely to give the incumbent the benefit of the doubt than those voters who are ideologically opposed. For example, PiS supporters may interpret a reform proposal as democratic if

¹⁵ Of our online sample, constructed with the help of the Center for Public Opinion Research (CBOS), one of Poland’s most reputable pollsters, 60 % indicated interest in voting in the upcoming elections. The specifics of the survey are described in more detail in the Appendix B.

they associate it with PiS and as undemocratic if they associate it with a party they are ideologically opposed to. To isolate the effect of citizen uncertainty about the intentions behind a reform, we therefore require a measure of how a voter interprets the intentions behind an institutional reform that is *independent* of the voter’s ideological leanings and partisan preferences. At the same time, the reform needs to be plausible and resemble a real world scenario. To address this challenge, we designed the survey to obtain a voter’s evaluation of a generic, yet realistic, institutional reform devoid of partisan context *before* conducting our survey experiment. Specifically, respondents were presented with the following vignette:

“Electoral reforms have been initiated in several countries. They consist of introducing single-member districts, meaning that only one candidate can be elected from a given constituency - the one who receives the most votes. As a result, on the scale of the whole country, candidates supported by large parties have greater chances of being elected than candidates of small parties. Some believe that this solution strengthens democracy because it enables large parties to govern effectively. Others believe that this solution weakens democracy by making it difficult for smaller parties to participate in governance.”

Critically, such a reform was never proposed by PiS, the potential “closet autocrat” in the Polish context. Using a reform actually proposed by PiS would increase the risk of partisan bias in evaluations. Moreover, the proposal has the two key properties that can induce citizen uncertainty. On the one hand, it benefits the party that is capable of securing a majority of seats in the legislature. On the other hand, it can be defended as an effort to make government more efficient and accountable. Following the vignette, respondents were asked if the proposed institutional reform was, in their opinion, democratic (“Reforms will make governments more accountable and effective”) or undemocratic (“Reforms are an undemocratic way to weaken the opposition and allow the rulers to stay in power”). We asked respondents to place themselves on a scale running from 1 (“disagree completely”) to 10 (“completely agree”). The resulting measure of a respondent’s perception of the reform provides us with an assessment of how a respondent evaluates the democratic nature

of an ambiguous reform that is independent of partisan influences. Reassuringly for our purposes, respondents are indeed divided in the assessment of the reform, as shown in Figure ?? in Appendix B, supporting the plausibility of uncertainty over the incumbent’s intentions behind this proposal.

4.1. The Conjoint Experiment. Our conjoint experiment follows the protocols established in the literature (Graham and Svolik 2020; Svolik 2020; Graham et al. 2022). Respondents in a control and a treatment group are asked to choose between two hypothetical candidates. Each candidate is characterized by an ideological position and a proposal for an institutional reform. In the control group, the two candidates differ ideologically, but propose the same institutional reform. Thus, the support for each of the candidates in the control provides us with a baseline of the extent to which the choice between candidates can be attributed to the the difference in ideology as perceived by the respondent. In the treatment group, respondents are presented with candidates with the same ideological positions as in the control. But in contrast to the control, the more conservative candidate proposes an arguably undemocratic reform. Again, respondents are asked to choose between the two candidates. Tables 2 (control) and 3 (treatment) provide the candidate profiles presented to respondents.

Given random assignment, comparing support for the candidates across the treatment and control groups provides us with an estimate of how much voters are willing to punish undemocratic maneuvers. Moreover, since the candidates carry no party labels, any difference in respondent choice is driven by the institutional reform proposed and not partisan attachments. This willingness by citizens to punish incumbents for undemocratic behavior is the focus of analysis by Svolik and Graham, and we find comparable levels of punishment.¹⁶ Critically, evaluating our argument requires more. The key issue is whether this willingness to punish is a function of voter uncertainty about the intentions behind an incumbent’s institutional reforms. We expect that respondents who believe that reforms are not motivated by authoritarian ambitions will be less likely to punish a candidate (corresponding to Citizen A in Figure 5, the pooling equilibrium) than respondents who

¹⁶ See Figure ?? in the Appendix.

