Political Science Research and Methods (2026), 1–21
doi:10.1017/psrm.2026.10109

ORIGINAL ARTICLE

Backsliding under contested oversight

Robert Gulotty and Zhaotian Luo

Political Science, University of Chicago, Chicago, Illinois, USA
Corresponding author: Robert Gulotty; Email: gulotty@uchicago.edu

(Received 6 December 2024 UTC; revised 15 July 2025 UTC; accepted 24 August 2025 UTC)

Abstract
Citizens do not directly observe democratic backsliding and, as a result, may hesitate to respond to subversion. We develop a model of third-party oversight bodies, such as the media or courts, that detect and assess actions that may be subversive to democracy and inform citizens. Oversight deters subversion, disciplines incumbents, and enables corrective actions by providing credible information about ambiguous incumbent behavior to citizens. However, when the oversight body is contested, citizens may doubt the intent behind its criticisms. When the oversight body is cautious in its criticisms, it elicits negative inferences about its intentions, what we term a fake news effect. The consequences are severe, undermining oversight and enabling backsliding. Democratic accountability depends on reliable sources of information and elected officials' commitment to upholding norms of conduct.

Keywords: democratic backsliding; electoral accountability; oversight

Democracy depends on oversight by the news media, courts, and investigators to evaluate the actions of politicians and persuade citizens to act (Weingast and Moran, 1983; McCubbins and Schwartz, 1984; Ferejohn and Shipan, 1990; Zaller, 2003). However, oversight bodies can become the subject of political contestation. In the 1998 Venezuelan elections, public warnings about Hugo Chávez's continued antidemocratic intentions were countered by accusations that the media was mere cover for the opposition (Lander and Maya, 1999; Kronick *et al.*, 2023). Following his historic victory in the 2001 Thai general election, Thaksin Shinawatra accused the oversight body constitutionally charged with policing corruption of political opportunism, turning an "innocent mistake" in asset reporting into a witch-hunt (Barnes, 2000; Cumming-Bruce, 2001). In the United States Presidential Election of 1972, the public was told that Congress and the press were overreacting to a "third-rate burglary" at Watergate. In each case, citizens had to judge the backsliding candidate's behavior and act on the word of an oversight body that was itself contested.

In the last of these cases, citizens were able to hold backsliding incumbents to account. In the others, oversight is rendered ineffective. We argue that this occurs because not only are citizens uncertain about the incumbents' subversive actions, but that uncertainty is compounded by the strategic behavior of the oversight body. In the face of public skepticism about the intentions of an oversight body, it must walk a fine line between being too cavalier to be persuasive and being so cautious as to turn any criticism into evidence of malicious intent. In sum, we find that in its efforts to inform citizens, an oversight body can end up undermining its own credibility.

Much of the literature on democratic backsliding assumes that either citizens directly observe politician behavior (Svolik, 2020) or that the oversight body is non-strategic

© The Author(s), 2026. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

Table 1. Kinds of backsliding

Kind of backsliding	Cause of failure	Effect of criticism	Examples
Stealth	Non-detection	Removal	Nixon
Excusable	Fake news effect	None	Thaksin, Chávez

(Li, Raiha and Shotts, 2022). We find that weakening these assumptions has significant consequences for the prospects of democracy. Democratic backsliding is much more difficult to avoid if citizens are uncertain about both information quality and the intentions of the oversight body. If the oversight body can be malicious, intent on criticism or witch-hunts, citizens must discern whether criticism is evidence of subversion or instead produced by such a malicious oversight body. That is, when oversight is contested, voters also use the presence of criticism to determine whether the oversight body is malicious, raising a dual inference problem, creating a difficult strategic dilemma for oversight bodies.

Our formal analysis begins with a model of electoral accountability in which a representative citizen may remove the backsliding incumbent, creating what Grillo *et al.* (2024) term a vertical restraint. Following Svolik (2020), we assume society is polarized, so accountability depends on persuading citizens to forgo policy benefits to hold the incumbent to account.[1] Building on this baseline, we introduce uncertainty about whether a controversy amounts to backsliding. Even if citizens hear of a controversy, they may still need to determine whether there are sufficient grounds to act. This is made difficult when controversial policy choices or rhetoric are ambiguous, either subversive and worthy of removal or non-subversive and excusable (Stokes, 2025).

Our model emphasizes two ways that backsliding may not be obvious to the citizen, displayed in Table 1. First is because of a failure to detect, what we term "backsliding by stealth." A power grab may be unremarked upon and corruption can go undetected, allowing incumbents to harm democratic institutions and citizen interests (Levitsky and Ziblatt, 2019; Luo and Przeworski, 2023). Second is backsliding under the cover of an excuse. That is, citizens must not only hear criticism and be alerted to suspicious behavior but also believe that accusations are accurate. That is a difficult task, not only because institutional degradation can be difficult to evaluate but also because citizens also face uncertainty about the intention of the oversight body that issues the criticism. Citizens hear criticism but are not moved by "fake news."

In our analysis, we distinguish cases of effective and ineffective oversight by its effects on incumbent behavior, voter learning, and the type of backsliding that occurs. We find that the effects of uncertainty over the incumbent's action and uncertainty regarding the intention of the oversight body are endogenous to the ability of the oversight body to induce removal. In an equilibrium with effective oversight, any criticism guarantees removal. This fully deters actions that can be mistaken for subversion—the incumbent is endogenously disciplined. As a result, the task of the oversight body is simple: rather than resolving ambiguity, it can assume that any ambiguous action merits removal and take any opportunity for criticism. From the perspective of the citizen, the type of oversight body does not matter. Both types behave identically: there is no reputational problem for the oversight body if incumbents always avoid behavior that can be confused for backsliding. The only risk is that the incumbent takes a gamble on escaping detection and backsliding by stealth.

When oversight is ineffective its warnings are completely ignored. Free from the risk of removal, the incumbent is undisciplined, taking actions that can be mistaken for subversion. The cause of this breakdown originates in the dual demands on oversight: it must decide whether an action is subversive and it must act so as to ensure that its criticisms are believed. The need to be believed drives a wedge in the behavior between the two types of oversight body, which creates the opportunity for new inferences on the part of the citizen. The non-malicious oversight body can affect the degree

[1]Here polarization is interpreted as a preference for the incumbent versus a challenger. For the distinction between affective and ideological polarization, see Iyengar *et al.* (2012); Broockman *et al.* (2023).

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

of citizen learning by requiring stronger evidence of subversion before issuing criticism, what we term the caution effect. Caution has a downside, however, because when the oversight body becomes cautious, from the citizenry's perspective, criticism is proportionally more likely to originate from a malicious oversight body, what we term the fake news effect. In combination, the caution and fake news effects turn the oversight body against itself: if it attempts to adjust to avoid raising false alarms, it makes any criticism sound like a witch-hunt.

Our analysis complements existing research on the role of citizen uncertainty in enabling democratic backsliding (Miller, 2021; Chiopris *et al.*, 2025; Gratton and Lee, 2024). Li, Raiha and Shotts (2022), for example, study the role of media oversight in democratic backsliding and its consequences for candidate selection and incumbent effort. That analysis assumes the media is not strategic and faces no uncertainty over the incumbent's actions. We find that the level of uncertainty that the oversight body has about the incumbent's actions is a crucial determinant of its effectiveness. Miller (2021) considers a model of backsliding in which citizens cannot observe democratic subversion and must rely on the opposition to sound an alarm, but then citizens receive a second a signal of whether subversion occurred. As with our model, when citizens see a challenge, they are uncertain whether the incumbent did act to subvert democracy or if, instead, the opposition chose to cry wolf. Unlike our setup, Miller (2021) assumes voters have an exogenously determined level of uncertainty about whether subversion had occurred. In our setting, even small amounts of uncertainty are magnified by strategic dynamics to enable backsliding.

The main results of the model are widely applicable to settings with strategic oversight. Gibilisco and Horz (2024) study law enforcement agencies who must simultaneously choose to inspect criminal activity and inform the public about their efforts. That inspection game with endogenous reporting has multiple equilibria, complicating the interpretation of misreporting. In our setting, criticism is adjusted to improve the persuasive impact of the criticism. However, we show that it is the strategic oversight body's interest in being persuasive that ends up undercutting its message.

1. Knowledge, accountability, and fire alarm oversight

Democracies rely on publicity, that is, public knowledge of government decision-making, to help hold politicians to account (Dahl, 2005). This information enables vertical accountability, between the elected officials and voters, which is centered in formal models of democratic erosion. For instance, Svolik's model of democratic backsliding assumes that pre-election manipulation is observed by some voters, who in turn, take the incumbent's manipulation into account when deciding how to vote (Svolik, 2020). However, in practice, such direct knowledge becomes difficult, if not impossible, to obtain as the distance between the elected and the electors grows and as the complexity of government rises (Bobbio, 1987, p. 83). As Levitsky and Ziblatt put it, "Democracy's erosion is, for many, almost imperceptible" (2019, p. 6). Instead, voters must rely on third-party oversight bodies to detect and disseminate concerns about backsliding.

In the following, we argue uncertainty is a common constraint on vertical accountability and a typical feature of democratic backsliding. We then discuss the way that a number of institutions common across democracies serve as oversight bodies, including the news media and the courts, to help citizens hold politicians accountable.

1.1. Citizen uncertainty over democratic backsliding

Scholars of accountability have long questioned whether citizens have the capacity or information needed to hold politicians to account. In many cases, citizens may not have the expertise or personal experience to judge incumbent behavior, particularly in complex policy domains.[2] The literature on economic voting suggests that voters use market outcomes as a noisy proxy of candidate quality or

[2]See Ashworth and Bueno de Mesquita (2014).

