

Week 5: Democratic backsliding

A second introduction to formal political economy, Trinity Term 2026

Jacob Edenhofer

04/06/2026

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

Overview

The lecture is organised around two cross-cutting distinctions:

- **Demand vs. supply:** who acts (or fails to act) – voters or elites?
- **Target of backsliding:** electoral manipulation vs. executive aggrandizement (**grilloetal2024**)

1. Conceptualising and measuring democracy

What is backsliding, and how do we recognise it?

2. Demand-side theories

Why do voters fail to defend democracy at the ballot box?

3. Supply-side theories

When do elites attempt to undermine democracy, and what restrains them?

4. Consequences of backsliding

5. Conclusion and outlook

Where does the debate stand? What is the risk to established democracies?

Motivation: Is democracy in decline?

Concerned voices

- Levitsky and Ziblatt (2018): *How Democracies Die*
- Freedom House (2022): “democracy under siege”
- V-Dem (2023): “the level of democracy enjoyed by the average global citizen in 2022 is down to 1986 levels”
- M. Marshall, Polity Project (Feb. 2025): “Congratulations America! The Polity Project has downgraded its score for the USA from +8 . . . to 0.”

Sceptical voices

- Treisman (2023): share of democracies is at or near an all-time high; US breakdown hazard remains very low
- Little and Meng (2024): “if the world is experiencing major backsliding in the aggregate, we should expect to see evidence on the objective measures – but we do not”
- Reinforced by Viktor Orbán’s (surprising) decision to concede the election to Peter Magyar, despite Hungary having been the paradigmatic case of backsliding in the literature

Motivation: If so, or if not, why?

(a) Voters (*demand side*)

- Citizens tolerate, excuse, or actively reward authoritarian behaviour
- Partisan loyalty, motivated reasoning, divergent conceptions of democracy
- “Just vote them out” fails

(b) Elites (*supply side*)

- Politicians and parties exploit institutional weaknesses
- Horizontal constraints fail
- Business and economic elites acquiesce / benefit

(c) Structural forces

- Economic shocks, rising inequality, cultural backlash
- Affective polarisation, social media, declining trust
- These cut across (a) and (b)

(d) Wrong question

- Reports of backsliding reflect observer bias, not regime change
- “Objective” indicators tell a different story (cf. Little & Meng)

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

Defining and operationalising our terms

Two types of conceptions of democracy (Shapiro, [2020](#)): procedural (“minimalist”) vs. substantive.

- **Minimalist** conceptions classify countries as democratic *if and only if* they have a certain set of formal democratic procedures / institutions in place, most notably *free and competitive elections*.
 - These formal institutions are regarded as both a *necessary* (uncontroversial) and *sufficient* (controversial) condition for democracy.
 - Implication: outcomes produced by these institutions have no bearing on classification.
- **Substantive** conceptions regard formal procedures as necessary, but not sufficient, instead requiring other outcomes to be present as well, such as the *protection of civil liberties and political rights*.
 - Intermediate conceptions possible.

These conceptions can be operationalised via **three types of scales**: dichotomous / binary; ordinal; continuous.

⇒ Any **measure of democracy** = (i) a conception + (ii) a scale.

How to assess a measure of democracy?

Three criteria for assessing the relative goodness of measures of democracy (and other variables in social science more generally):

1. **Conceptual validity**

- Degree to which measures — both conception and scale — correspond to the concepts they are intended to represent.

2. **Reliability / reproducibility**

- “Reliability refers to the extent to which the *measurement process repeatedly and consistently produces the same score* [or classification] for a given score” (Clark, M. Golder, and S. N. Golder, 2025, p. 166).

3. **Interpretability of scale**

- Do (unit) changes in scale represent *theoretically meaningful* quantities?

Overview of prominent measures of democracy

Name	Conception	Scale	Temporal	Spatial
Boix-Miller-Rosato (BMR), v4.0	Minimalist	Dichotomous	1800–2020	222
Democracy-and-Dictatorship (DD)	Minimalist	Dichotomous	1950–2020	210
EIU Democracy Index	Substantive	Continuous	2006–2020	167
Freedom House	Substantive	Ordinal	1972–2021	221
Geddes–Wright–Frantz (GWF)	Minimalist	Dichotomous	1742–2010	155
Polity V	Intermediate	Ordinal	1776–2020	186
Svolik	Minimalist	Dichotomous	1946–2008	198
V-Dem (v11.1)	Intermediate	Continuous	1789–2020	202

(Dis)advantages of different measures

Conception	Scale	Examples	Advantages	Disadvantages
Minimalist	Dichotomous	DD, Svolik, BMR	High reproducibility; high interpretability	High risk of type I errors (if H_0 : non-democratic) $\Rightarrow$ concept validity contentious
Minimalist	Dichotomous + sub-types	DD, GWF	High reproducibility; high interpretability	Shortcomings of ideal types / typologies (Svolik, 2012) $\Rightarrow$ concept validity contentious
Substantive / intermediate	Ordinal	FH, Polity V	High risk of type II errors $\Rightarrow$ picks up subtle trends; early-warning function	Low reproducibility (expert-coded); low interpretability of intermediate values
Substantive / intermediate	Continuous	EIU, V-Dem	Lower type II risk than ordinal scales; still higher than for minimalist conceptions	Low reproducibility; low interpretability

Interlude: measuring democracy vs. building an early-warning system

Task 1: Measuring democracy

- Goal: estimate the level / type of regime at time t .
- Desiderata: high reproducibility, conceptual validity across cases and time, interpretable scale.
- Type-I and type-II errors are roughly symmetric in cost.
- L&M's preferred trade-off (minimalist + observable) is *defensible* for this task.

Task 2: Early-warning system

- Goal: detect incipient backsliding *before* it shows up in headline electoral outcomes.
- Desiderata: sensitivity to subtle, hard-to-observe shifts; tolerance for false positives.
- Type-II errors (missing a real episode) very costly; type-I (false alarm) much less so.
- Judgement-based measures (V-Dem, FH) may be *better suited* here even if worse as point estimates.

Does it matter which measure we use? (i)

Yes – results can be sensitive to the measure chosen.

- Assuming insensitivity by default is unwarranted.
- Cheibub, Gandhi, and Vreeland (2010) reproduce Rodrik and Wacziarg (2005) on the effects of regime transitions on economic growth.
 - Only Polity yields a significant positive effect of democratisation on growth.
 - With DD (binary scale + minimalist conception), the effect becomes insignificant.

Table 1 The effect of regime transition on economic growth

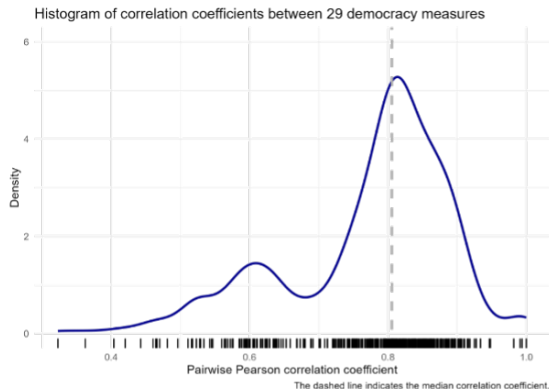
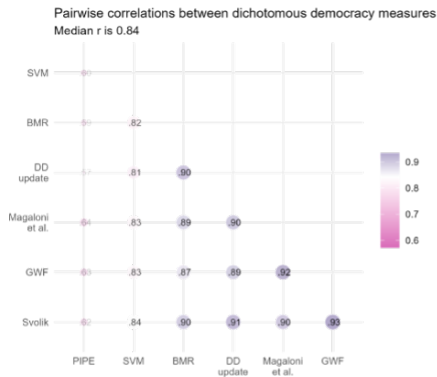
Variable	POLITY	DD
New Democracy	0.89** (0.42)	0.15 (0.40)
Established Democracy	-0.14 (0.47)	0.03 (0.44)
New Autocracy	0.17 (0.46)	-0.21 (0.56)
Established Autocracy	-0.23 (0.47)	1.05* (0.61)
Small Regime Change	0.94** (0.28)	
State Failure	-6.14** (0.95)	-6.50** (0.94)
Constant	1.63** (0.79)	1.78** (0.79)
Number of observations	5,465	5,465
Number of groups	154	154

Notes: The label of each column indicates the measure of political regime employed to generate the variables listed in the variable column. Both models allow for serially correlated (AR1) disturbances. Standard errors in parenthesis

**Indicates significance at the 0.05 level

*Indicates significance at the 0.10 level

Does it matter which measure we use? (ii)



The current debate (i.a): Little & Meng's challenge

- Conventional account: “Democracy Dies in Darkness”; V-Dem reports the average global citizen's democracy level has fallen to 1986 levels.
- L&M's hypothesis: *existing studies rely almost entirely on subjective, expert-coded indicators*; if there were major backsliding in the aggregate, we should also see it in *objective* indicators — but we do not.
- Two candidate explanations for the divergence:
 1. **Time-varying coder bias**: experts have become more pessimistic about democracy, biasing recent ratings downward.
 2. **Strategic substitution**: leaders have learned to use subtle violations that elude objective measurement.

“If the world is experiencing major backsliding in the aggregate, we should expect to see evidence of this effect on the objective measures — particularly incumbent leaders and their parties winning elections — but we do not.”

(Little and Meng, [2024](#))

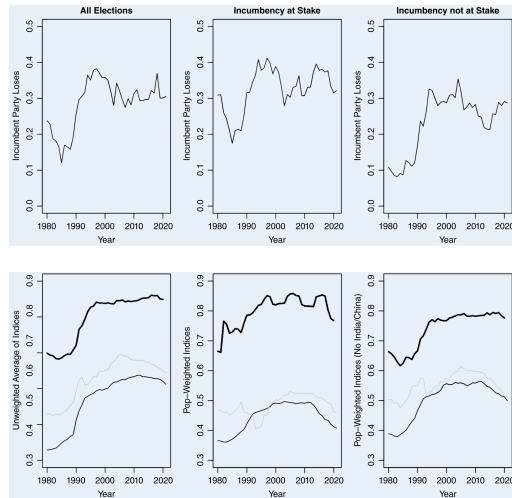
The current debate (i.b): Little & Meng's evidence

L&M's *objective* indicators:

- Incumbent party vote / seat shares / loss rate
- Multi-party competition; process violations
- Term limits; succession and dismissal rules
- Journalists jailed / murdered

Findings:

- Incumbent turnover *has not declined* since 1990. Multi-party indices and process-violation indices show *no recent deterioration*.
- Constitutional executive constraints (term limits, dismissal, succession) are *stable or improving*. Only attacks on journalists show clearly worsening trends.



The current debate (ii.a): Response by Knutsen et al., 2024

1. **“Objective” vs “subjective” is a false dichotomy.**

- Even L&M's indicators (e.g. process-violation index, term-limit evasions) require judgement calls; inter-coder agreement is far from perfect.
- Better terminology: *observer-invariant* vs *judgement-based*. Observer-invariance is not the same as being unbiased.

2. **L&M's quasi-minimalist conception misses what backsliding targets.**

- Contemporary backsliding works precisely through degrading civil liberties, civil society, freedom of expression, judicial independence.
- These dimensions are largely *outside* L&M's index by construction ⇒ low concept validity for the backsliding question.

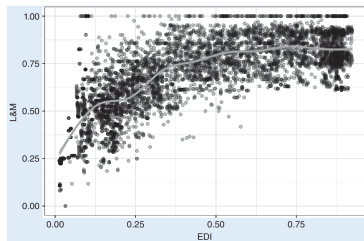
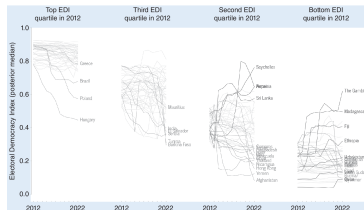
3. **No empirical evidence of “bad-vibes bias” in V-Dem.**

- V-Dem's measurement model corrects for coder reliability and scale-perception.
- Tests for systematic downward revision by V-Dem experts: *no consistent pattern*.