Candidate A	Candidate B
[profession],[age]	[profession], [age]
Wants to strengthen relations with the European Union and limit the role of the Catholic Church in deciding about abortion rights	Is in favor of upholding Christian values, also in public life, and wants greater guarantees of Polish sovereignty within the European Union.
Wants to popularize participation in elections by opening polling stations both on Saturday and Sunday of the election weekend (currently the presidential and the Sejm and Senate elections are held only on Sundays)	Wants to popularize participation in elections by opening polling stations both on Saturday and Sunday of the election weekend (currently the presidential and the Sejm and Senate elections are held only on Sundays).

TABLE 2. Conjoint Experiment Table: Control Group

Candidate A	Candidate B
[profession],[age]	[profession], [age]
Wants to strengthen relations with the European Union and limit the role of the Catholic Church in deciding about abortion rights	Is in favor of upholding Christian values, also in public life, and wants greater guarantees of Polish sovereignty within the European Union.
Wants to popularize participation in elections by opening polling stations both on Saturday and Sunday of the election weekend (currently the presidential and the Sejm and Senate elections are held only on Sundays)	Wants the government to be able to extend the terms of the Sejm and Senate for a year or more in situations of exceptional state emergency and to postpone the election date without consulting other state authorities. ¹⁸

TABLE 3. Conjoint Experiment Table: Treatment Group.

interpret the reform as undemocratic (corresponding to Citizen B in Figure 5, the separating equilibrium). To evaluate these expectations, we divide the respondents in our treatment and control groups into four different groups based on their interpretation of the reform vignette offered earlier in the survey. Specifically, we create four sub-groups, separating respondents into those who saw the reform as largely democratic up to those who perceived it as highly undemocratic, i.e., we divide respondents into four groups based on how “suspicious” those within a group are regarding the potentially undemocratic intentions behind an institutional reform. In Table ?? in the Appendix, we show that random assignment resulted in covariate balance across these sub-groups across the treatment and control groups.¹⁷

¹⁷ Specifically, the ‘low’ group is comprised of those who responded that their belief about the reform being undemocratic is 1-3, ‘mid-low’ those who responded 4-5, ‘mid-high’ 6-8, and ‘high’ 9-10. We chose to group respondents into four categories in order to maintain sufficient power.

Reform interpreted as undemocratic	% Choosing B		Difference	% Change
	Control	Treatment		
Low agreement	0.45	0.40	-0.05	-0.13
Mid-low agreement	0.35	0.29	-0.06	-0.18
Mid-high agreement	0.32	0.22	-0.10	-0.31
High agreement	0.26	0.16	-0.10	-0.41

TABLE 4. Changes in support for candidate B for introducing an undemocratic reform, depending on interpretation of the vignette reform.

We are now in position to proceed to the analysis. Table 4 reports the fraction of respondents who choose Candidate B (ideologically, the more PiS-like candidate) across the treatment and control within each sub-group (defined by interpretation of the reform). Consistent with our expectations, as respondents view the vignette reform as more undemocratic, the less likely they are to support the PiS-like candidate in response to the treatment (i.e., comparing across control and treatment group). Among those who view the reform as largely democratic (the first group), support for Candidate B drops from 45% to 40% (about a 13% reduction in support). In contrast, among those in high agreement that the reform is undemocratic, support drops from 26% to 16%, a reduction of approximately 40%. These results clearly suggest that willingness to punish depends on the level of certainty that citizens have with respect to the intentions behind institutional reforms.