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

effect. Even then, accountability is possible insofar as there is "clarity of responsibility" to help voters identify the politicians responsible for economic performance.[3] In the case of democratic subversion, it is often only clear which politicians benefit from backsliding after the backsliding has occurred and when it is too late to act.

In the case of backsliding, often voters must infer both incumbent behavior and the underlying intent of that behavior. For example, facing a likely defeat in the 2022 election in Brazil, then-President Jair Bolsonaro pushed the Brazilian legislature to amend the constitution to adopt paper ballots.[4] News reporting on these efforts alleged that they were designed to set the ground for Bolsonaro to challenge the results of the upcoming election (Milhorance and Londoño, 2021). For Bolsonaro's supporters, paper ballots improve trust in the election, not undermine it. In this case, the legislative vote failed, but Bolsonaro's supporters were not persuaded; they would storm the capital less than two years later.

Uncertainty over intent does not only arise from public ignorance. Even experts who share the same conceptualization of democracy may disagree about the interpretation of a case of democratic subversion (Waldner and Lust, 2018). For example, Little and Meng (2024) argue that expert coders in the Varieties of Democracy project face significant difficulty in reaching consensus about even narrow features of democracy. Among the five to ten coders for each country-year, the average standard deviation across coders asked the same question ranged from 20% to 25% of the scale. This is in part because the same behavior that "deepens" democracy in one political context could, in others, enable its undoing. If trained experts are unable to apply definitions in a consistent way, we might expect citizens to face similar difficulties in addressing democratic subversion.

1.2. Horizontal accountability and oversight

Many democratic systems rely on expert bodies to detect and adjudicate subversion. The formal literature identifies such bodies as providing "fire alarm oversight": a third party that monitors and decides whether to sound the alarm (Epstein and O'Halloran, 1995). Such third parties can include the news media, specialized prosecutors, review boards, or committees. For instance, in 2017, South Africa's Public Protector, a quasi-judicial administrator, compiled a report exposing the corruption of then-President Jacob Zuma. This report prompted judicial action and eventually led to impeachment proceedings by the national legislature (Ginsburg and Huq, 2018). Similarly, Richard Nixon's impeachment proceedings began with reporting by *The Washington Post* and the Bofors scandal in India, which led to the defeat of Rajiv Gandhi's Indian National Congress in 1989, was brought to light by coverage in *The Hindu* newspaper (Hazarika, 1989).

In each case, the oversight body must decide whether to advance a criticism of the incumbent. To avoid this, backsliding leaders may preemptively shut down criticism, such as in the case of Turkey's President Recep Erdoğan, who initiated criminal prosecution of journalists, blocked websites, and restrained the judiciary (Bermeo, 2016). We see similar behavior in Venezuela, Poland, Russia, and Thailand, where politicians and elites sought to undermine competing sources of information that would otherwise threaten their hold on power (Little, 2012; 2015; Gehlbach and Sonin, 2014; Shadmehr and Bernhardt, 2015). Even when these actions were only threatened, they can have a chilling effect on criticism (Stanig, 2015).

In contrast to research on outright interference, our analysis proceeds by introducing adverse selection of the oversight body. In particular, we allow for the possibility for the oversight body to be malicious, a condition we refer to as contested oversight. Reasons for this malicious behavior may include partisanship, corruption, or profit motives. Existing research on oversight bodies argues

[3] See Lewis-Beck (1988); Powell and Whitten (1993); Anderson (1995).

[4] Preceding the vote, Bolsonaro held a 10-minute tank parade past the national congressional building, allegedly to formally invite himself to an annual naval training exercise (Boadle, 2021b). For some, the result was a scandal, labeled "real, clear and unconstitutional intimidation" (Boadle, 2021a).

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

that they often aim to satisfy an audience's preferences, which may include a preference for accuracy (Gentzkow and Shapiro, 2006). The extent of these strategic incentives may vary because of external factors, such as having a media market structure that allows an outlet to reach a new audience, or the extent of career concerns on the part of individual overseers (Andina-Díaz and García-Martínez, 2020; Larreguy *et al.*, 2020).

Whatever the source, imperfections, inaccuracy, or bias on the part of oversight bodies are widely expected to have strong negative consequences on political accountability. For instance, Li, Raiha and Shotts (2022) model malice and error on the part of the media in a fragile democracy, finding that voters who expect false accusations can fail to sanction subversion. In a similar vein we allow voters to infer the type of oversight body on the basis of its behavior. Nonetheless, we find that uncertainty over the type of oversight body need not affect accountability. Whether it does or not is endogenous to the effect of its criticisms. When criticisms from the oversight body can induce removal, its type is irrelevant. However, when oversight fails, even minor doubts over the intent of criticisms are exacerbated by the interplay of strategic caution and a fake news effect, enabling democratic backsliding.

2. Model

The players are an *incumbent* (referred to as "he"), a representative *citizen* (referred to as "she"), and an *oversight body* (referred to as "it"). The model has two periods divided by an election in which the citizen chooses whether to *remove* the incumbent, $r \in \{0, 1\}$. Who holds office in period 2 depends on the citizen's choice: if $r = 0$, the incumbent stays in office, while if $r = 1$, a *challenger* becomes the office holder. In period 2, the players receive their payoffs depending on the choices made in period 1 and during the election.

In period 1, the incumbent takes one of three possible actions, $a \in \{n, 0, 1\}$. The first, $a = n$ is a "*clean action*" that removes ambiguity about potential subversion. Otherwise the action is ambiguous: $a = 1$ is a "*subversive action*" that undermines democracy; $a = 0$ is a "*non-subversive action*," in which the incumbent does not attempt to undermine democracy, but neither does he exert any extra effort to avoid ambiguity.

We assume the clean action $a = n$ and the non-subversive action $a = 0$ is always available. However, opportunities to undermine democracy only occur with probability $\pi \in (0, 1)$. Let $\theta \in \{0, 1\}$ represent whether such a subversive action is available, so that $\Pr(\theta = 1) = \pi$. If $\theta = 1$, the incumbent has an opportunity to undermine democracy by taking the subversive action. Whether such an opportunity arises or not, θ, is private information to the incumbent.

The ambiguous actions are distinguished by their effect on democratic institutions. That is, subversion has negative post-election consequences for the citizen.[5] Alternatively, the cost of subversion can come from the direct value that citizens place on democratic institutions. Democracy-minded citizens may act to punish efforts to remove checks on unilateral action or bias electoral procedures to advantage the incumbent. By contrast, while the non-subversive action may raise suspicion, we assume such actions are neither corrupt nor do they impose additional consequence for democratic institutions after the election.

Democratic backsliding occurs when the incumbent takes the subversive action, $a = 1$, and is subsequently retained by the citizen, $r = 0$. In other words, the citizen can prevent democratic backsliding by removing the incumbent from office.

The incumbent gets the payoff of 1 by holding office. In addition, he benefits from democratic backsliding. If the incumbent takes the subversive action in period 1, $a = 1$, being the office holder in period 2 provides him with an extra benefit of $b > 0$. The clean action, $a = n$, requires effort

[5]This can be micro-founded by assuming that under democratic backsliding, in period 2 the incumbent can take an additional action that benefits himself at the expense of the citizen—the benefit the incumbent receives is b while the expense on the citizen is κ. This setup is widely applied in models of democratic backsliding, see Li, Raiha and Shotts (2022).

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

Table 2. Payoffs of the incumbent and citizen

(Incumbent, citizen)	$r = 0$	$r = 1$
$a = n$	$1 - \epsilon, x$	$-\epsilon, y$
$a = 0$	$1, x$	$0, y$
$a = 1$	$1 + b, x - \kappa$	$0, y$

by the incumbent to avoid ambiguity, imposing a disutility $\epsilon > 0$, which we treat as infinitesimal.[6] By contrast, we assume the non-subversive action, $a = 0$, brings no intrinsic benefit or cost. The idea that incumbents may be disciplined, taking effort to resolve ambiguity or losing out on opportunities for modest advantage, is akin to what Levitsky and Ziblatt (2019) term institutional forbearance. However, as originally conceived, institutional forbearance merely requires that incumbents avoid actions that "obviously violate" the spirit of the law. Here we allow the clean action to have a further effect, guaranteeing no accusations of backsliding, closer to the sort of forbearance characterized by Helmke *et al.* (2022).

The citizen cares about who sits in office. She gets the payoff of $x > 0$ if the incumbent holds office, which measures the extent to which the incumbent *appeals* to the citizen in terms of policy or valence. Similarly, the citizen gets the expected payoff of $y > 0$ if the challenger takes power. The difference in these payoffs, $x - y$, measures *political polarization*. In addition, the citizen pays a cost for retaining a backsliding incumbent, $\kappa > 0$. This can be because she places non-instrumental value on democratic institutions or because she faces some other consequence of democratic backsliding by having this incumbent in office in the future (Nalepa and Pop-Eleches, 2022).[7]

The payoffs of the incumbent and of the citizen are summarized in Table 2.

In sum, the problem of oversight is to ensure that citizens can use elections to remove subversive incumbents.

The citizen is uninformed. She does not know whether the incumbent has gotten any opportunity to undermine democracy, θ, nor the action he has taken, a. The only information she has comes from the oversight body. After the incumbent has taken an *ambiguous action*—either the subversive action $a = 1$ or the non-subversive one $a = 0$, the oversight body detects it with a known probability $q \in (0, 1)$, what we term the capability of the oversight body.