The current debate (ii.b): Knutsen et al.'s specifics

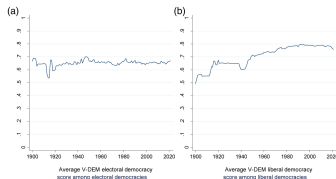
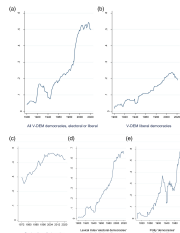
- 32 countries with significant negative changes in V-Dem EDI over 2012–2022; 11 with significant positive changes.
- Most backsliders in the upper half of the democracy distribution (Brazil, Poland, Hungary, USA, India, Serbia...).
- Decline visible *within* V-Dem subindices: freedom of expression, censorship, civil society repression. Electoral subindices show *less* decline.

Summary: the global aggregate is modest, but country-level backsliding episodes are real – and concentrated in dimensions L&M's index does not capture.



The current debate (iii.a): Treisman's (2023) historical perspective

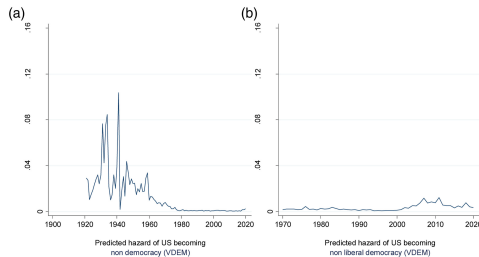
- The *proportion* of democracies is at or near an all-time high; recent declines bring us back to late-1990s levels.
- The *quality* of liberal democracies has fallen modestly ($\sim 3\%$ on V-Dem since 2010) – not catastrophically.
- Recent breakdowns are concentrated in *Third-Wave* democracies with lower incomes and less consolidated institutions.
- Among V-Dem's 11 “liberal-democracy backsliders”, three (Poland, Hungary, Mauritius) crossed thresholds on political institutions; the other eight on legal-transparency subcomponents.



The current debate (iii.b): Treisman — what predicts breakdown?

Survival models on V-Dem / Polity / FH / LIED.

- Robust predictors of breakdown: *lower income, shorter democratic experience, higher prior autocratisation rate among neighbours.*
- Falling popular support for democracy correlates with breakdown in *electoral* democracies; **no relationship** in liberal democracies (replicating Claassen 2020).
- Hungary, Poland, USA, Spain, CZ: support for democracy actually *rose* 2009–2017 while democracy scores fell.
- Models predict actual breakdown events well out-of-sample (AUC-ROC 0.79–0.98).
- For a country with US-level income and democratic experience, the posterior breakdown hazard is *very low*.



Przeworski's warning about extrapolating from history

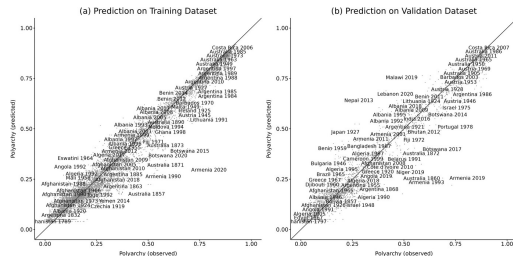
The 1930-33 period in Weimar was like a coin toss. By 1930 there was not a single precedent of a democratic regime that had lasted 11 years and fell. History could not have been a guide. But by now we have a lot of history behind us: this is why I, among many others, was able to conduct statistical studies of breakdowns of democracy. And the lesson of history was that the probability that democracy would fall in a country like the United States was practically zero. So perhaps now lessons from history only mislead. There is a Polish proverb which says "Pessimism is but informed optimism." But informed by what?

Now I need to put on the garb of a political scientist. There are, I think, three ways to confront the failure of the statistical results. One is to shrug shoulders: extremely unlikely events are not impossible and they may occur, so the US is just a fluke. The second one is to suspect that all the statistical models we have been using are conditioned on something we did not include or observe. The third one is that we have correctly identified the conditions that matter but have never encountered their particular combination that occurs now in the US. I lean toward the last possibility. When I regress breakdowns of democracy on per capita income, the number of past partisan alternations in office during the current spell of democracy, and income share of the top 1% recipients, I predict that the probability of it breaking down in the US is of the order of 1 in 2.6 million country years. When I introduce the interaction of the three variables, this probability jumps to 1 in 263 country years, by the order of 10,000. It is still very low but no longer as surprising: I would not want to get into a plane that has a 0.0038 chance of crashing. The exact numbers depend on the period under consideration, data sources, and the estimators so they indicate only orders of magnitude. But my intuition is that what is happening in the US is a consequence of an extremely rare, perhaps never encountered before, combination of historical conditions. This is why lesson from history fail and why we, as observers of the ongoing developments, are unable to predict where they will end.

Substack Post, April 2025

Beyond the dichotomy: Observable-to-Subjective Mapping (OSM) via machine learning (Weitzel et al. (2026))

- Train a random forest to predict V-Dem's *Polyarchy* (continuous) using only *observable* indicators (turnover periods, vote shares, suffrage, multipartyism. . .).
- Out-of-sample fit: $R^2 = 0.95$. The original V-Dem index is largely *reconstructible* from observables alone.
- Generalises across Polity2, UDS, FH, BMR – though FH shows time breakpoints consistent with coding-protocol shifts.
- Some evidence of ideological coder bias: right-wing heads of government coded ~ 0.5 –1 SD lower on Polyarchy / UDS / Polity2 than observables predict. No such bias for FH.



Where does this leave us?

1. **The aggregate global picture is modest.**

Population-weighted V-Dem shows decline; objective indicators and the proportion of democracies do not move much. The claim of “mass backsliding” seems too strong.

2. **But backsliding is real.**

~ 32 countries (V-Dem) show statistically discernible 10-year declines, including Hungary, Poland, Turkey, India, Brazil, Serbia, the United States. Weitzel et al. (2025) confirm that the V-Dem *electoral* signal can be reconstructed from observables — unlikely just an artefact of coder bias.

3. **Steepest decline where measurement is most contested.**

Civil liberties, civil society, freedom of expression, judicial independence — precisely the dimensions outside L&M's index. The measurement debate *is* a conceptualisation debate: minimalist conceptions register less decline by construction.

4. **Historical base rates remain reassuring for established democracies (but for how long?).**

Treisman: hazard of breakdown for high-income, long-experienced democracies is small. The Third-Wave democracies regressing today were always the most fragile. “It can happen here” is possible but not the most likely outcome.

⇒ **If backsliding is real where it is real, the puzzle is: why do voters re-elect those who do it?**

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

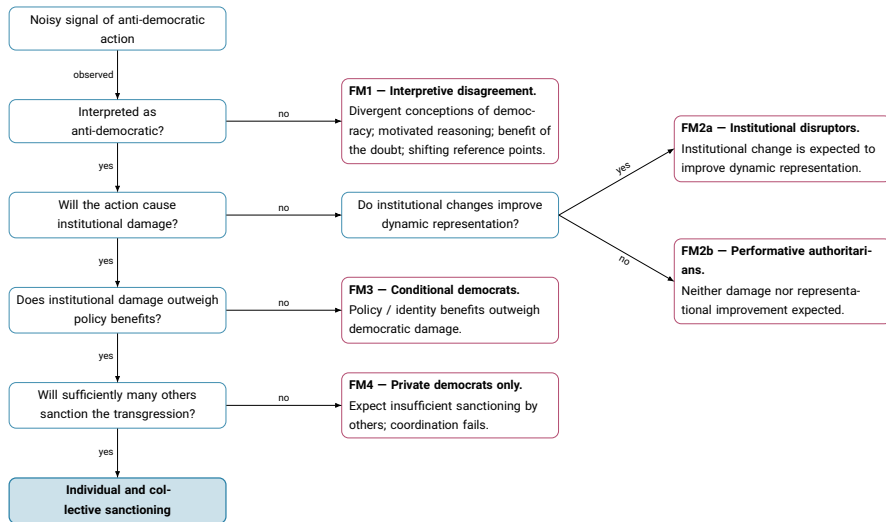
The puzzle: why doesn't (electoral) accountability save democracy?

The hope. If would-be autocrats need elections to take and to keep power, voters should be able to punish them. Backsliding by *elected* leaders ought to be self-limiting — the ballot box does the work.

Discussion.

- In a case you know (Hungary, Turkey, Poland, the US, India, ...), what was the *most important reason* the ballot box failed to stop the backslider?
- Is the failure best understood as one of *information, preferences, or coordination*?
- Can you name a case where accountability *did* work — where voters unseated a backsliding incumbent before serious damage was done? What was different?

An overview of the demand-side literature



Mapping voters' conceptions of democracy

The first failure mode (“FM1”) concerns *interpretation*. Before voters can punish an action, they have to interpret it as a violation — which depends on what they take democracy to mean.

Three families of conceptions in the empirical literature (Gidron et al., 2025; Wunsch, Jacob, and Derksen, 2025):

- **Liberal** — constraints: courts, minority protections, media independence, rule of law.
- **Majoritarian** — responsiveness: unconstrained implementation of the popular will (Grossman et al., 2021).
- **Substantive** — outcomes: social rights, redistribution, physical security.

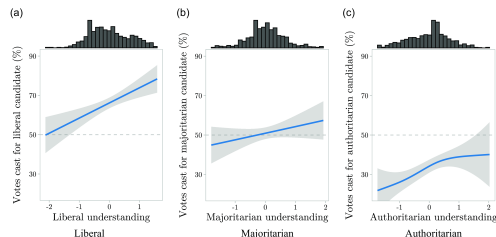
Wunsch, Jacob, and Derksen (2025). Across a large multi-country survey, all three families command non-trivial mass support *within every country studied* — liberal democracy is not the modal conception anywhere. Heterogeneity is at least as large *within* as *between* countries.

How divergent conceptions impede sanctioning

The mechanism. An act counts as a "democratic violation" only *relative to a conception*. Voters whose conception does not include the rule the incumbent has just broken do not register a violation — and so have nothing to sanction.

Wunsch, Jacob, and Derksen (2025).

- Voters who emphasise *majoritarian* or *substantive* elements of democracy are systematically more tolerant of *liberal*-democratic violations (judicial constraints, media independence, minority protections).
- Supporters of backsliding parties are disproportionately drawn from these conception clusters — it is not that they assign low weight to democracy, but that they classify the act differently.
- Conditional on conception, the supporter / opponent gap in evaluating the same violation *shrinks markedly*.



Examining the implications for sanctioning backsliding via a toy model

Goal. Pin down formally when divergent conceptions impede *collective* sanctioning.

Primitives.

- **Population.** Continuum of voters of unit mass. Share $\pi \in (0, 1)$ hold a *minimalistic* conception M ; share $1 - \pi$ hold a *substantive* conception S .
- **Conceptions.** An action is classified as a democratic violation iff it crosses the conception's threshold. By assumption the violation sets are nested, $V_M \subset V_S$: anything M flags, S flags too.
- **Actions.** The incumbent takes one of two actions $\omega \in \{\text{ambiguous, clear}\}$:
 - *Ambiguous* – violates S but not M (e.g., a contested judicial appointment; a media-regulation tweak with policy cover).
 - *Clear* – violates both M and S (e.g., refusing to leave office; jailing the opposition leader).
- **Decision rule.** Voter i sanctions iff her conception classifies the action as a violation.
- **Collective threshold.** The incumbent is removed iff the mass of sanctioners is at least $\theta \in (0, 1)$.

The “marginal conception” determines overall sanctioning capacity

- **Clear act.** Both M and S classify it as a violation; sanctioning mass = $1 > \theta$. Incumbent *always* removed.
- **Ambiguous act.** Only S -voters classify it; sanctioning mass = $1 - \pi$. The incumbent is removed iff

$$1 - \pi \geq \theta \iff \pi \leq \pi^* \equiv 1 - \theta.$$

Regime	Condition	Sanctioning capacity	Marginal conception
Substantive coalition suffices	$\pi \leq \pi^*$	ambiguous <i>and</i> clear acts	S
Minimalists are pivotal	$\pi > \pi^*$	clear acts only	M

The result. As π rises past the critical mass $\pi^* = 1 - \theta$, the binding (marginal) conception jumps discontinuously from S to M , and the population loses the ability to sanction ambiguous acts (Goetz Briefs’ “Grenzmoral”).