A different method for approaching these data is to regress the decision to vote for Candidate B on the interaction of treatment and vignette reform interpretation, thus estimating the conditional average marginal component effect of introducing the undemocratic reform (Hainmueller, Hopkins and Yamamoto 2014). Specifically, we condition on the interpretation of the vignette reform, which is our measure of uncertainty. The effects for each ‘reform interpretation’ group are reported in Figure 6, along with 95% confidence intervals. Respondents who did not interpret the vignette reform as undemocratic (low and mid-low) are insensitive to the treatment: introducing a candidate B who is prone to undemocratic moves does not produce significant differences relative to the purely ideological control. On the other hand, those voters who interpreted the vignette reform as undemocratic (mid-high and high) resolve the trade-off between ideology and democracy in a significantly different way: they are 10-12% less likely to vote for Candidate B.

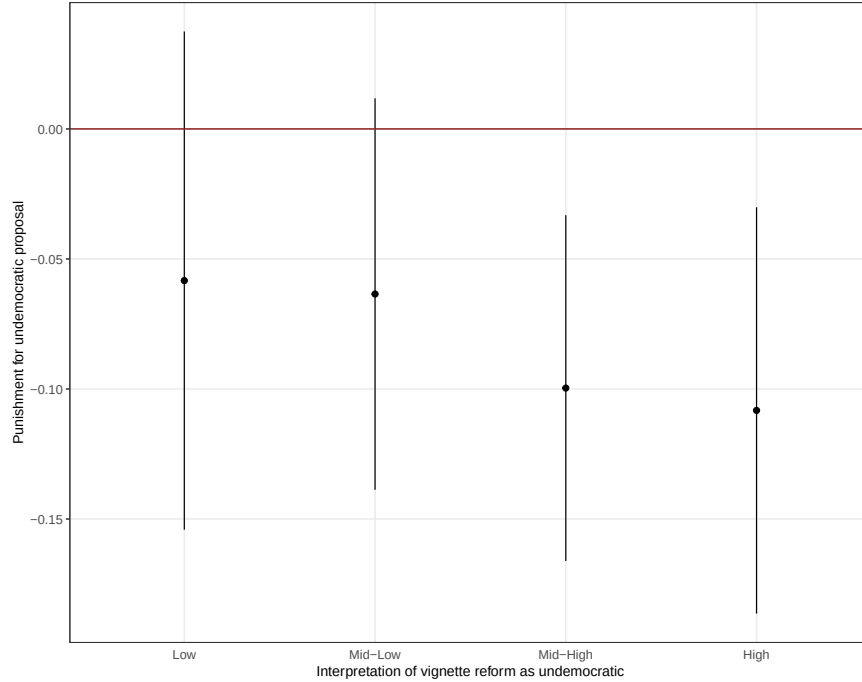


FIGURE 6. Conditional ACME, with 95% confidence intervals.

4.2. Correlational Analysis. The advantage of the conjoint experiment we have just presented is that it confronts respondents with an actual choice between concrete candidates. The obvious downside is directly connected: It is a choice that is disconnected from the political context in Poland. To strengthen confidence in the external validity of our results, we therefore turn to a correlational analysis to examine whether respondents' support for PiS (the potential closet autocrat in the Polish context) is conditioned by how respondents interpret our vignette reform. Specifically, we want to know whether, controlling for ideology as well as other relevant covariates, Polish citizens are less likely to vote for PiS as they become more suspicious that the vignette reform is motivated by undemocratic aims. To do this, we draw on our survey, which asked respondents a battery of standard questions including self-reported ideology and voting intention.

We estimate a logit model of the probability that a respondent intends to vote for PiS as a function of demographic variables, a number of covariates related to support for PiS, and the *interaction* between the respondent's ideology and his evaluation of the vignette reform:

TABLE 5. Predictors of voting for PiS if elections were Sunday; logistic regressions.