Just like the citizen, the oversight body does not directly observe the action of the incumbent. That is, we assume that the oversight body cannot distinguish the clean action from its own failure to detect a controversial action. If no ambiguous action is detected, it can be that the incumbent took pains to avoid it, or that the incumbent got lucky in that it was not detected. Let $d \in \{0, 1\}$ represent whether the oversight body detects an ambiguous action, then

$$\Pr(d = 1|a) = \begin{cases} q, & a \in \{0, 1\} \\ 0, & a = n \end{cases}.$$

However, after detecting an ambiguous action, the oversight body investigates whether the action is subversive or not. We model the outcome of that investigation as a signal s, drawn from a conditional density f_a that has full support on $\mathbb{R}$. We assume throughout that the densities f_0 and f_1 satisfy

[6] This assumption can be relaxed so long as the disutility is positive and sufficiently low in comparison to the payoff of holding office.

[7] For instance, the consequence could be that in the second period the citizen becomes less able to remove the incumbent and, therefore, less able to select a potentially more appealing candidate into office, as in Luo and Przeworski (2023). The model can be extended to incorporate this consequence of backsliding. In particular, suppose in period 2, there is a probability $\phi \in (0, 1)$ that a more appealing challenger who brings the citizen the payoff of $y' > y$ is available for the citizen to select into office; however, if the citizen retains a backsliding incumbent in period 1, she loses the ability to remove the incumbent in period 2. The stake for the citizen to retain a backsliding incumbent in this extension, therefore, is $\phi y'$, which provides a micro-foundation for κ. All results of the benchmark model hold in this extension.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

the *monotone likelihood ratio property*:

$$\ell(s) := \frac{f_1(s)}{f_0(s)}$$

is strictly increasing in s.[8] A larger s, therefore, represents a stronger piece of evidence that the detected action is subversive, $a = 1$, rather than non-subversive, $a = 0$.

Finally, after detection and investigation, the oversight body chooses whether to *criticize* the incumbent for seeking to undermine democracy, $c \in \{0, 1\}$. That is, we assume that criticism against the incumbent can only occur after the detection, $d = 1$, of an ambiguous action: it is ruled out by either the incumbent taking the clean action $a = n$ or the failure to detect an ambiguous action. The oversight body detects, investigates, and interprets controversy, it cannot invent it out of whole cloth.

At the election stage, the citizen first observes whether the oversight body has issued criticism, $c \in \{0, 1\}$, then decides whether to retain or remove the incumbent, $r \in \{0, 1\}$.

The oversight body is privately informed about its type, $m \in \{0, 1\}$, drawn from a common prior that $\Pr(m = 1) = \mu \in (0, 1)$. Type $m = 0$ seeks to criticize the incumbent for taking the subversive action. This type of oversight body is characterized by a weight $z \in (0, 1)$ on false alarms versus neglected subversion. The oversight body loses z for inaccurately criticizing the incumbent that did not engage in subversion, a false alarm, and loses $(1-z)$ for failing to criticize an incumbent that took a subversive action, what we term neglected subversion. Given detection of a controversial action d, and a particular signal s, the oversight criticizes the incumbent if and only if

$$\Pr(a = 1 | d = 1,\ s) > z,$$

that is, if and only if it believes the incumbent has committed a subversive action with a probability that exceeds the threshold z.

In contrast, type $m = 1$ is *malicious*, a behavioral type which always wants to criticize the incumbent, engaging in a "witch-hunt," regardless of whether the incumbent has acted to subvert democracy or not.[9] We say oversight is contested when the probability of this type is positive, $\mu > 0$.

The timing of the game is summarized in Figure 1. In period 1, the incumbent first observes whether there is an opportunity to undermine democracy, θ, and chooses an action $a \in \{n, 0, 1\}$. Then, the oversight body learns its type m and if an ambiguous action is detected, $d = 1$, it observes signal s and chooses whether to criticize the incumbent, $c \in \{0, 1\}$. At the election stage, the citizen observes whether the oversight body criticizes the incumbent, c, and chooses whether to remove the incumbent, $r \in \{0, 1\}$. In period 2, the action that the incumbent has taken, a, is revealed and all the players receive their payoffs. The solution concept is *perfect Bayesian equilibrium* ("equilibrium").

2.1. Formal definition of equilibrium

The behavioral strategies of the incumbent, oversight body, and citizen are, respectively, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1) \in \Delta(\{n, 0, 1\})^2$, $\sigma : \mathbb{R} \to [0, 1]$, and $\boldsymbol{\rho} = (\rho_0, \rho_1) \in [0, 1]^2$, where

[8] To ease exposition, we assume that ℓ is differentiable, $\lim_{s\to-\infty} \ell(s) = 0$, and $\lim_{s\to\infty} \ell(s) = \infty$. Also, for both $a \in \{0, 1\}$, we use F_a to denote the distribution function of f_a. An example is that $s = a + \sigma\epsilon$, where ϵ is drawn from the standard normal distribution and $\sigma > 0$ inversely measures the precision of signal s. In this example,

$$f_a(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-a)^2}{2\sigma^2}}$$
$$\ell(s) = e^{\frac{2s-1}{2\sigma^2}}.$$

[9] We obtain qualitatively similar results if the malicious oversight body is strategic but favors criticism.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

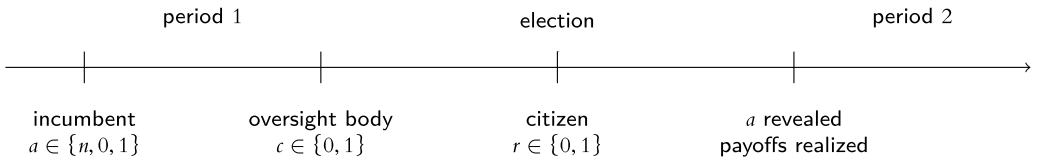


Figure 1. Timing.

- $\alpha_\theta(a) = \Pr(a|\theta)$ is the probability the incumbent takes action $a \in \{n, 0, 1\}$ conditional on whether the ambiguous action would undermine democracy, θ;[10]
- $\sigma(s) = \Pr(c = 1|s, d = 1, m = 0)$ is the probability the non-malicious oversight body criticizes the incumbent after detecting an ambiguous action, $d = 1$, and observing signal s;
- $\rho_c = \Pr(r = 1|c)$ is the probability the citizen removes the incumbent conditional on criticism from the oversight body, c.

A strategy profile $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$ constitutes an equilibrium if it is sequentially rational under a consistent belief system that consists of beliefs formed through Bayesian updating given $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$.

Definition 1. An equilibrium $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$ has effective oversight if $\rho_0^* \neq \rho_1^*$. An equilibrium has ineffective oversight if it does not have effective oversight.

Our definition of effective oversight is minimal: it is effective if criticism of the incumbent has any consequence for the citizen's decision of whether to remove the incumbent. Oversight is only ineffective if the citizen completely disregards the presence or absence of criticism by the oversight body during the election.

2.2. Preliminary analysis

While both established democracies and electoral autocracies can experience subversion, recent interest in democratic backsliding is driven by the breakdown of democracy, rather than changes in autocratic institutions (Little and Meng, 2024). Our analysis begins with assumptions that make oversight necessary to sustain democratic institutions. Specifically, we focus on cases that satisfy the following assumption.

Assumption 1. *The incumbent is more appealing to the citizen than the challenger, $x > y$. The citizen is willing to remove the incumbent to prevent democratic backsliding, $\kappa > x - y$. The opportunity to undermine democracy is rare, $\pi < (x - y)/\kappa$.*

The first two conditions in Assumption 1 ensure the citizen's ideal election decision depends on incumbent behavior. On the one hand, if the incumbent is less appealing to the citizen than the challenger, $x \leq y$, the citizen always removes the incumbent in the election. This trivially prevents backsliding. On the other hand, if effects of subversion on democracy are insufficient relative to polarization, so that $\kappa \leq x-y$, the citizen would always retain the incumbent. In this case, backsliding does not have to be done by stealth, it happens openly. Only when the citizen prefers the incumbent more than the challenger, $x > y$, but places a sufficiently high value on democracy, $\kappa > x - y$, is information about whether the incumbent is trying to undermine democracy crucial for her reelection decision. In this case, the citizen would retain the incumbent if she were certain that the incumbent has not chosen to subvert democracy, while she would remove the incumbent despite polarization if

[10]Because the subversive action $a = 1$ is available only when $\theta = 1$, $\alpha_0(1) = 0$ always holds.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

she knows for sure that the incumbent has taken subversive actions. As a result, the incumbent can engage in democratic backsliding, but he must do so by stealth, concealing his true action from the public.

The last condition in Assumption 1 guarantees that if the citizen was fully uninformed about the incumbent's action, democratic backsliding would happen at every opportunity. Under this condition, the uninformed citizen believes opportunities to undermine democracy are sufficiently rare that even if the incumbent were to be fully undeterred from subversion, the citizen's prior belief would lead her to retain the incumbent. In this case, the oversight body plays an important role in preventing backsliding, using its criticism to warn the citizen that the incumbent has gotten that rare opportunity and taken the subversive action.

3. Oversight effectiveness

In the following, we characterize how oversight works when it is effective. We show that in these cases, politicians have an incentive to take costly actions to reassure voters, as there is an ever-present risk of criticism and removal. Even if the oversight body may be malicious, subversion can be detected and deterred because the citizen is confident that criticisms are warranted regardless of their source.