Why this matters

1. **The equilibrium logic of gradualism.** An incumbent whose anti-democratic steps all fall inside the ambiguous range (above τ_S , below τ_M) cannot be sanctioned whenever $\pi > \pi^*$. Heterogeneity in conceptions therefore becomes a *strategic resource* the autocrat exploits: rather than test the limits of the binding minimalist conception, design every step to land in the gap (τ_S, τ_M) , where substantivists perceive a violation but are too few to remove him.
2. **Cross-country variation without preference variation.** Two countries with the same average level of democratic support, but different distributions of conceptions ($\pi_A \leq \pi^* < \pi_B$)m sanction the same ambiguous act differently.

Comparative statics. The regime is determined by the single inequality $1 - \pi \geq \theta$.

- **More minimalists** ($\pi \uparrow$): the substantive coalition shrinks; once π crosses π^* , ambiguous acts become unsanctionable – with no change in any voter's commitment to democracy.
- **Tougher coordination** ($\theta \uparrow$): $\pi^* = 1 - \theta$ shrinks.
- **Substitutability.** Either can undermine sanctioning capacity.

Endogenous conceptions of democracy and motivated reasoning

The question. The toy model treated π as exogenous. But are conceptions of democracy really fixed prior dispositions, or do they shift in response to what one's own party does?

The design (Krishnarajan, 2023). Respondents in several Western democracies rate identical hypothetical incumbent actions — restricting media licences, court packing, ignoring an unfavourable court ruling — with the partisan identity of the actor randomly varied. Respondents are then also asked, in the abstract, what does and does not count as democratic.

Main result — conceptions are endogenous to partisanship.

- Identical actions are rated as *more compatible with democracy* when the actor is a co-partisan.
- The shift concerns not just the evaluation of the act, but also respondents' abstract conceptions of democracy, which adjust to accommodate co-partisan behaviour.
- The conception of democracy changes to “rationalise” the action.

A model of motivated reasoning: Little, Schnakenberg & Turner

The question. Krishnarajan documents that conceptions of democracy adjust to fit co-partisan actions. Little, Schnakenberg, and Turner (2022) examine the consequences of motivated reasoning for electoral accountability (recall lecture 1).

The set-up. Voters care about two things:

- **Accuracy** — the standard Bayesian motive: beliefs about the incumbent should track the truth.
- **Directional utility** — a non-Bayesian motive: beliefs should accord with partisan attachments.

A model of motivated reasoning: Little, Schnakenberg & Turner

The question. Krishnarajan documents that conceptions of democracy adjust to fit co-partisan actions. Little, Schnakenberg, and Turner (2022) examine the consequences of motivated reasoning for electoral accountability (recall lecture 1).

The set-up. Voters care about two things:

- **Accuracy** — the standard Bayesian motive: beliefs about the incumbent should track the truth.
- **Directional utility** — a non-Bayesian motive: beliefs should accord with partisan attachments.

Two distinct effects in equilibrium.

- **Divergence** — voters with different affinities form *different* conclusions about the same incumbent.
- **Desensitisation** — any given voter's conclusion becomes *less responsive* to new performance information. The *within-voter* effect: the slope of beliefs on signals flattens.

Key claim. Desensitisation is the more pernicious of the two. Divergence alone can leave incumbent effort unchanged; desensitisation always undercuts it, and it bites every voter, including centrists — which is what makes it both general and (empirically) under-studied.

The gradualism puzzle and reference-dependent preferences

The empirical puzzle. Contemporary backsliding is almost always *gradual*. There has been a shift, with *executive aggrandisement* – slow, incremental, legalistic erosion – having replaced *coups* and the *electoral fraud* as the dominant mode of democratic decline since the 1990s (Bermeo, [2016](#); Huq and Ginsburg, [2018](#); Levitsky and Ziblatt, [2018](#); Scheppele, [2018](#)).

The puzzle. If voters' preferences over democracy are stable, only the cumulative decline should matter. Either voters care enough to sanction the trajectory (and the autocrat is removed at some point along the path), or they do not (and the path is sustainable from step 1).

The gradualism puzzle and reference-dependent preferences

The empirical puzzle. Contemporary backsliding is almost always *gradual*. There has been a shift, with *executive aggrandisement* – slow, incremental, legalistic erosion – having replaced *coups* and the *electoral fraud* as the dominant mode of democratic decline since the 1990s (Bermeo, 2016; Huq and Ginsburg, 2018; Levitsky and Ziblatt, 2018; Scheppele, 2018).

The puzzle. If voters' preferences over democracy are stable, only the cumulative decline should matter. Either voters care enough to sanction the trajectory (and the autocrat is removed at some point along the path), or they do not (and the path is sustainable from step 1).

The building block: Köszegi–Rabin reference-dependent preferences. Voters evaluate outcomes not against an absolute, stable standard but against an *endogenously-formed reference point*.

- **Reference-dependence** – utility depends on the *gap* between the outcome and the reference point, not on the outcome's absolute level.
- **Loss aversion** – outcomes below the reference point hurt more than equivalent gains feel good.

Preview of Grillo and Prato (2023). Each incremental backsliding step lowers what voters expect the next step to deliver. *Gradualism is the rational way for an autocrat to exploit endogenous reference points.*

Grillo & Prato: set-up

Players.

- A continuum of citizens with ideology θ_i drawn from F , uniform on $[-\frac{1}{2\psi}, \frac{1}{2\psi}]$ with $\frac{1}{2\psi} > 1$ (so a majority – but not all – oppose backsliding; lower $\psi \Rightarrow$ greater *mass polarisation*).
- An incumbent with privately-observed ideology $\theta_I \sim F_I$, uniform on $[\tau - \frac{1}{2\phi}, \tau + \frac{1}{2\phi}]$ with $\tau \in (0, 1)$ (lower $\phi \Rightarrow$ greater *voter uncertainty* about θ_I).

Incumbent's two-stage choice.

1. **Challenge:** $c \in \{0, 1\}$ – whether to attack democratic norms.
2. **Escalation:** $d \in [\delta, 1]$ if $c = 1$ – how far to “double down”.

Denote the resulting outcome by $\mathbf{q} = (c, d)$. The parameter $\delta \in (0, 1)$ captures *institutional checks and balances*: lower $\delta \Rightarrow$ stronger checks (the incumbent cannot easily retreat all the way).

Grillo & Prato: set-up

Players.

- A continuum of citizens with ideology θ_i drawn from F , uniform on $[-\frac{1}{2\psi}, \frac{1}{2\psi}]$ with $\frac{1}{2\psi} > 1$ (so a majority – but not all – oppose backsliding; lower $\psi \Rightarrow$ greater *mass polarisation*).
- An incumbent with privately-observed ideology $\theta_I \sim F_I$, uniform on $[\tau - \frac{1}{2\phi}, \tau + \frac{1}{2\phi}]$ with $\tau \in (0, 1)$ (lower $\phi \Rightarrow$ greater *voter uncertainty* about θ_I).

Incumbent's two-stage choice.

1. **Challenge:** $c \in \{0, 1\}$ – whether to attack democratic norms.
2. **Escalation:** $d \in [\delta, 1]$ if $c = 1$ – how far to “double down”.

Denote the resulting outcome by $\mathbf{q} = (c, d)$. The parameter $\delta \in (0, 1)$ captures *institutional checks and balances*: lower $\delta \Rightarrow$ stronger checks (the incumbent cannot easily retreat all the way).

Citizen material utility. With policy outcome $y(c, d) = 1 + cd$,

$$u(\mathbf{q}; \theta_i) = \theta_i y(c, d) - cd = \theta_i + (\theta_i - 1) cd.$$

Citizens with $\theta_i > 1$ favour backsliding; those with $\theta_i < 1$ oppose it.

Incumbent utility. Combines policy utility and popular support: $u_I(\mathbf{q}; \theta_I) = \theta_I(1 + cd) - cd + R \cdot \pi(\mathbf{q})$,

where $\pi(\mathbf{q}) \in [0, 1]$ is the share of citizens supporting the incumbent and $R > 0$ scales the popularity concern.

Reference dependence and the fear-and-relief mechanism

Total utility with reference dependence. Citizens' total utility adds a Köszegi–Rabin gain-loss term:

$$v(\mathbf{q}; \theta_i \mid \underline{u}) = u(\mathbf{q}; \theta_i) + \eta \cdot [u(\mathbf{q}; \theta_i) - \underline{u}], \quad \eta \geq 0.$$

Outcomes above the reference point $\underline{u}$ produce *relief*; outcomes below produce *disappointment*.

Reference dependence and the fear-and-relief mechanism

Total utility with reference dependence. Citizens' total utility adds a Köszegi–Rabin gain-loss term:

$$v(\mathbf{q}; \theta_i \mid \underline{u}) = u(\mathbf{q}; \theta_i) + \eta \cdot [u(\mathbf{q}; \theta_i) - \underline{u}], \quad \eta \geq 0.$$

Outcomes above the reference point $\underline{u}$ produce *relief*; outcomes below produce *disappointment*.

Reference point = rational expectation, formed after c . The reference point is formed *after* citizens observe c but *before* they observe d , and equals the rational expectation of material utility under the incumbent's equilibrium strategy:

$$\underline{u}(c; \hat{\mathbf{q}}, \theta_i) = \mathbb{E}[u(\hat{\mathbf{q}}; \theta_i) \mid c].$$

The fear-and-relief mechanism. After the incumbent challenges ($c = 1$), citizens expect escalation $\underline{d}_1 = \mathbb{E}[d \mid c = 1] > \delta$. The reference point adjusts *downward*. If the incumbent then *partially retreats* ($d = \delta < \underline{d}_1$), realised utility *exceeds* the reference point. Citizens experience a psychological gain (relief) — which can outweigh the material loss from the partial backsliding that did occur.

Personal equilibrium and the benchmark

Personal equilibrium (Köszegi–Rabin). A profile $(\hat{q}, \underline{u}(0; \hat{q}, \theta_i), \underline{u}(1; \hat{q}, \theta_i))$ such that:

(SR) The incumbent's strategy $\hat{q}$ is sequentially rational given citizens' reference-dependent support.

(RE) Each citizen's reference point equals her rational expectation of material utility under $\hat{q}$.

Benchmark without reference dependence ($\eta = 0$, **Proposition 2**). Define

$$\theta^{\dagger} \equiv 1 + \frac{R\psi}{2}.$$

- *Extreme autocrats* ($\theta_I > \theta^{\dagger}$): $c = 1, d = 1$ – challenge and fully escalate.
- *Restrained autocrats* ($\theta_I \in (1, \theta^{\dagger}]$): $c = 0$ – would prefer to backslide but are deterred by popularity loss.
- *Democrats* ($\theta_I \leq 1$): $c = 0$.

In this benchmark, *whoever challenges fully escalates*. Backsliding is only ever attempted by extreme types – and once attempted, is total.

The opportunistic-authoritarian result

Key condition (Proposition 4). When reference dependence is strong enough,

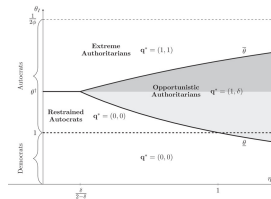
$$\eta \geq \frac{\delta}{2 - \delta},$$

the equilibrium changes qualitatively. There exist cut-offs $\underline{\theta} < \theta^\dagger < \bar{\theta}$ with:

- *Extreme authoritarians* ($\theta_I > \bar{\theta}$): $c = 1, d = 1$.
- *Opportunistic authoritarians* ($\theta_I \in (\underline{\theta}, \bar{\theta}]$): $c = 1, d = \delta$ – challenge then *partially retreat*.
- Otherwise ($\theta_I \leq \underline{\theta}$): $c = 0$.

Comparative statics.

