

Talking democrats, loyal partisans: when opinion polls fail to warn of democratic rupture*

Marcio Firmo[†]

Pedro Hemsley[‡]

Abstract

Polarization makes punishing an undemocratic co-partisan costly. We show that conformity amplifies this familiar trade-off. When its maximum feedback strength exceeds one, democratic enforcement disappears at a tipping point. We then compare costly behavior with a survey that may itself respond to polarization. Near the tipping point, revealed support becomes arbitrarily more sensitive to polarization than any survey measure with a smooth, bounded response. The poll need not be false or flat; it lacks the strategic multiplier that governs behavior. Conventional surveys may therefore move smoothly while enforcement collapses discontinuously.

Keywords: democratic backsliding; social identity; surveys; polarization; strategic complementarities.

JEL: D72; D91; P16.

1 Introduction

Conventional survey measures show that public support for democracy is broad and persistent (Claassen, 2020). Yet this commitment falters when defending democratic norms requires voters to oppose their own party. In candidate-choice experiments, only about 3.5% of respondents vote against a co-partisan candidate who violates democratic principles (Graham and Svolik, 2020). Support for norm-eroding policies also increases when respondents' own party, rather than the opposition, holds power (Simonovits et al., 2022). Why can stated support for democracy remain strong while partisan voters become reluctant to enforce democratic norms?

We study a mechanism that can reconcile these patterns: identity conformity. Existing work shows that polarization weakens democratic accountability by increasing what voters must sacrifice to punish an undemocratic co-partisan (Svolik, 2020). Earlier theory shows that democratic enforcement can itself be a coordination problem among citizens (Weingast, 1997). We connect these insights by allowing the identity cost of opposing one's party to depend on how many fellow partisans remain loyal. When others continue to support the party, defection becomes more costly. Each partisan's loyalty therefore makes loyalty more attractive to others: loyalty choices are strategic complements (Akerlof and Kranton, 2000).

*Declarations of interest: none. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Proofs are in the Online Appendix.

[†]State University of Rio de Janeiro and Brazilian Development Bank (BNDES), Rio de Janeiro, Brazil. E-mail: marciogoldfirmo@gmail.com.

[‡]Corresponding author. Institute of Economics, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: pedro.hemsley@ie.ufrj.br.

This feedback follows the familiar logic of threshold models (Granovetter, 1978; Kuran, 1995). A gradual increase in polarization may initially have little visible effect on democratic enforcement. When conformity feedback is sufficiently strong, polarization eventually reaches a critical level at which the equilibrium sustaining enforcement disappears and behavior shifts abruptly toward partisan loyalty.

This tipping mechanism also creates a measurement problem. Gaps between survey responses and costly action are well documented (Valentim, 2026), but they need not imply that respondents are lying. Expressive responding appears modest on average (Graham, 2026), and recent evidence finds no systematic social-desirability inflation in survey measures of democratic attitudes (Magalhães and Aarslew, 2025). We therefore allow survey responses to vary with polarization and assume that stated and revealed support reflect the same underlying democratic values. They differ in the threshold at which those values are elicited: surveys apply their own elicitation cutoff, whereas costly behavior depends on an endogenous identity price determined partly by the behavior of other partisans.

Our contribution is to identify when the gradual erosion documented in existing work becomes a discontinuous rupture. If maximum conformity feedback exceeds one, partisan loyalty becomes self-reinforcing, democratic enforcement exhibits tipping and hysteresis, and anticipated loyalty can suddenly cross the threshold that deters incumbent violations. Because conventional surveys do not contain this strategic feedback, they may remain smooth and informative about individual values while providing little warning about the collapse of collective enforcement.

2 Loyalty and democratic enforcement

An incumbent leads a group whose supporters form a continuum $i \in [0, 1]$. After a democratic violation, supporter i stays loyal (L) or defects and punishes (D). Payoffs are

$$u_i(L) = \theta + \gamma\ell - \delta_i, \quad u_i(D) = 0, \quad (1)$$

where ℓ is the loyal share, θ is the partisan stake (our polarization index), $\gamma \geq 0$ measures identity conformity, and $\delta_i \in \mathbb{R}$ is the citizen's democratic value, with continuously differentiable c.d.f. F and positive density f . Higher δ_i makes punishment more likely. Following Svoblik (2020), θ captures what voters sacrifice by punishing their own party; $\gamma\ell$ is an identity payoff that increases with the share of co-partisans who conform (Akerlof and Kranton, 2000). The incumbent violates when anticipated loyalty exceeds a reduced-form safety threshold $\bar{\ell} \in (0, 1)$.¹