3.1. The benefit of the clean action

Crucial to our analysis is the assumption that politicians always have the option of preventing ambiguity from rising to the attention of the oversight body by taking the clean action but must pay a price to do so. The following lemma, which is a result of Assumption 1, explains why incumbents might be willing to pay this price.

Lemma 1. *In any equilibrium* $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$*, the citizen retains the incumbent if the incumbent receives no criticism from the oversight body,* $\rho_0^* = 0$*. If oversight is effective, criticism generates a positive possibility of removal,* $\rho_1^* > 0$*.*

Lemma 1 indicates that criticism is a prerequisite for removal. Intuitively, the only reason for the citizen to remove the incumbent is to prevent backsliding. By Assumption 1, the opportunity to undermine democracy is rare, $\pi < (x - y)/\kappa$ so that the uninformed citizen prefers to reelect the incumbent. Knowing that the oversight body chose not to criticize the incumbent would only provide additional reassurance that the incumbent has not taken the subversive action.

Because the citizen never removes the incumbent absent the oversight body's criticism and because the oversight body can only criticize him based on ambiguous actions, the incumbent is always able to guarantee reelection by taking the clean action. However, doing so comes at two costs, first, taking the clean action may require the incumbent to forgo an opportunity for subversion, and second, it always comes with a cost ϵ, representing the difficulty of clarifying ambiguity or avoiding it in the first place.

By contrast, taking any ambiguous action always generates a risk of getting criticized by the malicious oversight body. Hence, when oversight is effective, the incumbent risks losing office by taking either of the two ambiguous actions—there is a chance that his action gets detected, a chance that this action is then criticized by the oversight body, which, because oversight is effective, could influence the citizen to remove him in the election.

3.2. Inexcusable backsliding under effective oversight

Given Lemma 1, oversight is effective in equilibria in which there is any positive probability of criticism, $\rho_1^* > 0$. The following proposition characterizes such an equilibrium and its existence:

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

Proposition 1. *There exists an equilibrium* $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$ *with effective oversight and in each such equilibrium:*

1. $\alpha_0^*(n) = 1$, *while* $\alpha_1^*(n) = 1$ *if* $q > b/(1+b)$ *and* $\alpha_1^*(1) = 1$ *if* $q \leq b/(1+b)$;
2. $\sigma^*(s) = 1$ *for all* $s \in \mathbb{R}$;
3. $\rho_1^* = 1$.

We find that when oversight is effective, it removes uncertainty about the nature of incumbent actions. Still, the incumbent may find the risk of removal worth bearing for the additional benefits of subversion. However, because the incumbent's cost for the clean action, ϵ, is trivial in comparison to the value of office, the risk of removal is sufficient to fully deter ambiguous, non-subversive actions that may be mistaken for democratic backsliding. That is, given that the citizen may remove the incumbent when he is criticized by the oversight body, the incumbent never takes the non-subversive action; he always takes the clean action when there is no opportunity to undermine democracy, $a = n$ when $\theta = 0$; and when such an opportunity arises, he chooses either the subversive action or the clean action, $a \in \{n, 1\}$ when $\theta = 1$, depending on the capability of the oversight body, q.

Here the two types of oversight body behave identically, endogenously removing the need to distinguish between types. That is, subject to detection, either type of oversight body will issue criticism with certainty. Knowing that the incumbent has taken the subversive action, both types of oversight body prefer to criticize him, either to warn the citizen about backsliding or simply to enjoy witch-hunting. The citizen remains uncertain about the type of oversight body, but that uncertainty does not matter because there is no uncertainty about the incumbent's action.

When both types of oversight body behave identically, it also endogenously removes the ability for the citizen to update her beliefs about its reputation. In the absence of such reputational concerns, criticisms are fully persuasive: whenever the citizen observes that oversight body criticizes the incumbent, she immediately infers that the incumbent must have been caught taking an ambiguous action and the action must be subversive, $\Pr(a = 1 | c = 1) = 1$. Hence, to prevent democratic backsliding, the citizen would always remove the incumbent after criticism by the oversight body, $\Pr(r = 1 | c = 1) = 1$.

Svolik (2013) suggests that democratic backsliding occurs due to citizens' inability to monitor the actions of incumbents. We find that this is not the case in equilibria with effective oversight. Citizen uncertainty alone does not necessarily weaken deterrence. Instead the extent of deterrence depends on the detection capability of the oversight body, q, relative to the incumbent's incentive to engage in backsliding, b. Whenever the oversight body is sufficiently capable of detecting the ambiguous action, $q > b/(1+b)$, the incumbent is fully deterred and disciplined. If the oversight body is not capable enough, $q \leq b/(1+b)$, the incumbent may take subversive action if he thinks he won't get caught. When the incumbent does get caught, however, he would be removed from office and backsliding would fail.

4. Ineffective oversight

We have established that effective oversight always exists, now the question is whether there are equilibria in which oversight is ineffective. Suppose an equilibrium with ineffective oversight exists. In such an equilibrium the citizen disregards the oversight body's criticism and, under Assumption 1, retains the incumbent. As a result, the incumbent never bothers exerting any effort to avoid ambiguity, no matter how trivial the cost of that effort. Facing no repercussions for doing so, he opts for subversion whenever the opportunity arises. We formally characterize such equilibria in the following proposition.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

Proposition 2. *In any equilibrium with ineffective oversight* $(\boldsymbol{\alpha}^*, \boldsymbol{\rho}^*, \sigma^*)$,

1. $\alpha_0^*(0) = \alpha_1^*(1) = 1$;
2. *for each* $s \in \mathbb{R}$,

$$\sigma^*(s) = \begin{cases} 1, & s > s^*(z) \\ 0, & s \leq s^*(z) \end{cases},$$

where

$$s^*(z) := \ell^{-1}\left(\frac{1-\pi}{\pi}\frac{z}{1-z}\right);$$

3. $\rho_0^* = \rho_1^* = 0$.

In addition to the incumbent's choices discussed above, the second part of Proposition 2 indicates that the non-malicious oversight body follows a cutoff strategy in deciding whether to criticize the incumbent. Unlike the case when oversight is effective, here the oversight body faces uncertainty after detecting an ambiguous action. Given that the incumbent is completely undeterred, when an opportunity arises, $\theta = 1$, the action is subversive, if no opportunity arises, $\theta = 0$, the action is non-subversive. Hence, in contrast to the case of equilibrium with effective oversight, what action was taken remains ambiguous. Therefore, the oversight body must rely on its investigation, represented by signal s, to make an inference. Recall that a larger s indicates a stronger piece of evidence that the detected action is subversive. As shown in Proposition 2, the oversight body would criticize the incumbent after detecting an ambiguous action if the evidence that the action is subversive is sufficiently strong; and it would remain silent if the evidence is weak.

The effect of criticism on the citizen's beliefs depends on a combination of the citizen's learning about two unknown quantities: the incumbent's action and the oversight body's type. Recall, the citizen who observes criticism faces ambiguity about whether the incumbent has taken the subversive action. It is possible that instead, the incumbent engaged in non-subversive actions that are incorrectly labeled subversion. Under such ambiguity, the citizen is willing to remove the incumbent if and only if

$$\Pr(a = 1|c = 1) > \frac{x-y}{\kappa},$$

that is, if and only if the oversight body's criticism convinces the citizen that it is sufficiently likely that the incumbent has taken the subversive action. It must overcome polarization $(x - y)$ net of the costs for the citizen from democratic backsliding (κ).

As in the case of effective oversight, the citizen is uncertain about the oversight body's type m—particularly whether the criticism is made by the malicious type that benefits from witch-hunting. However, now that uncertainty matters. By the law of total probability,

$$\Pr(a = 1|c = 1) = \sum_{m=0,1} \Pr(m|c = 1)\Pr(a = 1|c = 1, m).$$

Hence, the citizen's belief about whether the incumbent has taken the subversive action can be affected by her belief about the oversight body's type.

In the following, we unpack the conditions that render oversight ineffective. We first show how political polarization, a stronger preference for the incumbent, can discourage criticism from the oversight body, which, in turn, makes criticism that does occur more informative. This *caution effect* demonstrates why ambiguity over the incumbent's action is insufficient to make oversight ineffective. We then show how uncertainty over the oversight body's type enables a fake news effect, as criticism has consequences for its reputation. The combination of the two effects drives a non-monotonicity in beliefs about subversion after the oversight body has criticized the incumbent, the net effect being to lower the degree of polarization necessary to make oversight ineffective.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

4.1. The caution effect

Assume that the citizen is certain that the oversight body is not malicious and is only uncertain about the incumbent's action. In such cases, the ability of the oversight body to persuasively warn the citizen about subversion only reflects its criteria for evidence, which itself is driven by its trade-off between false alarms and neglected subversion, z. This posterior is

$$\Pr(a = 1|c = 1, m = 0) = \overline{a}(z),$$

where

$$\overline{a}(z) := \frac{\frac{\pi}{1-\pi}\frac{1-F_1(s^*(z))}{1-F_0(s^*(z))}}{1 + \frac{\pi}{1-\pi}\frac{1-F_1(s^*(z))}{1-F_0(s^*(z))}}.$$

This term incorporates standard Bayesian updating on the part of the citizen, combining information about the prior belief that subversion could occur, π, and assumptions about the informativeness of signals to the oversight body, F_a. As we might expect, the equilibrium effect of increasing weight on avoiding false alarms on the part of the oversight body is to increase the citizen's confidence that criticism implies that subversion has occurred:

Lemma 2. *$\overline{a}(z)$ is strictly increasing in z and for all $z \in (0, 1)$, $\overline{a}(z) > z$.*

Lemma 2 generates the "caution effect": the greater the preference for avoiding false alarms, the higher the bar is set for evidence of subversion, the more persuasive is any particular criticism. In practice, oversight bodies use robust editorial standards and prosecutorial discretion to ensure criticism is informative. In this setting, caution is only necessary when oversight is ineffective, as it is only under those conditions that ambiguous actions occur.