- $\eta \uparrow$ (stronger reference dependence): opportunistic interval $(\underline{\theta}, \bar{\theta}]$ expands.
- $\delta \downarrow$ (stronger checks): *ambiguous* – less material damage per retreat, but more relief.
- $\psi \downarrow$ (higher polarisation): can simultaneously decrease frequency and increase severity of backsliding.
- $\phi \downarrow$ (higher voter uncertainty about θ_I): more opportunistic types.



Two surprising comparative statics: checks and polarisation

Stronger checks ($\delta \downarrow$).

- *Material* (intuitive): lower δ caps the damage from partial retreat.
- *Relief* (counterintuitive): lower δ widens the gap $\underline{d}_1 - \delta$ that determines the relief payoff $\Rightarrow$ fear-and-relief becomes *more* profitable.
- *Net*: stronger checks reduce severity per episode but expand the set of opportunists. Frequency can rise even as severity falls.

More polarisation ($\psi \downarrow$).

- *Svolik* (intuitive): lower ψ shrinks $\theta^\dagger$; more extreme types find full escalation worth the support loss. Pushes toward *more full escalation*.
- *Responsiveness* (subtle): lower ψ also makes citizens less responsive to incumbent behaviour. The relief mechanism delivers less; *fewer opportunists*.
- *Net*: polarisation shifts the *composition* of backsliders from opportunists ($d = \delta$) to extreme authoritarians ($d = 1$). Frequency falls; severity rises.

Two empirical predictions of the model. (i) Stronger institutional checks $\Rightarrow$ *more* attempts at backsliding but *less* severity per attempt – against the Madisonian intuition. (ii) Polarisation and backsliding should be *negatively* correlated for mild episodes but *positively* correlated for severe ones.

Backsliding as a corrective to systemic rot

The shift from FM1 to FM2. Voters *do* perceive an act as anti-democratic — but support it anyway, here for institutional reasons (FM2a).

Institutions are captured. The populist analysis:

- Counter-majoritarian institutions — courts, central banks, party machines, legacy media — are interpreted as the instruments for the entrenchment of elites (Gratton and Edenhofer, [2025](#)).
- Conventional democratic competition is incapable of dislodging entrenched interests.

The instrumental step: backsliding as a corrective.

- Only by weakening or bypassing the captured institutions can space be cleared for better representation. “Backsliding” is construed as the precondition for an institutional reset, as it were.

Stylised example. Milei supporters may not deny that emergency decrees and the bypassing of Congress violate democratic norms. They support them anyway because they see Argentine political institutions as so captured by Peronist patronage networks that aggressive disruption is the only credible route to reform.

Discriminating predictions and scope conditions

The empirical challenge. Detect voters who (i) recognise the act as anti-democratic, (ii) support it anyway, and (iii) do so on *instrumental* grounds — not by redefining democracy (FM1), trading it off for policy (FM3), or doubting damage (FM2b).

Discriminating predictions.

- *vs FM1.* Does the voter classify the act as anti-democratic?
- *vs FM2b.* Expected institutional consequences: FM2a predicts directional change; FM2b predicts “not much will happen”.
- *vs FM3.* Justification: FM2a invokes *institutional* improvement.
- *vs partisanship.* The FM2a voter accepts the same logic when applied against her own side.

Scope conditions.

1. Visible recent history of elite capture — the diagnosis needs a credible target.
2. **Macro-level corruption that extends to the voter's own coalition** — without this, FM2a collapses into rationalised partisanship.

Backsliding as expression: institutional damage in the error term

For FM2b voters, support is *expressive*: the benefit they get from endorsing the act is symbolic.

What makes institutional damage low-salience.

- *Perceived resilience*. Voters have an implicit Treisman base rate: in consolidated democracies, breakdown hazards are very low. Courts hold; federalism contains; civil society mobilises. “It can’t happen here.”

The expressive benefit. Support delivers identity expression, anti-establishment signalling, solidarity with a charismatic leader, or the pleasure of seeing despised institutions / opponents discomfited. None of these require institutional consequences to materialise.

Stylised example. The component of Trump-era support driven by cultural signalling — “owning the libs”, defying conventional political decorum, breaking from elite norms — delivered expressive benefit independently of policy delivery or institutional consequences. Some voters acknowledged norm violations *as such* but treated their downstream institutional effects as a secondary concern.

Conditional democrats

The shift from FM1 and FM2 to FM3. FM3 voters *do* perceive the act as anti-democratic and *do* expect institutional damage — but support the incumbent anyway, because the policy or identity benefits outweigh the democratic cost.

The trade-off. Voter utility has two components: democracy gets positive but finite weight δ ; policy / identity gets weight 1. The voter supports the violator when policy benefit $> \delta \times$ democratic cost.

Two implications.

- δ is the primitive Svolik, [2020](#) will pin down — the “civic virtue” parameter.
- As polarisation rises, the policy stakes from losing rise, and the trade-off tilts against democracy.

The puzzle. Mass publics consistently profess strong support for democracy in surveys — yet re-elect incumbents who subvert it. Conventional “support for democracy” indices fail to predict actual behaviour. The model by Svolik helps to explain why.

Svolik: set-up

Players.

- Incumbent A and challenger B , with policy ideals $\theta_A > 0 > \theta_B$.
- A continuum of voters: *informed* (share $1 - \alpha$) vote on platforms and electoral fairness; *uninformed* (share α) swayed by manipulation alone.

Voter preferences (informed voters). A voter with ideal point x_i receives

$$u_i(x_j, \mu) = \begin{cases} -(x_i - x_A)^2 - \delta\mu^2 & \text{if incumbent wins,} \\ -(x_i - x_B)^2 & \text{if challenger wins,} \end{cases}$$

where $\mu \geq 0$ is the incumbent's manipulation of the electoral process and $\delta \geq 0$ is the *civic virtue* parameter – voters' sensitivity to electoral malpractice.

Svolik: set-up

Players.

- Incumbent A and challenger B , with policy ideals $\theta_A > 0 > \theta_B$.
- A continuum of voters: *informed* (share $1 - \alpha$) vote on platforms and electoral fairness; *uninformed* (share α) swayed by manipulation alone.

Voter preferences (informed voters). A voter with ideal point x_i receives

$$u_i(x_j, \mu) = \begin{cases} -(x_i - x_A)^2 - \delta\mu^2 & \text{if incumbent wins,} \\ -(x_i - x_B)^2 & \text{if challenger wins,} \end{cases}$$

where $\mu \geq 0$ is the incumbent's manipulation of the electoral process and $\delta \geq 0$ is the *civic virtue* parameter – voters' sensitivity to electoral malpractice.

Polarisation. A fraction $(1 - \pi)$ of ideal points are uniform on $[-\frac{1}{2}, \frac{1}{2}]$; the remaining π split equally at $\pm\frac{1}{2}$. Higher $\pi \Leftrightarrow$ more bimodal electorate.

Manipulation technology. The incumbent chooses $\mu \in [0, 1]$. He gains share $(1 + \mu M + \epsilon)/2$ of the uninformed vote, with M scaling effectiveness and $\epsilon \sim U(-\sigma, \sigma)$ a small shock.

Timing. Candidates simultaneously announce x_A, x_B ; incumbent picks μ ; shock ϵ realised; voters vote; winner implements his platform.

Svolik: the swing voter's trade-off

The informed voter's choice. For platforms $x_A > x_B$ and manipulation μ , informed voter i prefers the incumbent iff

$$-(x_i - x_A)^2 - \delta\mu^2 \geq -(x_i - x_B)^2.$$

The swing voter. The voter for whom this binds with equality is at

$$x_S(\mu) = \underbrace{\frac{x_A + x_B}{2}}_{\text{midpoint}} + \underbrace{\frac{\delta\mu^2}{2(x_A - x_B)}}_{\text{shift toward incumbent}}.$$

Without manipulation ($\mu = 0$), x_S is exactly at the midpoint. As μ rises, x_S shifts *toward* the incumbent's ideal: voters who would prefer the incumbent on policy alone are put off enough by manipulation to switch to the challenger instead.

Svolik: the swing voter's trade-off

The informed voter's choice. For platforms $x_A > x_B$ and manipulation μ , informed voter i prefers the incumbent iff

$$-(x_i - x_A)^2 - \delta\mu^2 \geq -(x_i - x_B)^2.$$

The swing voter. The voter for whom this binds with equality is at

$$x_S(\mu) = \underbrace{\frac{x_A + x_B}{2}}_{\text{midpoint}} + \underbrace{\frac{\delta\mu^2}{2(x_A - x_B)}}_{\text{shift toward incumbent}}.$$

Without manipulation ($\mu = 0$), x_S is exactly at the midpoint. As μ rises, x_S shifts *toward* the incumbent's ideal: voters who would prefer the incumbent on policy alone are put off enough by manipulation to switch to the challenger instead.

The trade-off. A voter just to the right of $x_S(\mu)$ faces a choice between

- *punishing manipulation* – and getting policy x_B , which she dislikes;
- *tolerating manipulation* – and keeping policy x_A , which she prefers.

Where polarisation enters. As π rises, voter ideal points cluster at the extremes $\pm \frac{1}{2}$. The incumbent's “core” supporters – voters far from the challenger's platform x_B – have a policy distance too large to swallow even for substantial manipulation. The marginal voter at $x_S(\mu)$ becomes a smaller share of the electorate, and the incumbent can manipulate more without losing the median.

Svolik: the incumbent's calculus

What the incumbent maximises. The incumbent picks μ to maximise his probability of victory. With a small symmetric electoral shock, that probability is monotone in the deterministic vote margin

$$\underbrace{\alpha\mu M}_{\text{gain from uninformed}} - \underbrace{2(1-\alpha)(1-\pi)x_S(\mu)}_{\text{loss from informed}}.$$

Two countervailing effects of raising μ .

- *Marginal benefit.* A unit of manipulation increases the incumbent's vote-share advantage among uninformed voters by αM — a clean linear gain.
- *Marginal cost.* The same unit shifts the swing voter $x_S(\mu)$ further toward the incumbent. Informed voters between the midpoint and $x_S(\mu)$ defect. The cost is *convex* in μ (since $x_S \propto \mu^2$): the first unit of manipulation is cheap, later units expensive.

The polarisation result. The marginal cost scales with $(1-\pi)$: the higher the polarisation, the fewer informed voters sit in the defector band, and the smaller the marginal cost of manipulation. Equilibrium μ^* equates marginal benefit and marginal cost — so as π rises the cost curve falls, and μ^* rises with it. The closed form on the next slide makes this precise.

Svolik: equilibrium and comparative statics

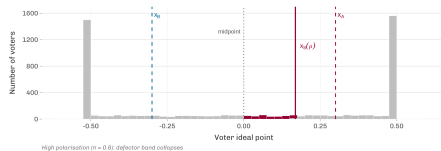
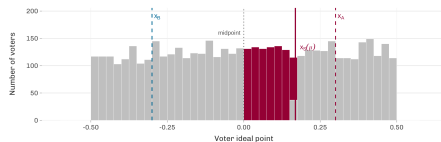
Equilibrium manipulation. For fixed platforms, the incumbent's optimal level of manipulation is

$$\mu^* = \frac{\alpha}{1 - \alpha} \cdot \frac{M x_A}{\delta (1 - \pi)}.$$

Comparative statics. μ^* rises in

- *polarisation* π – the central result. As voters sort into the extremes, the mass in the defector band collapses (figure, right), and the incumbent can manipulate more without losing the median;
- *share of uninformed* α – more voters who do not detect manipulation;
- *manipulation technology* M – a more effective tool;
- *incumbent extremism* x_A – larger policy stakes for the incumbent's base.

And falls in δ – civic virtue is the only check on manipulation in this model.



Svolik (2020): the swing voter and the role of polarisation

The dynamic trade-off: liberty against security

Gratton and Lee (2024): A forward-looking, infinitely-lived voter chooses each period between

- a **liberal** government — bound by the constitution; delivers liberty and transparent information, but cannot respond to security shocks of probability;
- an **illiberal** government — breaches the constitution; delivers security *and* engages in Bayesian persuasion (Kamenica and Gentzkow, 2011) over what the voter learns.