	<i>Dependent variable:</i>		
	Would vote for PiS if elections were Sunday		
	(1)	(2)	(3)
Undemocratic interpretation	−0.145*** (0.029)	−0.129 (0.104)	−0.143 (0.109)
Self-placement in left-right spectrum	0.986*** (0.062)	0.970*** (0.131)	0.988*** (0.140)
Undemocratic interpretation x left-right		−0.003 (0.019)	−0.0004 (0.020)
Expressed Preferences for Democracy	−0.250*** (0.089)	−0.236*** (0.088)	−0.250*** (0.089)
Age	0.023*** (0.006)	0.023*** (0.006)	0.023*** (0.006)
Education	−0.189*** (0.029)	−0.178*** (0.029)	−0.189*** (0.029)
Town Size	−0.038 (0.037)	−0.039 (0.036)	−0.038 (0.037)
Unemployed	0.115 (0.175)	0.090 (0.172)	0.115 (0.175)
Dissatisfaction with one's condition	−0.393** (0.185)	−0.407** (0.188)	−0.393** (0.185)
Gender	0.218 (0.151)	0.214 (0.149)	0.218 (0.151)
Religiosity	0.335*** (0.060)	0.342*** (0.059)	0.334*** (0.061)
Interest in Politics	0.143 (0.090)	0.139 (0.088)	0.143 (0.090)
Province FE	yes	no	yes
Observations	1,879	1,879	1,879

Note:

*p<0.1; **p<0.05; ***p<0.01

$$y_i = \sigma(\alpha + \beta \text{ Ideological Self-Placement}_i + \gamma \text{ Reform Interpretation}_i + \delta \text{ Reform Interpretation}_i \times \text{Ideological Self-Placement}_i + \eta X_i + \theta_w + \epsilon_i)$$

where y_i is a dummy variable indicating whether respondent i intends to vote for PiS if elections were held on Sunday; the ‘Reform Interpretation’ variable is our measure of how the voter interprets the reform vignette; θ_w are fixed effects for Poland’s administrative units; and X_i are standard control variables.¹⁹ We report the results in Table 5. Our primary interest is in the interactive effect of ideology and uncertainty regarding the aims of reform efforts. To show this interactive effect, Figure 7 plots the predicted probabilities of voting for PiS as a function of the respondent’s assessment that the reform is undemocratic, focusing on the most right-leaning voters in our survey (those self-placing in categories 5,6, and 7 on the left-right scale and, consequently, ideologically closest to PiS). The results provide clear evidence consistent with our expectations: Even among voters who are ideologically close to PiS, the probability of voting for this party decreases with scepticism regarding incumbents’ reform intentions. As a voter at a given ideological position interprets the vignette reform as more undemocratic, she is less willing to cast her vote for PiS. For example, for those who place themselves at the right end of the ideological spectrum, the probability of voting for PiS decreases from 84% among those who believe the reform is largely democratic to 58% among those who regard it as largely undemocratic.

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 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 have examined

¹⁹ These include age, gender, education, size of the town of residence, employment status, religiosity, dissatisfaction with one’s condition, and interest in politics. We also solicit voters’ preferences for democracy using standard questions on democratic values from the World Values Survey and construct a z-score index. Details on how the index was put together and robustness checks can be found in Appendix B.

