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

Marcio Firmo† Pedro Hemsley‡

September 2026

Abstract

Polarization makes punishing an undemocratic co-partisan costly. We show that conformity amplifies this familiar trade-off: loyalty becomes a strategic complement and 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

Citizens overwhelmingly tell surveyors that they support democracy, yet few punish a co-partisan who violates democratic norms. In candidate-choice experiments, only about 13.1% are willing to defect from an undemocratic co-partisan when doing so means voting against their own party (Graham and Svolik, 2020); citizens also excuse norm erosion when their own side governs (Simonovits et al., 2022). Meanwhile, conventional measures of democratic support are high and persistent, and on average they do predict democratic survival (Claassen, 2020). Our question is why such measures may nonetheless miss a collapse.

Recent episodes display the pattern. In the United States, about two thirds of Republicans said in 2021 that the 2020 election had been stolen (PRRI, 2021), and the Capitol was stormed on 6 January 2021. In Brazil, the share of voters calling democracy always preferable reached 79% in October 2022, the highest since the series began in 1989 (Datafolha, 2022); within three months, after the incumbent had contested the result, his supporters stormed the seats of the three branches of government on 8 January 2023.

*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, 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.

Existing work already explains why polarization weakens democratic punishment: voters trade democratic principles against partisan stakes (Svolik, 2020). Earlier theory also treats democratic enforcement as a coordination problem among citizens (Weingast, 1997). Our narrower contribution combines these forces through *identity conformity*. Punishing one's group is more costly when many co-partisans remain loyal. Loyalty choices are therefore strategic complements (Akerlof and Kranton, 2000). The feedback follows familiar threshold-model logic (Granovetter, 1978; Kuran, 1995); applied here, it implies that a smooth increase in polarization can destroy democratic enforcement discontinuously.

The discontinuity creates a measurement problem. Survey/action gaps are well documented (Valentim, 2026), but they need not reflect lying. Recent evidence finds expressive responding to be modest on average (Graham, 2026), and finds no systematic social-desirability inflation in survey measures of democratic attitudes (Magalhães and Aarslew, 2025). We accordingly allow the survey to respond smoothly to polarization and to reflect the same democratic values that govern costly action. The survey evaluates those values at its own elicitation cutoff; behavior evaluates them at an endogenous *identity price*.

Three results follow. First, stated and revealed support are tail probabilities evaluated at different cutoffs. Second, conformity creates two stable branches and hysteresis. Third, the behavioral cutoff contains a strategic multiplier that diverges at the tipping point. Any survey cutoff with bounded sensitivity becomes locally uninformative relative to costly behavior. Thus polls *may fail* to warn not because they are assumed constant, but because they do not internalize the feedback that sustains democracy.

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

$$u_i(L) = \theta + \gamma\ell - \delta_i, \qquad u_i(D) = 0, \tag{1}$$

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. The incumbent violates when anticipated loyalty exceeds a reduced-form safety threshold $\bar{\ell} \in (0,1)$; the Appendix gives a payoff foundation.

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

$$\ell = F(\theta + \gamma\ell). \tag{2}$$

Let $p(\theta) \equiv \theta + \gamma\ell^*(\theta)$ denote the endogenous identity price on a selected stable branch. Revealed support is $R(\theta) = 1 - \ell^*(\theta) = 1 - F(p(\theta))$.

To discipline equilibrium selection, let loyalty adjust according to

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

with polarization changing slowly. An interior fixed point is locally stable when $\gamma f(p) < 1$; the selected equilibrium continues from the preceding stable branch.

A survey asks whether democracy is supported. We use the reduced-form elicitation rule

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

where $q(\theta)$ is a differentiable survey cutoff. A conventional anonymous question has constant q; allowing $q'(\theta) \neq 0$ admits partisan pressure, framing, or other smooth movement in expressed attitudes.