Key twist. Censorship means there is an **accountability cost** for illiberal incumbents. Even rational voters may be *persuaded* to re-elect them when, under full information, they would not. The voter accepts this in exchange for greater security.

Equilibrium taxonomy. Four regimes — stable liberal democracy; *inefficient* stable illiberal democracy; cycling liberal–illiberal; and capture into autocracy. Voters go for illiberal candidates when expected security costs are high and may be *kept* there by manipulation even after the crisis has passed.

Stronger checks and balances: a double-edged sword

A central comparative static in Gratton and Lee (2024): tightening constraints on the executive has **two opposing effects** on the conditional democrat.

- + **Slow-down effect.** Stronger constraints make it harder for an elected illiberal government to consolidate into outright autocracy.
- **Demand-inducing effect.** Stronger constraints *also* reduce a liberal government's capacity to deliver security; voters, anticipating this, become **more** willing to elect an illiberal government in the first place.

Implication. The institutional fix that makes autocratisation slower can simultaneously make it more likely to begin. Excessive constraints — by hobbling the liberal alternative — feed precisely the conditional-democrat logic of FM3.

Empirical resonance. Speaks to Howell, Shepsle, and Wolton (2019) on the case for executive flexibility; cautions against assuming "more veto points $\Rightarrow$ more democratic resilience".

Conclusion. The conditional-democrat trade-off is **endogenous to institutional design**: what the constitution permits a liberal government to do shapes how attractive the illiberal alternative looks.

Graham & Svolik 2020: the canonical conjoint design

Design. Respondents (US, $N \approx 1,500$) see pairs of hypothetical candidates with attributes varying along party, policy positions, demographics, and *one democratic violation* drawn from a fixed menu (gerrymandering, voter suppression, executive overreach, jailing journalists, ignoring court rulings). They choose which to vote for. The conjoint structure identifies the *causal* effect (AMCE) of a violation on vote choice, independent of party and ideology.

Main findings.

- Co-partisan candidates with democratic violations lose only ≈ 11.7 percentage points of vote share — a small penalty given the explicit anti-democratic framing.
- Only $\approx 13.5\%$ of voters *always* punish democratic violations by their own party — the “true democrats”.
- Punishment is concentrated among *moderates* with weaker partisan attachment.
- Punishment varies sharply by type of violation: jailing journalists is punished much more than gerrymandering.

Has become the workhorse. The conjoint isolates the democratic-norms attribute from policy and partisan confounders. Albertus and Grossman, [2021](#); Grossman et al., [2021](#), and most subsequent empirical work on FM3 build on this design — some extending it cross-nationally, others adding perception or sequence dimensions.

Cross-national replications

Albertus and Grossman, 2021. Replicate the Graham–Svolik design across the US, Argentina, Brazil, and Mexico. Most respondents identify transgressions and overall support for violations is low. But 10–35% support *clear* democratic violations, with larger co-partisan tolerance for right-wing incumbents (US, Brazil). Punishment is rare even when voters disapprove – a finding that anticipates FM4.

Neundorff et al., 2026. $N \approx 35,000$ across 32 countries (including 16 autocracies). Forced-choice conjoint with seven attributes: elections, executive constraints, free speech, economic prosperity, crime, gender equality, ethnic diversity.

- Citizens defend *free elections* sharply – the most salient democratic feature, even in autocracies.
- *Executive constraints* are negotiable: citizens trade them for economic prosperity even in established democracies.

The pattern. Across countries and designs, the FM3 trade-off survives. Voters value democracy; they just do not value it enough to override sufficiently strong policy or economic considerations.

Jacob, 2025: the sequence design

The design. Survey experiment in Poland following the 2020 presidential election ($N \approx 2,900$) reproducing the temporal logic of an actual electoral cycle: voters first choose between two candidates; the chosen candidate's behaviour in office is then revealed (attacks on electoral institutions, or refrains from doing so); voters then face a follow-up election and either confirm or withdraw support.

Difference to Graham & Svolik. Standard conjoints collapse evaluation and choice into a single task. Jacob separates them: respondents first *evaluate* the conduct (is it anti-democratic?), then decide whether to *revise* their vote.

Main finding. Polish citizens *firmly disapprove* of attacks on electoral institutions, regardless of party. But the evaluation *does not translate into behaviour*: no segment of the electorate – not the politically educated, not liberals, not moderate partisans – withdraws electoral support from a co-partisan who attacks democratic norms more than the average voter does.

Why FM3 alone does not fully explain this. The trade-off in FM3 explains why voters who value democracy nonetheless tolerate violations – democracy is outweighed by policy. But Jacob's design holds policy distance roughly constant, so the FM3 trade-off should be *weak* here. The persistent gap between disapproval and action points elsewhere: voters who would prefer to sanction do not, because they expect no coordinated response. This is the FM4 coordination story we turn to next.

Private democrats and the coordination problem

FM4 isolates a quite different failure mode. Suppose **everyone privately values democracy** and would prefer it to survive. Backsliding can still succeed, because resisting it is costly and is individually rational **only if enough others resist**.

- Cost of action (turning out, protesting, refusing the order, defecting from the coalition) is borne privately.
- Benefit – democratic survival – is a public good that materialises only above a participation threshold.
- If I expect others not to resist, my best response is also not to resist.

Four ways this coordination can break – the structure of the rest of the block:

- (A) **Focal-point failure**: no salient violation triggers common knowledge of resistance.
- (B) **Information manipulation**: the incumbent undermines common knowledge.
- (C) **Salami tactics**: each move is too small to coordinate around.
- (D) **Breakdown of cross-cutting cleavages**: polarisation destroys solidarity across groups.

Democracy as a coordination equilibrium

Weingast (1997) models the rule of law as a coordination game between two citizen groups facing a transgressing sovereign. Resistance is costly; success requires **joint** resistance.

	<i>B</i> resists	<i>B</i> acquiesces
<i>A</i> resists	rule of law preserved	<i>A</i> alone purged
<i>A</i> acquiesces	<i>B</i> alone purged	sovereign transgresses; both lose

Two Pareto-ranked equilibria: **joint resistance** (democracy survives) and **joint acquiescence** (sovereign untrammelled). The constitution selects between them by acting as a **focal point** – telling *A* and *B* what counts as a violation worth resisting. Hadfield and Weingast (2014): the rule of law is a **classification institution** that supplies the shared categories needed for *A* and *B* to recognise the same act as a violation.

Implication. Backsliding can reflect a fragmented “classification” system, an opportunistically small transgression, or simply pessimistic beliefs about what others will do.

Coordination under incomplete information: global games

The Weingast game has **multiple equilibria** — a generic feature of coordination games and the central obstacle to sharp predictions about when collective action succeeds.

Global games (Morris and Shin, 2003) resolve this. The original complete-information game is embedded in a **global family** of games indexed by an unobserved fundamental θ , with each player receiving a small private noisy signal of θ . The "global" in the name refers to this embedding — the analysis is required to be robust across the whole family rather than tied to one payoff configuration — and is what delivers a unique equilibrium in the small-noise limit.

- Each citizen i observes $x_i = \theta + \varepsilon_i$, with ε_i idiosyncratic.
- Action: resist (costly) or acquiesce. Resistance succeeds if a threshold share resist.
- In the small-noise limit, a **unique cut-off equilibrium** obtains: resist iff $x_i \geq x^*$.

Why this is useful for thinking about backsliding.

- Both **transparency of the fundamental** and the precision of private signals shift x^* — institutions and the information environment determine *when* citizens coordinate.
- Edmond (2013) embeds this in a regime-change game where the incumbent **manipulates the public signal** to push x^* above what citizens privately observe.

Salami tactics: gradual erosion

If coordination requires a **focal violation** to trigger common knowledge of resistance, a sufficiently **small** violation cannot be coordinated against — even when everyone privately regards it as anti-democratic. The aggrandising leader chooses each move just below this threshold; the cumulative trajectory is autocratisation.

Empirical observation. Bermeo (2016) documents the post-Cold-War shift away from coups d'état and towards **executive aggrandisement** — incremental constitutional change, court-packing, prosecutorial harassment, electoral re-districting — none of which on its own suffices to mobilise resistance.

Formal foundations.

- Helmke, Kroeger, and Paine (2022) model a leader choosing a sequence of small institutional changes; citizens hold a threshold belief about what justifies resistance; equilibrium steps lie just below it.
- Luo and Przeworski (2023): a **strategy of stealth** in which the optimal trajectory is calibrated so that no individual step generates the common knowledge needed to coordinate.

Tension with FM4.3–4.4. Global games predict coordination on the *fundamental* θ ; salami tactics show how the leader endogenously manipulates the *observability* of θ .

Polarisation and coordination

Weingast's two-group setup assumes A and B can *in principle* coordinate. **Polarisation** destroys this presumption by making within-cleavage solidarity so much stronger than cross-cleavage commitment to democracy that the joint-resistance equilibrium becomes unattainable.

Why liberal democracy is rare. Mukand and Rodrik (2020) argue that liberal democracy requires an alliance between the **non-elite majority** and an **elite minority** against the majority-elite coalition. The alliance is fragile: it dissolves whenever non-elites perceive the minority elite as a closer threat than the majority elite, or vice versa.

Polarisation as the destroyer.

- Svoboda (2019): under high polarisation, voters refuse to punish co-partisan violations because the out-party is perceived as the greater threat — pre-empting joint resistance.
- Graham and Svoboda (2020) (cross-referenced from FM3): the ≈ 11.7 pp average penalty for norm violations masks a small "always punish" minority and a large pool willing to forgive co-partisans — the cross-cleavage breakdown in survey data.

Empirical implication. The same incumbent transgression has very different probabilities of triggering resistance depending on the cleavage structure — predicting more backsliding in deeply polarised polities at *any* given level of θ .

When shared information helps — and when it hurts

Basak, Deb, and Kuvalekar (2026) build a collective-action game with both coordination **and** free-riding. Two groups, each with a signal about the regime's **resilience** (the participation threshold $\bar{n} + 1$); cost is private, benefit public.

The new object: **similarity of information** (concentration along the diagonal, CAD). Information is more similar when, conditional on observing s , you assign higher posterior probability that others observed s too.

Main result. Two regimes:

- **Encouragement environments** (high threshold, resilient regime): more similar information **raises** maximal-equilibrium participation. Pivotality *rises* with others' participation — the coordination motive dominates.
- **Discouragement environments** (low threshold, fragile regime): more similar information **lowers** maximal-equilibrium participation. Pivotality *falls* with others' participation — the free-rider motive dominates.

Why this matters for FM4.

- It **reconciles** Enikolopov, Makarin, and Petrova (2020) (VKontakte $\Rightarrow$ more protest, Russia) with Cantoni et al. (2019) (Hong Kong: learning others participate $\Rightarrow$ free-riding).
- It implies that **press-freedom restrictions** are a double-edged instrument: useful only for resilient incumbents; counter-productive for fragile ones, who may face *larger* protests when they restrict.
- Combined with Gratton and Lee (2024): the same comparative-static logic — that an institutional or informational fix can backfire — recurs on the demand-coordination side.

The subversion dilemma in the lab

Braley et al. (2023) provide the cleanest experimental test of the FM4 mechanism. Three findings.

1. Voters systematically **overestimate** the share of out-partisans willing to subvert democratic norms — typically by a factor of two or more.
2. Their own willingness to subvert is **conditional** on this belief: a subversion dilemma in which each side defects to forestall expected defection by the other.
3. An **ask-tell intervention** — ask respondents what share of out-partisans will subvert, then provide the truth — reduces own willingness to subvert by roughly 5–7 percentage points, with effects persisting in follow-up.

Interpretation.

- Citizens are **private democrats**: most would prefer the joint-resistance equilibrium of Weingast.
- Beliefs about the other side place them in the **joint-acquiescence** equilibrium instead.
- The intervention shifts *beliefs*, not preferences — and that is enough to move behaviour.