This benchmark, which has no contestation, produces the same democracy-policy trade-off as Svolik (2020). While not all acts of subversion are detected, the citizen knows that whatever persuades the oversight body to issue criticism is more than sufficient for the citizen to believe both that subversion has occurred and that it merits removal. Moreover, because $\overline{a}(z) > z$, as $\mu \downarrow 0$, there is no equilibrium with ineffective oversight as long as $z \geq (x - y)/\kappa$. That is, so long as the oversight body is as concerned with avoiding false alarms as the citizen, the inability to observe the incumbent's action alone is insufficient to render oversight ineffective.

4.2. The fake news effect

It is natural to expect that the challenge of oversight is greater when the citizen faces uncertainty over the type of oversight body. We will consider oversight to be contested when the citizen believes that there is a positive probability that the oversight body is malicious. Surprisingly, any possibility of malicious oversight is sufficient to render oversight ineffective, given sufficient caution. In what we term the "fake news effect," strategic caution on the part of non-malicious oversight has an adverse effect on its own reputation. This is because as the non-malicious type of oversight body becomes cautious, criticism is more likely to have arisen from a malicious oversight body.

Allowing the oversight body to take on one of two types, malicious and non-malicious, and given that in equilibria with ineffective oversight the two types differ in their behavior, the citizen will use criticism to update her belief that the oversight body is malicious to:

$$\Pr(m = 1|c = 1) = \hat{m}_\mu(z),$$

where

$$\hat{m}_\mu(z) := \frac{\mu}{(1-\mu)\left(1-(1-\pi)F_0\left(s^*(z)\right)-\pi F_1\left(s^*(z)\right)\right)+\mu}.$$

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

The numerator follows from the fact that the malicious oversight body always issues criticism. The first term in the denominator combines the propensity for subversive opportunities, π and distribution of the signal, which in turn depends on the standard used by the oversight body.

Analyzing this posterior belief, we find that the oversight body's weight on avoiding false alarms is positively associated with the citizen inferring that it is malicious, what we term the "fake news effect":

Lemma 3. *$\hat{m}_\mu(z)$ is strictly increasing in z and $\lim_{z\to 1} \hat{m}_\mu(z) = 1$.*

That is, the more cautious the non-malicious oversight body, the higher z, the more negative the citizen's inference is about its type. The logic follows along the same lines as the caution effect: the non-malicious oversight body internalizes the citizen's democracy-policy trade-off and so is more cautious in issuing criticism. The malicious oversight body does not internalize that trade-off and only benefits from issuing criticism. The citizen then makes an inference about the source given the signal.

In summary, in equilibrium with ineffective oversight the two types of oversight body act differently, and there is updating on its type. This induces non-malicious oversight to become more cautious than the malicious oversight body. However, any criticism that is issued makes the citizen more suspicious that the oversight body is malicious. In the next section, we examine how these two processes affect deterrence or removal of subversive incumbents.

4.3. Excusable backsliding under ineffective oversight

While researchers of democratic backsliding and political ideology are justifiably concerned about polarization (see Broockman *et al.* (2023)), Lemma 3 gives reasons to suspect that the problem is significantly deepened by the combined uncertainty over the incumbent's actions and the oversight body's motivations. The caution effect and the fake news effect combine to enable democratic backsliding.

To show this, we analyze the citizen's posterior probability that subversion has occurred given criticism. By the law of total probability, the posterior belief of subversion can be decomposed into a weighted sum of learning from informative criticism originating from a non-malicious type of oversight body and uninformative criticism from the malicious type:

$$\begin{aligned}\Pr(a=1|c=1) &= \sum_{m\in\{0,1\}} \Pr(m|c=1)\Pr(a=1|c=1,m)\\ &= \left(1-\hat{m}_\mu(z)\right)\overline{a}(z) + \hat{m}_\mu(z)\pi\\ &=: \hat{a}_\mu(z).\end{aligned}$$

Because of the caution effect, the preference to avoid false alarms on the one hand increases citizen confidence in the validity of criticism; on the other hand, because of the fake news effect, that same preference leads the citizen to draw negative inferences about the type of oversight body. The following lemma describes the resulting non-monotonicity:

Lemma 4. *There exists a unique $\hat{z}_\mu \in (0,1)$ such that:*

1. *$\hat{a}_\mu(z)$ is maximized at $z = \hat{z}_\mu$ and $\hat{a}_\mu(\hat{z}_\mu) = \hat{z}_\mu$;*
2. *for $z < \hat{z}_\mu$, $\hat{a}_\mu(z) > z$ and is strictly increasing in z;*
3. *for $z > \hat{z}_\mu$, $\hat{a}_\mu(z) < z$ and is strictly decreasing in z.*

Moreover, $\hat{z}_\mu$ is strictly decreasing in μ, $\hat{z}_0 = 1$, and $\hat{z}_1 = \pi$.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

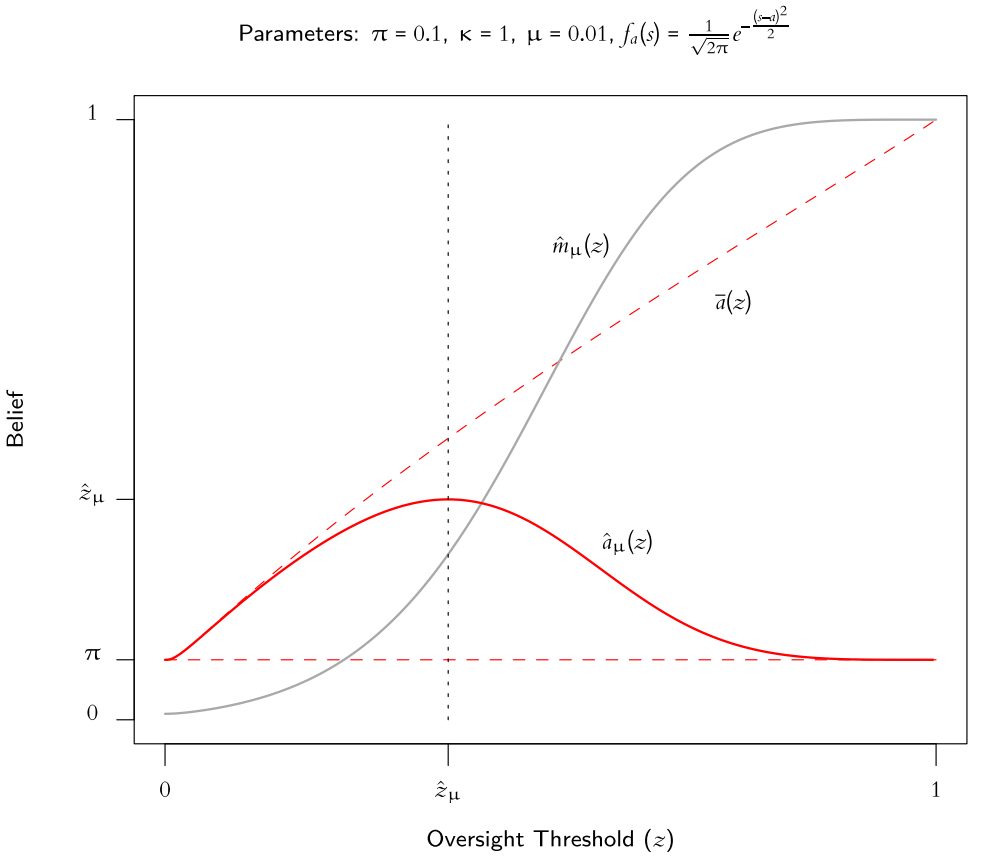


Figure 2. Beliefs regarding democratic subversion and the type of oversight body.

When the oversight body has a lower z, it is more likely to issue criticism, and that criticism, by the caution effect, will be discounted. When the oversight body is instead focused on avoiding false alarms, the fake news effect dominates. This transition is displayed in Figure 2 under particular values of π, κ, and μ and a normal distribution of signals. The x-axis represents the oversight body's weight on false alarms, z, or equivalently, its threshold on signals of democratic subversion. The y-axis displays the value of posterior beliefs for the oversight body's type, in solid gray, and the action taken by the incumbent following criticism, in solid red, decomposed into the contribution from non-malicious and malicious types as dashed lines. The point of transition, $\hat{z}_\mu$, marked by the vertical dotted line, occurs where $\hat{a}_\mu(z)$ is maximized.

From Figure 2, we can see that the oversight body can, at minimum, announce a controversy. That is, if the non-malicious oversight body places no weight on avoiding false alarms, such that z close to 0, the two types behave the same way, and all the citizen can learn is that an ambiguous action was taken. In that case, the citizen is left with her prior, π. Moving from left to right, initially, the oversight body's interest in avoiding false alarms increases the informativeness of criticism, but also begins raising the citizen's suspicion that the oversight body is malicious. The latter overcomes the former at $\hat{z}_\mu$. As z approaches 1, all criticisms become evidence that the oversight body is malicious, again leaving the citizen with no additional information over her prior.