Citizen i is loyal if and only if $\delta_i \leq \theta + \gamma\ell$, so equilibrium loyalty solves

$$\ell = F(\theta + \gamma\ell). \quad (2)$$

For $\gamma > 0$, the loyalty condition may be written as $\ell \geq t_i(\theta) \equiv (\delta_i - \theta)/\gamma$: each citizen conforms once the loyal share exceeds a personal threshold. The distribution of these thresholds is induced by F , as in Granovetter (1978). Because loyalty feeds back into itself, equation (2) may have several solutions (Section 3). A *regular-continuation* rule follows a fixed point continuously in θ while $\gamma f(\theta + \gamma\ell) < 1$, irrespective of the adjustment process. Our leading example is total adjustment to prior loyalty,

$$\ell_{t+1} = F(\theta_t + \gamma\ell_t), \quad (3)$$

¹The Appendix gives a payoff foundation for this threshold.

while polarization θ_t drifts slowly.² Under (3), a fixed point attracts nearby states when $\gamma f(\theta + \gamma \ell) < 1$. The interpretation is straightforward. A rise in loyalty of Δ raises the identity payoff by $\gamma \Delta$, which converts a further mass $\gamma f(\theta + \gamma \ell) \Delta$ of marginal citizens from punishment to loyalty. If this second-round increase is smaller than the first, the disturbance disappears over time; if it is larger, the fixed point is repelling. Thus stable solutions attract nearby states, and the same feedback term governs how strongly behavior responds to polarization (Proposition 1).

Regular-continuation tracks a stable branch until it ends at a fold, where the low-loyalty fixed point collides with its unstable neighbor. With no nearby low-loyalty solution, selection moves to the remaining high-loyalty equilibrium. Write $\ell^*(\theta)$ for loyalty on the branch selected in this way, and let $p(\theta) \equiv \theta + \gamma \ell^*(\theta)$ denote the endogenous identity price: the democratic value a citizen must exceed in order to punish, once both the partisan stake and the identity cost of breaking with loyal co-partisans are taken into account.

We turn now to the outcomes of surveys. Define revealed support as $R(\theta) = 1 - \ell^*(\theta) = 1 - F(p(\theta))$. A survey asks whether democracy is supported. We use the reduced-form elicitation rule

$$r_i(\theta) = \mathbf{1}\{\delta_i \geq q(\theta)\}, \quad S(\theta) = 1 - F(q(\theta)), \quad (4)$$

where $q(\theta)$ is a differentiable survey cutoff. Answers rank citizens by their true values, while the elicitation environment determines the cutoff. This differs from preference falsification (Kuran, 1995), in which public statements detach from private values. A conventional anonymous question has constant q ; allowing $q'(\theta) \neq 0$ admits partisan pressure, framing, or other smooth movement in expressed attitudes.³

Proposition 1 (Two cutoffs). *On any stable interior branch,*

$$S - R = F(p) - F(q), \quad R'(\theta) = -\frac{f(p)}{1 - \gamma f(p)}, \quad S'(\theta) = -f(q)q'(\theta).$$

Stated support exceeds revealed support if and only if the survey cutoff lies below the identity price, $q(\theta) < p(\theta)$.

Proposition 1 separates two objects. The first is a wedge: $S - R$ is the mass of citizens whose democratic values lie between the two cutoffs. They affirm democracy in the survey, $\delta_i \geq q$, but will not pay the identity price to punish, $\delta_i \leq p$. The wedge exists even without conformity: when $\gamma = 0$, $p = \theta$, so $S - R = F(\theta) - F(q)$ can be positive. A cross-sectional survey-action gap therefore does not identify conformity; that mechanism appears in the multiplier and the resulting nonlinear dynamics. Polarization moves revealed support through a direct effect, $f(p)$, and through the feedback $[1 - \gamma f(p)]^{-1}$: the marginal citizens who switch to loyalty raise the identity price for everyone else, who switch in turn. Stated support carries no such term, because one citizen's answer does not change the cutoff faced by another. The wedge explains why polls and behavior differ in a given period; the multiplier, which becomes unbounded at the tipping point characterized in Section 3, explains why their paths may differ over time.

²Partial and continuous adjustment are other implementations; the Online Appendix also contrasts global selection.

³For an example, Ferroni (2026) documents post-election increases in survey-measured violent and anti-democratic sentiments among extreme supporters of Brazil's losing candidate.