Closing out the demand side: Public opinion and the survival of democracy

- **Does mass support actually matter?** Claassen (2020a) – Bayesian latent-variable panel of 135 countries: yes. Support is significantly associated with subsequent democratic change, and especially with the [endurance](#) of established democracies.
- **But it is endogenous – and the measurement is contested.** Claassen (2020b): the relationship also runs the other way – democratic deepening [dampens](#) support, with the backlash concentrated on *counter-majoritarian* features. Tai, Hu, and Solt, 2024: latent-variable estimates are sensitive to modelling choices; treat magnitudes with care.
- **Restoration is feasible.** Frederiksen, Willer, and Petersen (2026), Polish 2023 panel: the election loss did not deepen anti-democratic sentiment on either side. Mass-level radicalisation is not the binding constraint – [elite strategy](#) is.

Conclusion. Mass attitudes do matter, but they are malleable to some extent and are not always a binding constraint. So: [who supplies the backsliding moves, and why?](#)

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

The supply side: An overview

The supply side asks when elites *choose* to provide, accommodate, or block backsliding. Three conditions must hold in sequence for backsliding to advance; **democracy is resilient whenever any one fails**.



Who (1): military officers

Characteristics. Senior officers within the coercive apparatus of the state who use, or credibly threaten, force to remove civilian rule.

Strategic logic.

- Reasons for intervention: disorder or dislike of left/liberal governments
- The military commands cheap access to the technology of regime change: [coercion](#). Coups exploit a low marginal cost of force.

Temporal scope condition. Post Cold War, the international environment (sanctions, regional-organisation clauses) has raised the price of coups while lowering the perceived cost of working through the electoral arena.

- Svoboda (2015): cross-national decomposition shows coups declining and [incumbent takeovers](#) dominating contemporary breakdowns.
- Bermeo (2016): documents the shift from coups towards executive aggrandisement.

Who (2): civilian-populist outsiders

Characteristics. A figure with no career stake in the existing party system who enters competitive politics by mobilising directly to a disaffected mass.

Strategic logic. Two reinforcing forces.

- **No rents to lose.** The outsider has no patronage network, no committee chair, no party machine.
- **Anti-institutional posture can be electorally appealing.** It is the technology by which the outsider builds the mass coalition needed to enter office at all.

The outsider's calculation is [dismantle constraints to win and govern continuously](#).

- Bernhardt, Krasa, and Shadmehr (2022): [economic fragility](#) leads to the selection of demagogic types — the value of an anti-system candidate to voters rises precisely as institutions are perceived to fail.
- Lehmann and Tyson (2021): radicalisation as a strategic *tool* of outsider candidates — the outsider chooses how aggressively to radicalise their base.

Who (3): insider party leaders who capture

Characteristics. An established politician inside a competitive party who consolidates control over it and uses incumbency to constrain the next round of competition — Orbán, Erdoğan, Kaczyński, Modi.

Strategic logic. The hardest case — the insider is, by construction, already succeeding under democracy.

- **Expectation about future stakes.** Losing fair competition can be costly, e.g. policy reversal, prosecution, loss of patronage, return to power of an existentially threatening opposition. When expected stakes of losing exceed expected costs of consolidating, the calculation **flips**.
- **Programmatic incompatibility.** Certain agendas — Orbán's illiberal cultural project, Kaczyński's judicial reconstruction, Erdoğan's religious-nationalist transformation — are not deliverable within liberal-democratic constraints.
- **Gradual erosion.** Once in office the marginal move is small.

Contrast with outsiders. The outsider attacks institutions because they have nothing to lose by doing so; the insider does so because they have a great deal to lose by *not* doing so.

When do challengers emerge?

	Sudden	Gradual
Exogenous	Exogenous shocks 2008 global financial crisis; COVID-19 pandemic; terrorist attacks.	Structural drift Skill-biased technological change; long-run demographic and value change; declines in trust.
Endogenous	Cascades and tipping points Corruption mega-scandals; sudden constitutional crises.	Policy drift Rising income and wealth inequality; long-term corruption; elite norm decay.

When do challengers emerge (1): exogenous shocks

Claim. Sudden adverse shocks — in particular deep recessions — raise the probability of democratic backsliding. The relationship is real but conditional.

Canonical empirical finding. Svoboda (2007): across young democracies, economic recessions roughly double the hazard of authoritarian reversal, though wealth and prior democratic experience attenuate the association almost entirely. The recession–breakdown link operates through elites and institutional fragility, not only through voter mobilisation.

Three reasons why a shock raises the probability of backsliding.

- **Selection of “demagogues”.** When the value of the status-quo alternative falls, voters become more willing to elect candidates who promise transformation at the cost of institutional commitments (see next slide).
- **Erosion of self-enforcing nature of democracy.** Crises raise the perceived stakes of losing office bringing incumbents closer to the threshold at which consolidation dominates fair competition.
- **Constitutionally available emergency instruments.** Shocks make state-of-emergency powers legally available.

Demagogues and democratic fragility (1): the model

The game. A repeated electoral contest between two parties:

- A **benevolent party** that maximises long-run voter welfare.
- A **demagogue** that cares only about winning office.

The voter. The median voter is **myopic** — evaluates candidates solely on current-period welfare (consumption plus a candidate-specific valence component).

The policy choice. Each period, the winning party allocates the existing capital stock between current consumption and investment. Investment grows future capital; disinvestment liquidates capital for current consumption, bounded by some maximum fraction of the current stock.

The demagogue's strategy. Always proposes **maximum disinvestment** — liquidate as much capital as is feasible in order to maximise current consumption and the probability of winning. The demagogue's policy is constant and known.

The benevolent party's dilemma. It would prefer the social planner's investment level. But the myopic voter rewards current consumption, so to beat the demagogue the benevolent party must propose *less* investment than the planner would. The strategic question is *how much less*.

Demagogues and democratic fragility (2): the death spiral

Central result. There is a threshold on the capital stock.

- **Above the threshold.** The benevolent party invests *more* than the planner would – building a buffer that will be needed in future periods when capital is lower.
- **Below the threshold.** The benevolent party must reduce investment to remain electorally competitive. As capital falls further, it converges towards the demagogue's policy – a self-reinforcing decline in capital that the authors call the **death spiral**.
- The spiral occurs *even when the demagogue never wins in equilibrium*. The mere presence of a credible demagogue is enough to force the benevolent party to mimic.

Lesson. Resilience must be built in good times – precisely when demagogues look least appealing and the political incentive to invest in institutional capital is weakest. Poorer or younger democracies face higher risk; tighter constraints on disinvestment, higher productivity, and parties that reliably select high-valence candidates all reduce this risk.

When do challengers emerge (2): structural drift

Claim. Slow-moving exogenous changes — e.g. in skill-biased technology and in political culture — raise the probability of democratic backsliding over decades, by generating durable electoral demand for outsider candidates and by eroding the inherited beliefs or norms.

Two sources of drift.

- **Material.** Skill-biased technological change and automation displace routine-task workers, generate persistent regional inequalities, and erode the base of mainstream parties.
- **Cultural.** Long-run value change and inherited beliefs about the political class change slowly and condition who is electable.

Connection to backsliding. Both channels raise the probability that an aspiring backslider is elected. Most of the evidence is about populist vote shares; the step from populist vote to backsliding is itself contested.

Bad democracy traps (1): the model (Gratton, Lee, and Yousaf, 2025)

The game. In each period the voter chooses between:

- An **ambitious candidate**, who promises — but cannot commit — to pursuing a risky long-term agenda whose payoff is large but uncertain and hard to monitor. A **non-ambitious candidate**, who delivers a safe default agenda.

Two primitives of the political environment.

- The **quality of the political class** — the probability that an ambitious candidate is honest, that is, will actually try to deliver the agenda rather than diverting to private rents.
- The **transparency of institutions** — the probability that, by the end of the candidate's term, the voter observes the agenda's outcome.

Political culture. The voter's **prior** about these two primitives. Inherited from history; possibly misspecified; not directly under the voter's control in any given period.

Information asymmetry. Candidates know the true environment *and* know the voter's culture. The voter knows only her culture; oblivious to its possible misspecification and candidates' attempts to exploit it.

Long-run dynamics. Culture is updated by observed outcomes through a **self-confirming equilibrium** (Battigalli). A culture survives if observed outcomes do not contradict it. Critically, a culture can be self-confirming *even when it does not match the true environment* — because the voter's choices, conditional on her culture, generate the observations that sustain it.

Bad democracy traps (2): equilibria and implications

Central result. Only two types of cultural equilibria can persist.

- **Bad democracy trap.** If the voter's culture is sufficiently negative, she never elects the ambitious candidate. Because no ambitious agenda is ever attempted, no observations contradict the negative prior. The culture persists — even *when the true environment would justify trust*. The polity under-performs its actual democratic potential.
- **Wishful-thinking democracy.** If the voter's culture is sufficiently positive, she always elects the ambitious candidate. With high-enough transparency, observed outcomes can sustain even a culture that systematically over-estimates the political class. Reality constrains culture only when the institutional environment supports it.

Bad democracy traps (2): equilibria and implications

Central result. Only two types of cultural equilibria can persist.

- **Bad democracy trap.** If the voter's culture is sufficiently negative, she never elects the ambitious candidate. Because no ambitious agenda is ever attempted, no observations contradict the negative prior. The culture persists — even *when the true environment would justify trust*. The polity under-performs its actual democratic potential.
- **Wishful-thinking democracy.** If the voter's culture is sufficiently positive, she always elects the ambitious candidate. With high-enough transparency, observed outcomes can sustain even a culture that systematically over-estimates the political class. Reality constrains culture only when the institutional environment supports it.

Two implications.

- Good institutions are **necessary but not sufficient**. A polity trapped in a negative culture can stay there, even after transparency-enhancing reforms, because the voter never elects the ambitious candidate whose outcomes would update the culture.
- More information is **double-edged**. It can liberate a bad-trap polity by letting rare ambitious episodes be observed, but it can also destabilise a wishful-thinking democracy whose culture over-rates the political class.

When do challengers emerge (3): cascades and tipping points

Sudden endogenous events — e.g. corruption scandals — can shift the supply of and demand for challengers within a single electoral cycle.

Three illustrative cases.

- *Corruption.* Italy 1992 (Tangentopoli): the Christian Democrats and Socialists collapsed in a single cycle. Brazil's Lava Jato delegitimised the established political class and opened space for Bolsonaro.
- *Constitutional crisis.* US 6 January 2021 and Brazil 8 January 2023 tested whether elites would publicly commit to defending democracy.
- *Polarisation tipping.* **svolik_polarization_2019**: a discrete shift in which the out-party comes to be perceived as existentially threatening.

Note. This cell is theoretically thin. Most evidence is case-based, and the formal theory that exists (informational cascades, critical-mass tipping) is not specifically about backsliding. Sudden endogenous events are best understood as **stress tests** that reveal the strength of the conditions established in the other three cells.

When do challengers emerge (4): policy drift

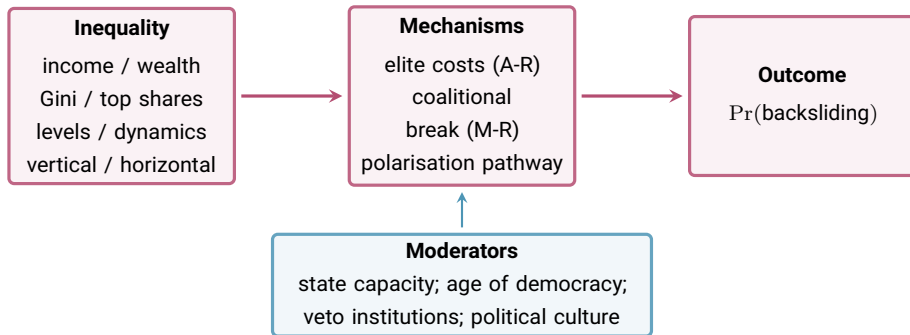
The fourth cell — **policy drift** — captures gradual, endogenous changes that the polity could in principle reverse through deliberate choice but typically does not.

Three prominent forms of policy drift.

- **Rising income and wealth inequality.**
- Long-term corruption.
- Elite norm decay.