The forces mentioned above worsen the trade-off between policy and democracy that was identified in Svolik (2020). In situations where the citizen only faces uncertainty about the incumbent's actions, oversight is guaranteed to be effective so long as polarization is less than or equal to κ.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

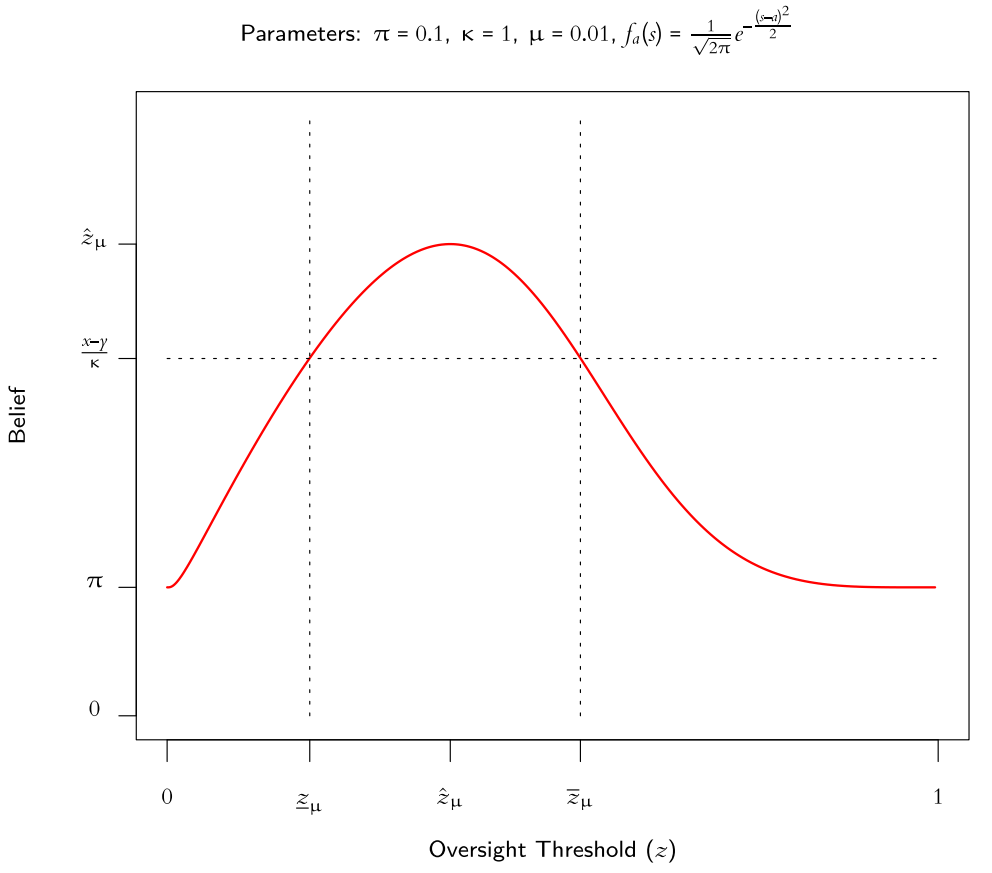


Figure 3. Equilibria with ineffective oversight.

However, that result is contingent on the assumption that the information about subversion comes from an uncontested source. The following proposition characterizes the combined effect of uncertainty and strategic mediation by the oversight body:

Proposition 3. *No equilibrium with ineffective oversight exists if and only if* $x - y < \hat{z}_\mu \kappa$ *and* $z \in (\underline{z}_\mu, \overline{z}_\mu)$, *where* $0 < \underline{z}_\mu < \overline{z}_\mu < 1$.

Proposition 3 shows that when oversight is contested, avoiding backsliding becomes more difficult than in the Svolik benchmark. We can see the relationship between polarization and effective oversight in Figure 3. Here we again display the preference of the oversight body on the x-axis and the citizen's belief that there has been democratic subversion following criticism on the y-axis. Following Proposition 3, in the region between $(\underline{z}_\mu, \overline{z}_\mu)$ oversight is effective in all possible equilibria, and recalling Proposition 1, in such equilibria, the two types of oversight body are behaviorally identical.

It is possible for polarization to be high enough that ineffective equilibria are always possible. Recall that the citizen's preference is to remove the incumbent so long as the posterior belief in subversion exceeds $(x-y)/\kappa$, indicated as a horizontal dashed line in Figure 3. The highest that polarization can be while still ruling out ineffective oversight occurs at $(x-y)/\kappa = \hat{z}_\mu$, as this is when the oversight body chooses the same threshold as the citizen. According to Proposition 3, if polarization exceeds this level $x-y > \hat{z}_\mu \kappa$, even when the oversight body selects a threshold most aligned with the citizen,

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

it is insufficient to rule out ineffective equilibria. As polarization declines, the range of oversight body preferences that can rule out ineffective equilibria expands

5. Illustrating effective and ineffective oversight

Above we have outlined two "ideal types" for oversight in the face of democratic subversion. In the following, we illustrate how these equilibria operate in two cases, the lead-up and consequence of the Watergate scandal under President Nixon and corruption issues in the 2001 Thai elections. These cases demonstrate the contrasting logic of effective and ineffective oversight and citizen responses.

In cases of effective oversight, the only trade-off for a backsliding incumbent lies between the costs of taking the clean action and the risk of detection. This logic appears to have been understood by President Nixon during the Watergate investigations. Already deeply concerned with discipline, following the publication of the Pentagon Papers President Nixon instructed his staff to establish the Special Investigations Unit, known as the Plumbers, to prevent "leaks" to the press and other oversight bodies. Following the Watergate break-in, President Nixon considered but dismissed cooperation with investigators to forestall controversy as doing so would expose other embarrassing misdeeds such as warrantless wiretaps and campaign finance violations. Effective oversight made for hard choices: either incur the cost of transparency and message discipline or hope for non-detection and risk triggering criticism, thereby persuading voters and incurring electoral consequences. Contemporary evidence suggests that Nixon and his aides believed they could subvert the investigation while avoiding detection (Genovese and Morgan, 2012).

To be clear, the model does not imply that Nixon could not try to discredit his opponents or frame investigations as political persecution. But it does suggest why such efforts would not succeed. Despite a coordinated campaign by Nixon's allies to undermine the press and congressional investigators, the public regarded the Watergate Committee with respect and found its allegations credible (Lang and Lang, 1983). In the wake of the scandal, the press saw a notable rise in national esteem: a Louis Harris survey from July 1973 found that only 17% of respondents believed "the press is just out to get President Nixon on Watergate" (Times, 1973). In the end, Nixon's eventual resignation was prompted by public pressure, including from his supporters, responding to effective oversight.

Our analysis of ineffective oversight suggests that when oversight fails, it allows politicians to maintain ambiguity. We can see that in practice in the trials of Thaksin Shinawatra, following the 2001 campaign for Prime Minister of Thailand (Leyland, 2008; Ginsburg, 2017). The 1997 Thai constitution established the National Counter-Corruption Commission (NCCC) to deter, detect, and prevent corrupt conduct on the part of officials. To do so, the NCCC was charged with collecting reports on assets from government officials to detect hidden payoffs. Those that failed to accurately report their assets would be referred to the Constitutional Court for a permanent ban from politics. However, as a new institution in a young democracy, the Court faced questions about its own legitimacy. It was supposed to ensure that public officials reported assets that could incentivize corrupt or undemocratic behavior, but doing so risked criticizing popular politicians.

Rather than gamble on a cover-up or come clean, Thaksin defended his failure to report assets as an "honest mistake" and a result of carelessness and confusion about the law (Tampa Bay Times, 2001). Thaksin's lawyers contended the commission was biased, and lacked the legal authority to bring charges. In effect, his defense argued that while Thaksin had not taken excessive care to avoid controversy, the issue was overblown, likely because of malicious and partisan oversight. In this case, despite the findings and referral of the NCCC, the Court lacked public support for a ban and voted 8–7 for acquittal. This case was widely seen as undermining the credibility of the Court, but its failure to prosecute was itself a function of the lack of consensus about the underlying corrupt behavior and doubts as to the motives of the NCCC.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