3 Tipping and hysteresis

We now study how equilibrium loyalty changes with polarization. At a candidate identity price $x = \theta + \gamma\ell$, a small initial rise in loyalty generates an approximately $\gamma f(x)$ -times-as-large second-round rise. Thus $\gamma f(x)$ measures the local strength of conformity feedback. This coefficient organizes the analysis below.

Assumption 1 (Shape of the value distribution). *The density f is single-peaked: it is strictly increasing up to an interior mode m and strictly decreasing thereafter. It also has vanishing tails, $f(x) \rightarrow 0$ as $x \rightarrow \pm\infty$.*

Assumption 1 concerns only the cross-sectional distribution of democratic values, not equilibrium behavior. Its first part is strict single-peakedness (unimodality) of the density: values form one central cluster and become progressively less common away from it. It does not require symmetry. The vanishing-tail condition additionally requires extreme values to be increasingly rare. Normal, logistic, Student- t , Cauchy, and Laplace densities satisfy these shape conditions, as do many skewed unimodal densities.

Define

$$\kappa \equiv \gamma f(m).$$

Because $f(m)$ is the density's maximum, κ is the strongest possible local conformity feedback. It is high when conformity is strong or democratic values are concentrated, and low when values are dispersed. For example, $\kappa > 1$ is equivalent to $\gamma > \sqrt{2\pi}\sigma$ under a normal distribution with standard deviation σ , and to $\gamma > 4s$ under a logistic distribution with scale s . We do not assume strong feedback; the next result characterizes equilibrium for every κ .

Proposition 2 (Equilibrium regimes). *Under Assumption 1:*

1. *If $\kappa < 1$, equilibrium is unique and locally stable for every θ , and loyalty varies smoothly with polarization.*
2. *If $\kappa = 1$, equilibrium remains unique and continuous in θ , although the local stability condition binds when the identity price equals m . There is no equilibrium coexistence or hysteresis.*
3. *If $\kappa > 1$, there are two identity prices $x_1 < m < x_2$ satisfying $\gamma f(x) = 1$. Define*

$$\theta_H = x_1 - \gamma F(x_1), \quad \theta_L = x_2 - \gamma F(x_2).$$

Then $\theta_L < \theta_H$. For $\theta \in (\theta_L, \theta_H)$, a low-loyalty enforcement equilibrium and a high-loyalty acquiescence equilibrium are locally stable, separated by an unstable equilibrium. Under this selection, loyalty jumps upward at θ_H as polarization increases along the enforcement branch. Reversing polarization does not restore enforcement until $\theta < \theta_L$.

The threshold $\kappa = 1$ therefore separates a weak-feedback region, in which polarization changes behavior continuously, from a strong-feedback region in which strategic amplification produces tipping and hysteresis.

Corollary 1 (Democratic rupture). *If $\kappa > 1$ and the safety threshold $\bar{\ell}$ lies between loyalty immediately before and after the fold, a continuous rise in polarization causes anticipated loyalty to jump across the incumbent's safety threshold at θ_H . A democratic violation that was deterred immediately before the tipping point becomes safe from effective punishment immediately afterward.*