When do challengers emerge (4): inequality

Strong intuition, weak theory & evidence. Rising income and wealth inequality is widely cited as a driver of contemporary backsliding.



Implications. The next slide presents the two prominent mechanisms; the one after sets out the measurement and identification challenges that make the empirical work harder than it looks.

Inequality and backsliding (1): two prominent mechanisms

- **Elite costs — Acemoglu and Robinson (2005).** Under democracy the median voter is poor, so equilibrium redistribution is high; rising inequality raises it further. Higher redistribution raises the cost to elites of operating within democratic constraints. When that cost exceeds the cost of consolidation, elites prefer non-democratic outcomes.
- **Coalitional structure — Mukand and Rodrik (2020).** Liberal democracy requires an alliance between the non-elite majority and a minority elite against the majority elite. Inequality aligned with the class cleavage pushes the non-elite majority to view the entire elite as the threat, breaking the alliance. Inequality undermines **liberal** democracy specifically — the rule-of-law and minority-rights components — not majoritarian democracy in general.

Note. Both predict that higher inequality raises backsliding risk. They differ on *through whom* the effect runs (elites in A-R; the cross-cleavage coalition in M-R) and on *what* gives way (democracy in general in A-R; the liberal components specifically in M-R).

Inequality and backsliding (2): empirical challenges

Measurement. “Inequality” is a family of measures; different choices give different answers.

- Income vs wealth; Gini vs top 1% / 10% / bottom 50%; pre-tax vs post-tax.
- Levels (cross-section) vs changes (within-country dynamics). The theoretical accounts implicate *dynamics* – a rise in inequality changing elite calculus or coalitional structure – but most empirical work exploits cross-sectional variation.
- Vertical (rich–poor) vs horizontal (group-based) inequality may operate through different channels.

Causal identification. The list of plausible confounders is long.

- *Reverse causality.* Backsliding incumbents attack progressive taxation and redistribution; measured inequality rises as a consequence of erosion, not as a cause.
- *Common drivers.* Skill-biased technological change, financialisation, and party-system decay jointly raise inequality and weaken the coalitions sustaining liberal democracy.
- *Selection.* The set of countries coded as eroding is itself non-random.

Summary. A descriptive correlation between inequality and erosion exists (Rau and S. Stokes, [2025](#); S. C. Stokes, [2025](#)). The causal claim that rising inequality *drives* backsliding remains under-identified.

When do challengers emerge (4): corruption and elite norm decay

Long-term corruption. Two channels:

- *Legitimacy.* Persistent corruption depresses diffuse support for democratic institutions, opening electoral space for outsider candidates running on anti-establishment platforms – the supply-side counterpart of FM2a.
- *Selection.* Corruption deters honest entrants and rewards opportunistic ones, reducing the supply of high-quality candidates.

Elite norm decay. The supply-side counterpart of FM4. When mainstream elites stop publicly committing to democratic rules – endorsing election denial, refusing to condemn co-partisan violations, treating norm violations as ordinary politics – the constraint on would-be backsliders weakens at the margin.

- Linz's *semi-loyalty*: parties that profess loyalty to democracy but remain open to allying with disloyal partners. The older intellectual lineage of the contemporary norms literature.
- Helmke, Kroeger, and Paine (2022): democracy by deterrence – formalises the equilibrium dynamics of incremental norm violations going unpunished (cited in FM4 with a different emphasis).

Elite support: who matters

Would-be backsliders need the cooperation, or at least the acquiescence, of other elites.

Primary focus: political elites.

- Mainstream parties decide whether to *contain* or *accommodate* the challenger – the central decision in the comparative-politics literature on backsliding.
- The decisions are strategic, repeated, and publicly observable. The theoretical and empirical record is correspondingly developed.

Other elite groups also matter.

- *Business elites.* Whether economic actors discipline backsliders (via market signals, exit) or accommodate them (via co-optation under pressure).
- *Bureaucrats.* Whether civil servants comply with constitutional orders, resign, or actively obstruct unconstitutional ones. Connects to the institutional veto-player logic of FM4.
- *Media.* Whether organisations frame violations as transgressions or as ordinary politics, shaping the information environment for both other elites and voters.

Political elites: strategic challenges

The basic strategic problem. A mainstream party faces a binary choice each period: **cooperate** with the would-be backslider or **contain** them.

Trade-off for the individual party.

- *Benefits of cooperation.* Access to office and policy; opportunity to constrain the challenger's agenda from within; avoidance of being outflanked by another mainstream party that does cooperate.
- *Costs of cooperation.* Normalising the challenger; raising the probability that the challenger ultimately dominates the coalition; surrendering the reputational claim to democratic loyalty.

Three strategic complications.

- **Coordination.** Even when each mainstream party prefers containment, they may fail to coordinate on it. If one party defects to cooperation, others may have to follow to avoid being outflanked.
- **Loyalty.** Linz's distinction between **loyal** democrats, **semi-loyal** actors (who profess loyalty but remain open to allying with disloyal partners), and **disloyal** opportunists.
- **Time inconsistency.** Containment may be optimal in each period but unsustainable across many periods if the challenger continues to grow.

Conservative parties (1): the conservative dilemma

The central question. Ziblatt (2017): why did democratic transitions in late-19th- and early-20th-century Europe end differently in different countries? The answer turns on the organisational form of the [conservative party](#).

The conservative dilemma. After 1848, old-regime elites had to win power through democratic competition. This required adapting to mass politics — but too much adaptation would erode the hierarchical order they sought to preserve. Two strategic responses emerged.

- **Formal party-building.** Build a hierarchical mass party with professional staff, local associations, wide geographic spread, and disciplined leadership. The party mobilises voters on its own organisational base. Cases: Brit. Conservatives, Danish Højre, Belgian Catholics, Dutch Anti-Rev. Party, Swedish Lantmannapartiet.
- **Informal manipulation.** Substitute clientelism, fraud, and corruption — Spanish *turno pacífico*, Italian *trasformismo*, German *Wahlbeeinflussung*, Portuguese *rotativismo* — for genuine mass mobilisation. The party retains office without building a base. Cases: German Deutsch-Konservative Partei, Spanish Partido Conservador, Italian Destra, Portuguese Partido Regenerador.

The decisive consequence. The strong-party route delivers an organisational base the leadership controls. The informal route leaves the leadership [structurally dependent](#) on outside reactionary groups for mobilisation. It is that dependence — not preferences or values — that determines whether the party can later say no to radicals.

Conservative parties (2): settled and unsettled democratisation

The mechanism. When the radical right emerged in the 1920s and 1930s, the two types of conservative party responded very differently.

- *Strong conservative parties* could **subordinate** the radicals. The British Conservatives absorbed and contained the right; the Danish, Belgian, and Dutch right held its own electoral base and used it to discipline outside groups.
- *Weak conservative parties* could not. Dependence on outside reactionary forces for mobilisation meant excluding the radicals would have cost the party its own electoral viability. The result: the radical right ultimately **captured or destroyed** the conservative party — Italy's Fascists, Germany's Nazis, Spain's Falange.

Two patterns of democratisation.

- **Settled.** UK, Belgium, Netherlands, Scandinavia. Conservatives built mass parties before universal male suffrage and the socialist challenge. Democracy consolidated through gradual, low-volatility steps.
- **Unsettled.** Germany, Italy, Spain, Portugal. Conservatives relied on informal manipulation. Democratisation was punctuated by reversals, coups, and interwar breakdown.

Elite cooptation as engineered fragmentation

Setting. Arriola, Devaro, and Meng (2021): an electoral regime in which the incumbent enjoys an advantage but still faces competitive multiparty elections. The aspiring backslider wants to prevent the formation of a unified opposition (evidence from African countries).

The mechanism. The incumbent uses **cabinet appointments as patronage**, offering ministerial posts to individual opposition politicians. Accepting brings private benefits (rents, state resources for one's constituency); *anticipating* a future offer gives each politician a reason to run independently rather than coalesce, in order to remain a credible target for cooptation. Weak incumbents need this strategy most and rely on it most.

Connection to the strategic-challenges slide.

- **Coordination.** Each opposition politician has a private incentive to defect. The aspiring backslider *actively engineers* the coordination failure rather than passively benefiting from it.
- **Loyalty.** The equilibrium does not require the incumbent to identify the semi-loyal types — only that such types exist since this is enough to make “stay independent” rational for every opposition politician.

Opposition strategy under aggrandisement

The question. Cleary and Öztürk (2022): when does executive aggrandisement lead to autocratic breakdown, and when does democracy survive? Five comparative cases – Venezuela (Chávez), Turkey (Erdoğan), Thailand (Thaksin), Bolivia (Morales), Ecuador (Correa).

Three possible outcomes.

- **Incumbent takeover** (Venezuela, Turkey): full consolidation into authoritarian rule.
- **Incumbent removal** (Thailand, Bolivia 2019): incumbent removed extra-constitutionally – also a form of breakdown.
- **Democratic survival** (Ecuador, Bolivia pre-2019): incumbent eventually replaced through normal electoral procedures.

The opposition's strategic dilemma. Two responses.

- *Moderate.* Reformist demands, mobilisation within the political arena – parliamentary challenges, electoral coalitions, public protest against constitutional changes. Activates democratic norms, maintains public legitimacy, buys time until the next election.
- *Radical.* Maximalist demands, mobilisation in bureaucratic, economic, and military arenas – election boycotts, calls for military intervention, attempts to ban the incumbent's party. Diminishes democratic norms, polarises public opinion; failed attempts boost the incumbent.

The key result. **Moderation works; radicalism backfires.** In all five cases the opposition was moderate in the first electoral cycle. Divergence appears after the *second* election: where the opposition turned radical (Turkey, Venezuela, Thailand), breakdown followed; where it remained moderate (Ecuador, Bolivia until 2019), democracy survived.

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

Outline

Motivation and overview

Conceptualising and measuring democracy

The demand side of democratic backsliding

The supply side of democratic backsliding

Conclusion

Appendix

References I

- Acemoglu, Daron and James A Robinson (2005).** *Economic Origins of Dictatorship and Democracy*. MAG ID: 2890297640 S2ID: 0e2071acc68a2bca51d1794a2778579d7b95b2f9. Cambridge: Cambridge University Press. DOI: [10.1017/cbo9780511510809](https://doi.org/10.1017/cbo9780511510809).
- Albertus, Michael and Guy Grossman (Jan. 2021).** "The Americas: When Do Voters Support Power Grabs?" *Journal of Democracy* 32.2. MAG ID: 3153683934, pp. 116–131. DOI: [10.1353/jod.2021.0023](https://doi.org/10.1353/jod.2021.0023).
- Arriola, Leonardo R., Jed Devaro, and Anne Meng (Nov. 2021).** "Democratic Subversion: Elite Cooptation and Opposition Fragmentation". en. *American Political Science Review* 115.4, pp. 1358–1372. ISSN: 0003-0554, 1537-5943. DOI: [10.1017/S0003055421000629](https://doi.org/10.1017/S0003055421000629).
- Basak, Deepal, Joyee Deb, and Aditya Kuvalekar (Apr. 2026).** "Similarity of Information and Collective Action". en. *American Economic Review* 116.4, pp. 1189–1233. ISSN: 0002-8282. DOI: [10.1257/aer.20241056](https://doi.org/10.1257/aer.20241056).

References II

- Bermeo, Nancy (Jan. 2016).** "On Democratic Backsliding". *Journal of Democracy* 27.1. MAG ID: 2296297756, pp. 5–19. DOI: [10.1353/jod.2016.0012](https://doi.org/10.1353/jod.2016.0012).
- Bernhardt, Dan, Stefan Krasa, and Shadmehr Shadmehr (Oct. 2022).** "Demagogues and the Economic Fragility of Democracies". *The American Economic Review* 112.10. MAG ID: 4298012868 S2ID: b9805e51ebaa2f71847eb2bdd87c59f0f8626f4c, pp. 3331–3366. DOI: [10.1257/aer.20211125](https://doi.org/10.1257/aer.20211125).
- Braley, Alia et al. (Aug. 2023).** "Why voters who value democracy participate in democratic backsliding". en. *Nature Human Behaviour* 7.8, pp. 1282–1293. ISSN: 2397-3374. DOI: [10.1038/s41562-023-01594-w](https://doi.org/10.1038/s41562-023-01594-w).
- Cantoni, Davide et al. (May 2019).** "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement". *The Quarterly Journal of Economics* 134.2, pp. 1021–1077. ISSN: 0033-5533. DOI: [10.1093/qje/qjz002](https://doi.org/10.1093/qje/qjz002).
- Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland (2010).** "Democracy and dictatorship revisited". *Public Choice* 143.1/2, pp. 67–101. ISSN: 0048-5829.