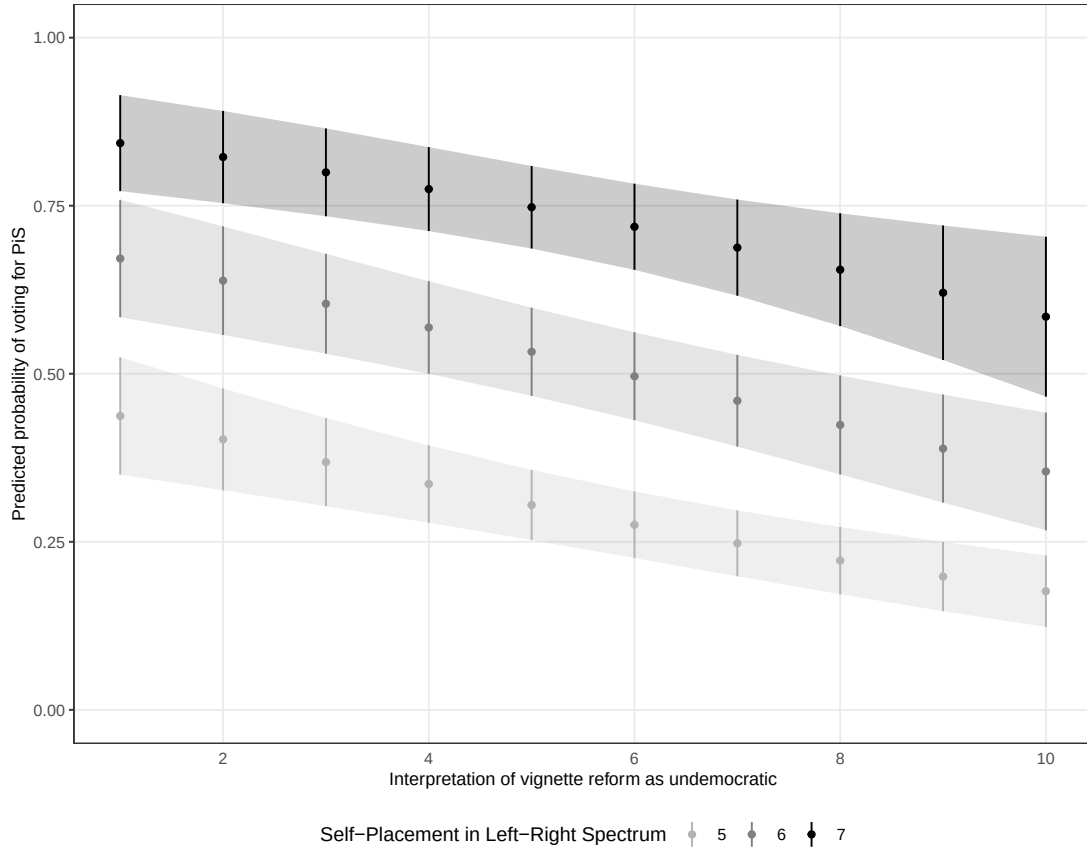


FIGURE 7. Predicted Probabilities of PiS Support for the three kinds of typical right-wing voters: very conservative (self-place at 7), moderately conservative (self-place at 6) and right of center (self-place at 5). The probabilities are calculated based on the model in column (2) of Table 5, setting the control variables at their median values.

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 2023).

Put differently, the argument we have presented demonstrates that increasing polarization matters for two distinct reasons. The first – an insight originally developed by Svoboda (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 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

- 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.
- Graham, Matthew, Austin Jang, Filip Milacic and Milan Svolik. 2022. "Commitment to Democracy: Revealed-Preference Evidence from Seven European countries."
- Graham, Matthew H and Milan W Svolik. 2020. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *American Political Science Review* 114(2):392–409.
- Gratton, Gabriele and Barton E Lee. 2020. "Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies."
- Grillo, Edoardo and Carlo Prato. 2020. "Reference points and democratic backsliding." *American Journal of Political Science* .
- Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.
- 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.
- Levitsky, Steven and Daniel Ziblatt. 2018. *How democracies die*. Crown.
- Lex Super Omnia Association of Prosecutors to the Prosecutor General*. 2019.
- Luo, Zhaotian and Adam Przeworski. 2023. "Democracy and Its Vulnerabilities: Dynamics of Democratic Backsliding." *Quarterly Journal of Political Science* .
- 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. 2019a. “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. 2019b. “The Polish legislation concerning the lowering of the retirement age of judges of the Supreme Court is contrary to EU law.”.
- Revised Statute of the European Commission for Democracy through Law*. 2002. Resolution(2002)3 adopted by the Committee of Ministers.
- Şaşmaz, Aytuğ, Alper H Yagci and Daniel Ziblatt. 2022. “How Voters Respond to Presidential Assaults on Checks and Balances: Evidence from a Survey Experiment in Turkey.” *Comparative Political Studies* p. 00104140211066216.
- 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.
- Voeten, Erik. 2016. “Are people really turning away from democracy?” *Available at SSRN 2882878*