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

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

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

3 Tipping and hysteresis

Assumption 1 (Fold condition)**.** *The density f is strictly increasing up to an interior mode m, strictly decreasing thereafter, satisfies $f(x) \to 0$ as $x \to \pm\infty$, and $\gamma f(m) > 1$.*

Assumption 1 guarantees two—not merely one—crossings of the unit-slope condition. Let $x_1 < m < x_2$ solve $\gamma f(x) = 1$, and define

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

Proposition 2 (Tipping and hysteresis)**.** *Under Assumption 1, $\theta_L < \theta_H$. 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 (3), increasing polarization along the enforcement branch produces an upward jump in loyalty at θ_H. Reversing polarization does not restore enforcement until $\theta < \theta_L$.*

Corollary 1 (Democratic rupture)**.** *If the safety threshold $\bar{\ell}$ lies between loyalty immediately before and after the jump at θ_H, a smooth increase in polarization makes violation safe discontinuously.*

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)**.** *Approach θ_H from below along the enforcement branch. If q is continuously differentiable and $f(q(\theta))q'(\theta)$ is bounded near θ_H, then*

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

If q is continuous, S also has no discontinuity when R jumps at θ_H.

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.

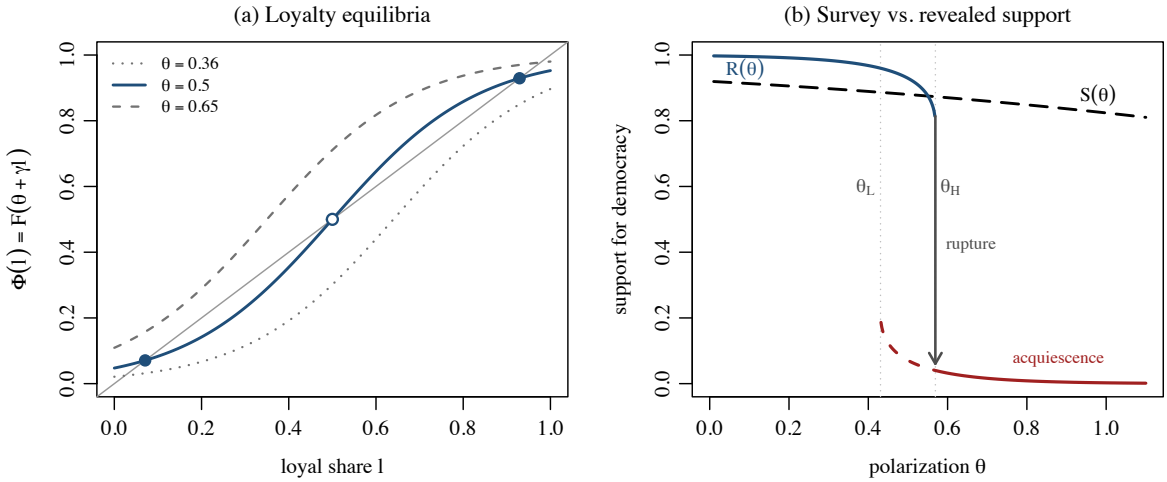

Figure 1: (a) Loyalty equilibria $\ell = F(\theta+\gamma\ell)$; filled points are stable and the open point is unstable. (b) A responsive survey $S(\theta)$ remains smooth while revealed support on the enforcement branch becomes steep and then collapses at θ_H. The dashed red segment is the coexisting acquiescence branch. The illustration uses a logistic F with location 1 and scale 1/6 (so $\gamma f(m) = 1.5$, satisfying Assumption 1), $\gamma = 1$, and a linear survey cutoff.

The divergence itself is the standard geometry of a saddle-node fold. What Proposition 3 adds is the identity of the denominator: it is the conformity multiplier that governs costly action, and no instrument whose elicitation environment does not price the group's response can inherit it. 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 LaTeX 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.

Datafolha, 2022. Apoio à democracia atinge recorde de 79%. Survey of 17–19 October 2022, n = 2,912. Folha de S.Paulo, São Paulo.

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., Svolik, 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.

Public Religion Research Institute, 2021. Competing Visions of America: An Evolving Identity or a Culture Under Attack? 2021 American Values Survey. PRRI, Washington, DC.

Simonovits, G., McCoy, J., Littvay, L., 2022. Democratic hypocrisy and out-group threat: explaining citizen support for democratic erosion. J. Polit. 84, 1806–1811.

Svolik, 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.