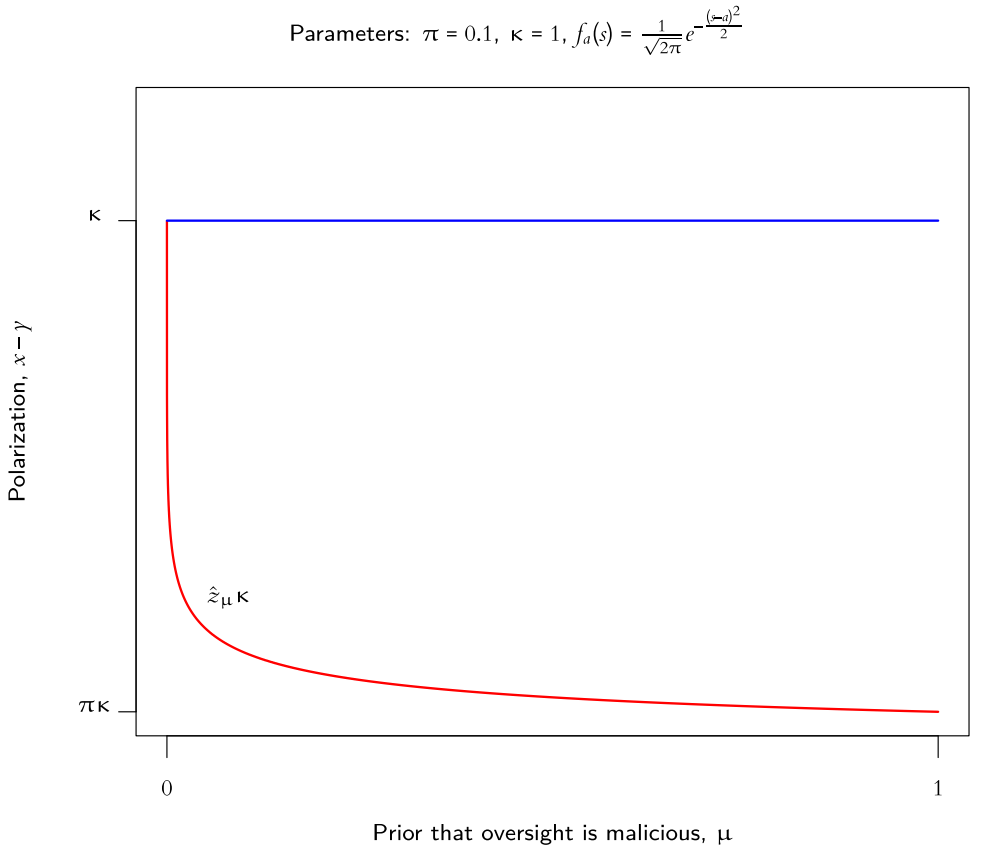


Figure 4. Threshold for backsliding.

6. Can oversight work?

In the preceding analysis, the oversight body plays the role of investigator, evaluator, and critic; its success in those roles depends on its emphasis on either avoiding false alarms or neglecting subversion, z. Oversight bodies take on a variety of institutional forms including profit-driven media, appointed ombuds, and standing legislative committees, whose incentives may generate higher or lower z. A non-partisan, unelected, and professional prosecutor may place as much weight on avoiding false alarms as neglected subversion, and have moderate z, whereas a more representative body may prioritize retaining popular incumbents, with z close to 1. In this section, we entertain two questions. First, there are political preferences or institutional prerogatives that would be effective in preventing backsliding, and second, are there versions of oversight that are not subject to the fake news effect?

To reiterate, Lemma 4 shows that the value of z that maximizes deterrence is $\hat{z}_\mu$. Is this maximal deterrence enough to prevent backsliding? Figure 4 compares the threshold for backsliding in the full information case characterized by Svolik (2020) with the threshold for backsliding by stealth. The x-axis indicates the common prior that the oversight body is malicious, μ, which ranges from 0 to 1. The y-axis indicates the extent of political polarization, $x - y$. The full information case is indicated by the blue line, where so long as $x - y$ is less than κ, the voter can prevent backsliding. The red line depicts the threshold given the introduction of the malicious type, assuming the maximal level of deterrence. Even then, the conditions are more permissive for backsliding, rapidly approaching the

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

no information threshold, $\pi\kappa$. In the following, we discuss how these thresholds are determined for two important classes of oversight body, the media and international monitors.

6.1. Editorial standards in the news media

In the case of media, we can think of z as governing coverage. Coverage decisions are determined by the selection of editorial standards, which, in turn, are a function of journalistic norms (Schudson, 2001), careerism among journalists (Baron, 2006), and market forces, such as government purchasing of advertisements (Szeidl and Szucs, 2021), corporate ownership (Gilens and Hertzman, 2000; DellaVigna and Hermle, 2017), or competition among outlets (Besley and Prat, 2006; Gentzkow and Shapiro, 2006).

Coverage is also affected by government regulations. In some cases, the news media could face significant *ex post* risks from raising a false alarm. Individual journalists often face intimidation, either by threats of violence or selective application of law. Analyzing patterns in local Mexican newspapers, Stanig (2015) finds that the severity of state-level defamation laws had a chilling effect on coverage of corruption and other illegal behavior by government officials. Under such restrictive laws, citizens would be particularly confident that any criticism was warranted, but criticism would also be rare.

Another determinant of media coverage is audience demand. In some cases, a public appetite for accountability can encourage aggressive reporting, even when false alarms are possible. One such period occurred in the United States at the beginning of the 20th century, when newspapers were known as muckrakers—a term popularized by President Theodore Roosevelt in response to a 1906 exposé on the influence of special interest groups in the U.S. Senate. Although intended as a critique, the term became emblematic of an era in which investigative journalism catalyzed major constitutional reforms and bolstered the broader Progressive movement (Weinberg and Weinberg, 2001). Walter Lippmann, echoing the logic of our model, argued that muckraking would not have thrived without strong audience demand: "The muckrakers spoke to a public willing to recognize as corrupt an incredibly varied assortment of conventional acts" (Lippmann, 1914, 5). In this case, the challenge of effective oversight was mitigated by a high public value placed on exposing subversion, a high κ.[11]

6.2. International democracy monitors

Above we focused on domestic sources of oversight, but our logic applies also to the variety of international organizations that are active in detecting and evaluating democratic backsliding. These independent, non-governmental organizations produce reports on the state of democracy around the world. For example, Freedom House's "Freedom in the World Report" produces an annual quantitative evaluation of the qualities of democracy and civil liberties across 195 countries and 15 territories. The Varieties of Democracy (V-Dem) Institute also produces annual reports for measures of democratic performance across 170 countries. These reports face similar editorial questions as the news media, drawing on expert analysis and expert surveys, along with various statistical corrections for differences across coders. These international scores tend to rely on comprehensive and expansive understanding of democracy to ensure that subversion does not go undetected.

Our model sheds new light on the debate over the interpretation of these scores and the role of subjective judgment among expert coders. Little and Meng (2024) raise the possibility that coder perceptions of democracy may shift over time in ways that make these scores more sensitive than is ideal. They recommend focusing instead on objective measures, such as *de facto* election outcomes. Following our analysis, relying on observed electoral returns, as opposed to a variety of more subtle

[11]Similarly, Mullainathan and Shleifer (2005) study the market for news and the role of consumer demand, arguing that media bias is shaped less by competition among outlets than by homogeneity in audience preferences.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

changes in society, would have the effect of setting a high z. If these international monitors are not contested, then such caution would improve the persuasive power of their reports. However, if there is some uncertainty about their motives, caution could backfire.[12]

7. Conclusion

Researchers, reporters, and other experts are necessary to help citizens detect and respond to subversion. One area of difficulty is ensuring that these organizations are able to detect subversion and announce criticism, but that is not the only challenge. As we have shown, backsliding can occur even when citizens are exposed to criticism. This is because citizens face uncertainty about the intentions of the oversight body that issues those criticisms. As a result, while democracies would certainly die in darkness, they can also die in the spotlight.

Our analysis complements past work that has focused on the role of political polarization in democratic backsliding. In our account, the democracy-policy trade-off depends on the information environment of voters. The problems we identify occur even under small amounts of uncertainty about the actions of the incumbent and the motives of the oversight body. These questions are not merely theoretical. As the examples at the outset of the paper show, backsliding events are often joined by high levels of political contestation. In Venezuela, for instance, Hugo Chávez was the target of criticism, but warnings from the media and opposition were undermined following the revelation of falsified audio of violent speech in an opposition advertisement (Kronick *et al.*, 2023). Such scandals can reinforce the dynamics we identify regarding the reputational risks from fake news. In an atmosphere of distrust, it can be difficult or even impossible for oversight bodies to overcome the fake news effect.

The caution effect and the fake news effect may apply outside the context of formal oversight institutions. Horizontal accountability can be provided by other states, academic researchers, or nongovernmental organizations, each of which faces their own potential reputational concerns (Bush and Prather, 2022). Beyond oversight of democracy, the dynamics we identify would apply to other cases of accountability where there is a risk of false alarms or neglected subversion and where failing to act has negative consequences for the future.

In our analysis of backsliding, we highlight some of the limitations of formal or legal constraints on power. The threat posed to democratic institutions depends on the operation of supporting institutions, such as the media, and adherence to norms of behavior on the part of incumbents (North *et al.*, 2009). A disciplined incumbent, one who takes the clean action in our setup, can be understood as a part of the informal, normative constraints on power. The need for discipline may be why ceremony is so important in conferring legitimacy to power transitions (Greif and Rubin, 2023). Unceremonious and undisciplined behavior introduces ambiguity over whether institutions have been subverted, even among those motivated to maintain those institutions. If incumbents could be forced to hold themselves to a higher standard and maintain discipline, detected subversion would not be controversial. As a result, following norms is important, whether in institutionalized democracy or in an autocracy. When those norms break down and incumbents engage in actions which, while not subversive, are possibly confused as such, backsliding is difficult to deter or correct.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2026.10109.

Competing interests. None.