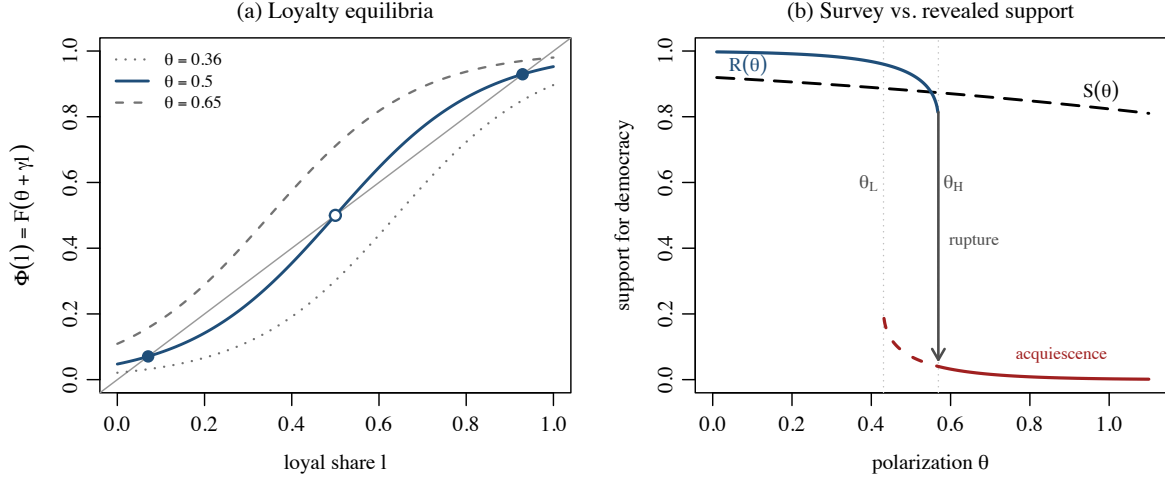


Figure 1: (a) Loyalty equilibria $\ell = F(\theta + \gamma\ell)$; filled points are stable and the open point is unstable. (b) A responsive survey $S(\theta)$ remains smooth while revealed support on the enforcement branch becomes steep and then collapses at θ_H . The dashed red segment is the coexisting acquiescence branch. The illustration uses a logistic F , $\gamma = 1$ ($\kappa = 1.5$), and a linear survey cutoff.

4 Why a responsive poll may still fail

The key contrast is not *truthful* versus *expressive* speech. It is nonstrategic measurement versus strategic behavior.

Proposition 3 (Vanishing relative warning). *Suppose $\kappa > 1$ and approach θ_H from below along the enforcement branch. If q is continuously differentiable near θ_H , then*

$$R'(\theta) \rightarrow -\infty, \quad \frac{|S'(\theta)|}{|R'(\theta)|} \rightarrow 0.$$

Moreover, S is continuous at θ_H , where R jumps.

The survey may decline with polarization and may perfectly rank citizens by their democratic values. Yet its cutoff moves smoothly. Costly behavior is different: as conformity brings the system toward the fold, a marginal rise in loyalty raises the identity price, which creates more loyalty and raises the price again. The multiplier $[1 - \gamma f(p)]^{-1}$ becomes arbitrarily large. If survey answers themselves carried an identity cost rising with the loyal share—preference falsification in the sense of [Kuran \(1995\)](#)—the cutoff would inherit the same multiplier and stated support would move with behavior; the result therefore characterizes instruments whose elicitation environment does not price the group’s response. To track resilience near rupture, an instrument must inherit this amplification or attach a real partisan cost to the response, as incentivized candidate-choice designs do ([Graham and Svolik, 2020](#); [Bullock et al., 2015](#)).

5 Concluding remark

Polarization has an ordinary direct effect on democratic punishment and an endogenous multiplier through conformity. That multiplier, not survey insincerity, is the source of the early-

warning failure. Abstract support can remain high and even move in the right direction while the costly action that sustains democracy approaches a discontinuous collapse.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) and Codex (OpenAI) to check computations, to assist with coding and L^AT_EX preparation, and to review and adjust the writing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Data availability

No data were used for the research described in this article. The code that reproduces Figure 1 is provided as supplementary material.

References

- Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. *Q. J. Econ.* 115, 715–753.
- Bullock, J.G., Gerber, A.S., Hill, S.J., Huber, G.A., 2015. Partisan bias in factual beliefs about politics. *Q. J. Polit. Sci.* 10, 519–578.
- Claassen, C., 2020. Does public support help democracy survive? *Am. J. Polit. Sci.* 64, 118–134.
- Ferroni, M.F., 2026. Election results and democratic discontent: expectations, extremism, and democratic values in post-election Brazil. *J. Public Econ.* 262, 105770.
- Graham, M.H., 2026. Partisan expressive responding: lessons from two decades of research. *Am. Polit. Sci. Rev.*, First View. <https://doi.org/10.1017/S0003055426101452>.
- Graham, M.H., Svobik, M.W., 2020. Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *Am. Polit. Sci. Rev.* 114, 392–409.
- Granovetter, M., 1978. Threshold models of collective behavior. *Am. J. Sociol.* 83, 1420–1443.
- Kuran, T., 1995. *Private Truths, Public Lies*. Harvard Univ. Press.
- Magalhães, P.C., Aarslew, L.F., 2025. Survey measures of democratic attitudes and social desirability bias. *Polit. Sci. Res. Methods*, First View. <https://doi.org/10.1017/psrm.2025.10044>.
- Simonovits, G., McCoy, J., Littvay, L., 2022. Democratic hypocrisy and out-group threat: explaining citizen support for democratic erosion. *J. Polit.* 84, 1806–1811.
- Svobik, M.W., 2020. When polarization trumps civic virtue: partisan conflict and the subversion of democracy by incumbents. *Q. J. Polit. Sci.* 15, 3–31.
- Valentim, V., 2026. Norms of democracy, staged democrats, and supply of exclusionary ideology. *Comp. Polit. Stud.* 59, 1367–1405.
- Weingast, B.R., 1997. The political foundations of democracy and the rule of law. *Am. Polit. Sci. Rev.* 91, 245–263.