[12]There is some evidence that this uncertainty has affected this form of international oversight of democratic outcomes. Bush and Prather (2022) survey Tunisia, the United States, and Georgia, finding wide variation in the reputations of election monitors and little evidence that the presence of these election observers increase trust in elections.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

References

Anderson CJ (1995) *Blaming the Government: Citizens and the Economy in Five European Democracies*. Armonk: M. E. Sharpe.
Andina-Díaz A and García-Martínez JA (2020) Reputation and news suppression in the media industry. *Games and Economic Behavior* **123**, 240–271.
Ashworth S and Bueno de Mesquita E (2014) Is voter competence good for voters?: Information, rationality, and democratic performance. *American Political Science Review* **108**(3), 565–587.
Barnes W (2000) Graft watchdog trips up ambitious tycoon: THAILAND. *South China Morning Post* 11.
Baron DP (2006) Persistent media bias. *Journal of Public Economics* **90**(1-2), 1–36.
Bermeo N (2016) On democratic backsliding. *Journal of Democracy* **27**(1), 5–19.
Besley T and Prat A (2006) Handcuffs for the grabbing hand? Media capture and government accountability. *American Economic Review* **96**(3), 720–736.
Boadle A (2021a) Brazil congress blocks changes to voting system despite military show of force. *Reuters* Available at https://www.reuters.com/world/americas/brazil-congress-blocks-changes-voting-system-despite-military-show-force-2021-08-10/ (accessed 26 September 2023).
Boadle A (2021b) Brazil's Bolsonaro rolls out tanks to cover his weak position. *Reuters* Available at https://www.reuters.com/world/americas/brazils-bolsonaro-rolls-out-tanks-cover-his-weak-position-2021-08-11/ (accessed 26 September 2023).
Bobbio N (1987) *The Future of Democracy: A Defence of the Rules of the Game*. Minneapolis: University of Minnesota Press.
Broockman DE, Kalla JL and Westwood SJ (2023) Does affective polarization undermine democratic norms or accountability? Maybe not. *American Journal of Political Science* **67**(3), 808–828.
Bush SS and Prather L (2022) *Monitors and Meddlers: How Foreign Actors Influence Local Trust in Elections*. New York: Cambridge University Press.
Chiopris C, Nalepa M and Vanberg G (2025) A Wolf in Sheep's clothing: Citizen uncertainty and democratic backsliding. *The Journal of Politics* **87**(4), 1272–1287.
Cumming-Bruce N (2001) Thaksin's lawyers accuse antigraft commission of bias — Thai agency defends its political neutrality. *Asian Wall Street Journal* 4.
Dahl RA (2005) What political institutions does large-scale democracy require? *Political Science Quarterly* **120**(2), 187–197.
DellaVigna S and Hermle J (2017) Does conflict of interest lead to biased coverage? Evidence from movie reviews. *Review of Economic Studies* **84**(4), 1510–1550.
Epstein D and O'Halloran S (1995) A theory of strategic oversight: Congress, lobbyists, and the bureaucracy. *The Journal of Law, Economics, and Organization* **11**(2), 227–255.
Ferejohn J and Shipan C (1990) Congressional influence on bureaucracy. *Journal of Law, Economics & Organization* **6**, 1.
Gehlbach S and Sonin K (2014) Government control of the media. *Journal of Public Economics* **118**, 163–171.
Genovese M and Morgan IW (2012) *Watergate Remembered: The Legacy for American Politics*. New York: Springer.
Gentzkow M and Shapiro JM (2006) Media bias and reputation. *Journal of Political Economy* **114**(2), 280–316.
Gibilisco M and Horz CM (2024) Juking the stats: Policing, misreporting, and policy evaluation. Working paper
Gilens M and Hertzman C (2000) Corporate ownership and news bias: Newspaper coverage of the 1996 Telecommunications Act. *Journal of Politics* **62**(2), 369–386.
Ginsburg T (2017) Constitutional afterlife: The continuing impact of Thailand's postpolitical constitution. In *Public Law in East Asia*. Abingdon: Routledge, pp. 289–311.
Ginsburg T and Huq A (2018) Democracy's near misses. *Journal of Democracy* **29**(4), 16–30.
Gratton G and Lee BE (2024) Liberty, security, and accountability: The rise and fall of illiberal democracies. *Review of Economic Studies* **91**(1), 340–371.
Greif A and Rubin J (2023) Endogenous political legitimacy: The Tudor roots of England's constitutional governance. ESI Working Paper 23-01 Economic Science Institute. Chapman University. https://digitalcommons.chapman.edu/esi_working_papers/381/.
Grillo E, Luo Z, Nalepa M and Prato C (2024) Theories of democratic backsliding. *Annual Review of Political Science* **27**, 381–400. https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-041322-025352.
Hazarika S (1989) New Delhi weapons scandal: Who knew what, and when? *The New York Times* Section A, Page. 14. https://www.nytimes.com/1989/11/02/world/new-delhi-weapons-scandal-who-knew-what-and-when.html.
Helmke G, Kroeger M and Paine J (2022) Democracy by deterrence: Norms, constitutions, and electoral tilting. *American Journal of Political Science* **66**(2), 434–450. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12668.
Iyengar S, Sood G and Lelkes Y (2012) Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly* **76**(3), 405–431.
Kronick D, Plunkett B and Rodriguez PL (2023) Backsliding by surprise: The rise of Chavismo. *Political Science Research and Methods* **11**(4), 838–854.
Lander LE and López Maya M (1999) Venezuela. La victoria de Chávez. El Polo Patriótico en las elecciones de 1998. *Nueva Sociedad*. Caracas: COYUNTURA, p.160.
Lang GE and Lang K (1983) *The Battle for Public Opinion: The President, The Press, and The Polls during Watergate*. New York: Columbia University Press.

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press

Larreguy H, Marshall J and Snyder JM (2020) Publicising malfeasance: When the local media structure facilitates electoral accountability in Mexico. *The Economic Journal* **130**(631), 2291–2327.
Levitsky S and Ziblatt D (2019) *How Democracies Die.* New York: Crown.
Lewis-Beck M (1988) *Economics and Elections: The Major Western Democracies.* Ann Arbor: University of Michigan Press.
Leyland P (2008) Thailand's constitutional watchdogs: Dobermans, bloodhounds or lapdogs? *Journal of Comparative Law* **2**, 151–177
Li A, Raiha D and Shotts KW (2022) Propaganda, alternative media, and accountability in Fragile Democracies. *The Journal of Politics* **84**(2), 1214–1219. https://doi.org/10.1086/715995
Lippmann W (1914) *Drift and Mastery.* New York: Mitchell Kennerly.
Little A (2012) Elections, fraud, and election monitoring in the shadow of revolution. *Quarterly Journal of Political Science* **7**(3), 249–283.
Little A (2015) Fraud and monitoring in non-competitive elections. *Political Science Research and Methods* **3**(1), 21–41.
Little A and Meng A (2024) Measuring democratic backsliding. *PS: Political Science & Politics* **57**(2), 1–13.
Luo Z and Przeworski A (2023) Democracy and its vulnerabilities: Dynamics of democratic backsliding. *Quarterly Journal of Political Science* **18**(1), 105–130. http://dx.doi.org/10.1561/100.00021112.
McCubbins MD and Schwartz T (1984) Congressional oversight overlooked: Police patrols versus fire alarms. *American Journal of Political Science* **28**(1), 165–179.
Milhorance F and Londoño E (2021) Bolsonaro prompts fears of a power grab with attacks on Brazil's voting system. *The New York Times* https://www.nytimes.com/2021/08/10/world/americas/bolsonaro-brazil-voting-system.html.
Miller MK (2021) A republic, if you can keep it: Breakdown and erosion in modern democracies. *The Journal of Politics* **83**(1), 198–213.
Mullainathan S and Shleifer A (2005) The market for news. *American Economic Review* **95**(4), 1031–1053.
Nalepa M and Pop-Eleches G (2022) Authoritarian infiltration of organizations: Causes and consequences. *The Journal of Politics* **84**(2), 861–873. https://doi.org/10.1086/715999.
New York Times (1973) Survey hails press on Watergate role. *New York Times* (July), p.20.
North DC, Wallis JJ and Weingast BR (2009) *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History.* New York: Cambridge University Press.
Powell GB and Whitten GD (1993) A cross-national analysis of economic voting: Taking account of the political context. *American Journal of Political Science* **37**(2), 391–414.
Schudson M (2001) The objectivity norm in American journalism. *Journalism* **2**(2), 149–170.
Shadmehr M and Bernhardt D (2015) State censorship. *American Economic Journal: Microeconomics* **7**(2), 280–307.
Stanig P (2015) Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican Press. *American Journal of Political Science* **59**(1), 175–193.
Stokes SC (2025) *The Backsliders: Why Leaders Undermine their Own Democracies.* Princeton, NJ: Princeton University Press.
Svolik MW (2013) Learning to love democracy: Electoral accountability and the success of democracy. *American Journal of Political Science* **57**(3), 685–702.
Svolik MW (2020) When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Quarterly Journal of Political Science* **15**(1), 3–31.
Szeidl A and Szucs F (2021) Media capture through favor exchange. *Econometrica* **89**(1), 281–310. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15641.
Tampa Bay Times (2001) Emotional Thai leader appeals for job. Published June 19, 2001, Updated Sept. 9, 2005. Available at: https://www.tampabay.com/archive/2001/06/19/emotional-thai-leader-appeals-for-job/.
Waldner D and Lust E (2018) Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science* **21**, 93–113.
Weinberg A and Weinberg LS (2001) *The Muckrakers.* Urbana: University of Illinois Press.
Weingast BR and Moran MJ (1983) Bureaucratic discretion or congressional control? Regulatory policymaking by the Federal Trade Commission. *Journal of Political Economy* **91**(5), 765–800.
Zaller J (2003) A new standard of news quality: Burglar alarms for the monitorial citizen. *Political Communication* **20**(2), 109–130.

Cite this article: Gulotty R and Luo Z (2026) Backsliding under contested oversight. *Political Science Research and Methods*, 1–21. https://doi.org/10.1017/psrm.2026.10109

https://doi.org/10.1017/psrm.2026.10109 Published online by Cambridge University Press