References III

- Claassen, Christopher (2020a).** "Does Public Support Help Democracy Survive?" en. *American Journal of Political Science* 64.1. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12452>, pp. 118–134. ISSN: 1540-5907. DOI: [10.1111/ajps.12452](https://doi.org/10.1111/ajps.12452).
- **(2020b).** "In the Mood for Democracy? Democratic Support as Thermostatic Opinion". *American Political Science Review* 114.1. MAG ID: 3123975835, pp. 36–53. DOI: [10.1017/s0003055419000558](https://doi.org/10.1017/s0003055419000558).
- Clark, William Roberts, Matt Golder, and Sona N. Golder (2025).** *Principles of Comparative Politics - International Student Edition*. English. Thousand Oaks, California London New Delhi Singapore: CQ Press. ISBN: 978-1-0718-5253-8.
- Cleary, Matthew R. and Aykut Öztürk (Mar. 2022).** "When Does Backsliding Lead to Breakdown? Uncertainty and Opposition Strategies in Democracies at Risk". en. *Perspectives on Politics* 20.1, pp. 205–221. ISSN: 1537-5927, 1541-0986. DOI: [10.1017/S1537592720003667](https://doi.org/10.1017/S1537592720003667).

References IV

Edmond, Chris (Oct. 2013). "Information Manipulation, Coordination, and Regime Change*". en. *The Review of Economic Studies* 80.4, pp. 1422–1458. ISSN: 1467-937X, 0034-6527. DOI:

[10.1093/restud/rdt020](https://doi.org/10.1093/restud/rdt020).

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova (2020). "Social Media and Protest Participation: Evidence From Russia". en. *Econometrica* 88.4. _eprint:

<https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14281>, pp. 1479–1514. ISSN: 1468-0262. DOI:

[10.3982/ECTA14281](https://doi.org/10.3982/ECTA14281).

Frederiksen, Kristian, Robb Willer, and Michael Bang Petersen (2026). "The Way Back After Backsliding: Public Opinion and the Restoration of Democracy". *Comparative Political Studies* Forthcoming.

Gidron, Noam et al. (Dec. 2025). *Resistance to Backsliding: The Role of Democratic Conceptions*. DOI:

[10.31235/osf.io/538be_v1](https://doi.org/10.31235/osf.io/538be_v1).

References V

- Graham, Matthew H. and Milan W. Svolik (May 2020).** "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States". en. *American Political Science Review* 114.2, pp. 392–409. ISSN: 0003-0554, 1537-5943. DOI: [10.1017/S0003055420000052](https://doi.org/10.1017/S0003055420000052).
- Gratton, Gabriele and Jacob Edenhofer (Nov. 2025).** "The rise and fall of technocratic democracies: unstable majorities and delegation to technocrats". *Journal of European Public Policy* 0.0. _eprint: <https://doi.org/10.1080/13501763.2025.2576160>, pp. 1–30. ISSN: 1350-1763. DOI: [10.1080/13501763.2025.2576160](https://doi.org/10.1080/13501763.2025.2576160).
- Gratton, Gabriele and Barton E Lee (Jan. 2024).** "Liberty, Security, and Accountability: The Rise and Fall of Illiberal Democracies". *The Review of Economic Studies* 91.1, pp. 340–371. ISSN: 0034-6527. DOI: [10.1093/restud/rdad030](https://doi.org/10.1093/restud/rdad030).
- Gratton, Gabriele, Barton E Lee, and Hasin Yousaf (Oct. 2025).** "Bad Democracy Traps". *The Economic Journal*, ueaf111. ISSN: 0013-0133. DOI: [10.1093/ej/ueaf111](https://doi.org/10.1093/ej/ueaf111).

References VI

- Grillo, Edoardo and Carlo Prato (2023).** "Reference Points and Democratic Backsliding". en. *American Journal of Political Science* 67.1. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12672>, pp. 71–88. ISSN: 1540-5907. DOI: [10.1111/ajps.12672](https://doi.org/10.1111/ajps.12672).
- Grossman, Guy et al. (2021).** "The Majoritarian Threat to Liberal Democracy". *Journal of Experimental Political Science*. MAG ID: 3126449539, pp. 1–10. DOI: [10.1017/xps.2020.44](https://doi.org/10.1017/xps.2020.44).
- Hadfield, Gillian K. and Barry R. Weingast (May 2014).** "Microfoundations of the Rule of Law". en. *Annual Review of Political Science* 17.1, pp. 21–42. ISSN: 1094-2939, 1545-1577. DOI: [10.1146/annurev-polisci-100711-135226](https://doi.org/10.1146/annurev-polisci-100711-135226).
- Helmke, Gretchen, Mary Kroeger, and Jack Paine (2022).** "Democracy by Deterrence: Norms, Constitutions, and Electoral Tilting". en. *American Journal of Political Science* 66.2. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12668>, pp. 434–450. ISSN: 1540-5907. DOI: [10.1111/ajps.12668](https://doi.org/10.1111/ajps.12668).

References VII

- Howell, William G., Kenneth A. Shepsle, and Stephane Wolton (2019).** "Executive Absolutism: The Dynamics of Authority Acquisition in a System of Separated Powers". MAG ID: 3122293430. DOI: [10.2139/ssrn.3440604](https://doi.org/10.2139/ssrn.3440604).
- Huq, Aziz Z. and Tom Ginsburg (2018).** *How to Save a Constitutional Democracy*. MAG ID: 3124143643.
- Jacob, Marc S. (2025).** "Citizens as a democratic safeguard? The sequence of sanctioning elite attacks on democracy". en. *American Journal of Political Science* 69.2. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12847>, pp. 455–470. ISSN: 1540-5907. DOI: [10.1111/ajps.12847](https://doi.org/10.1111/ajps.12847).
- Kamenica, Emir and Matthew Gentzkow (Oct. 2011).** "Bayesian Persuasion". en. *American Economic Review* 101.6, pp. 2590–2615. ISSN: 0002-8282. DOI: [10.1257/aer.101.6.2590](https://doi.org/10.1257/aer.101.6.2590).

References VIII

- Knutsen, Carl Henrik et al. (Apr. 2024).** "Conceptual and Measurement Issues in Assessing Democratic Backsliding". en. *PS: Political Science & Politics* 57.2, pp. 162–177. ISSN: 1049-0965, 1537-5935. doi: [10.1017/S104909652300077X](https://doi.org/10.1017/S104909652300077X).
- Krishnarajan, Suthan (May 2023).** "Rationalizing Democracy: The Perceptual Bias and (Un)Democratic Behavior". en. *American Political Science Review* 117.2, pp. 474–496. ISSN: 0003-0554, 1537-5943. doi: [10.1017/S0003055422000806](https://doi.org/10.1017/S0003055422000806).
- Lehmann, Todd C. and Scott A. Tyson (Apr. 2021).** "Sowing the Seeds: Radicalization as a Political Tool". *American Journal of Political Science*. MAG ID: 3156308290. doi: [10.1111/ajps.12602](https://doi.org/10.1111/ajps.12602).
- Levitsky, Steven and Daniel Ziblatt (Jan. 2018).** *How Democracies Die*. MAG ID: 2790793818.
- Little, Andrew T. and Anne Meng (Apr. 2024).** "Measuring Democratic Backsliding". en. *PS: Political Science & Politics* 57.2, pp. 149–161. ISSN: 1049-0965, 1537-5935. doi: [10.1017/S104909652300063X](https://doi.org/10.1017/S104909652300063X).

References IX

- Little, Andrew T., Keith E. Schnakenberg, and Ian R. Turner (May 2022).** “Motivated Reasoning and Democratic Accountability”. en. *American Political Science Review* 116.2, pp. 751–767. ISSN: 0003-0554, 1537-5943. DOI: [10.1017/S0003055421001209](https://doi.org/10.1017/S0003055421001209).
- Luo, Zhaotian and Adam Przeworski (Feb. 2023).** “Democracy and its Vulnerabilities: Dynamics of Democratic Backsliding”. *Quarterly Journal of Political Science* 18.1, pp. 105–130. ISSN: 1554-0626. DOI: [10.1561/100.00021112](https://doi.org/10.1561/100.00021112).
- Morris, Stephen and Hyun Song Shin (2003).** “Global Games: Theory and Applications”. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*. Ed. by Lars Peter Hansen, Mathias Dewatripont, and Stephen J. Turnovsky. Vol. 1. Econometric Society Monographs. Cambridge: Cambridge University Press, pp. 56–114. ISBN: 978-0-521-52411-7. DOI: [10.1017/CB09780511610240.004](https://doi.org/10.1017/CB09780511610240.004).

References X

- Mukand, Sharun W and Dani Rodrik (Apr. 2020).** "The Political Economy of Liberal Democracy". *The Economic Journal* 130.627, pp. 765–792. ISSN: 0013-0133. DOI: [10.1093/ej/ueaa004](https://doi.org/10.1093/ej/ueaa004).
- Neundorf, Anja et al. (Jan. 2026).** "Elections Without Constraints? The Appeal of Electoral Autocracy Across the World". en. *British Journal of Political Science* 56, e1. ISSN: 0007-1234, 1469-2112. DOI: [10.1017/S0007123425101129](https://doi.org/10.1017/S0007123425101129).
- Rau, Eli G. and Susan Stokes (Jan. 2025).** "Income inequality and the erosion of democracy in the twenty-first century". *Proceedings of the National Academy of Sciences* 122.1, e2422543121. DOI: [10.1073/pnas.2422543121](https://doi.org/10.1073/pnas.2422543121).
- Rodrik, Dani and Romain Wacziarg (May 2005).** "Do Democratic Transitions Produce Bad Economic Outcomes?" en. *American Economic Review* 95.2, pp. 50–55. ISSN: 0002-8282. DOI: [10.1257/000282805774670059](https://doi.org/10.1257/000282805774670059).

References XI

- Scheppele, Kim Lane (2018).** "Autocratic Legalism". *The University of Chicago Law Review* 85.2, pp. 545–584. ISSN: 0041-9494.
- Shapiro, Ian (2020).** "Appendix D: On Counting Democratic Countries II". *On Democracy*. Ed. by Robert A. Dahl. Veritas.
- Stokes, Susan C. (2025).** *The Backsliders: Why Leaders Undermine Their Own Democracies*. English. Princeton: Princeton University Press. ISBN: 978-0-691-27154-5.
- Svolik, Milan W. (Jan. 2007).** "Authoritarian Reversals and Democratic Consolidation". *American Political Science Review*. MAG ID: 2152218134 S2ID: 042640ad28de94c399114c767a89dc989f93b63a. doi: [10.1017/s0003055408080143](https://doi.org/10.1017/s0003055408080143).
- **(2012).** *The Politics of Authoritarian Rule*. Cambridge: Cambridge University Press. doi: [10.1017/CB09781139176040](https://doi.org/10.1017/CB09781139176040).

References XII

- Svolik, Milan W. (Oct. 2015).** "Which Democracies Will Last? Coups, Incumbent Takeovers, and the Dynamic of Democratic Consolidation". en. *British Journal of Political Science* 45.4, pp. 715–738. ISSN: 0007-1234, 1469-2112. DOI: [10.1017/S0007123413000550](https://doi.org/10.1017/S0007123413000550).
- **(Jan. 2019).** "Polarization versus Democracy". *Journal of Democracy* 30.3. MAG ID: 2962254002 S2ID: cc53a3e10f214a5fcdc0aabcb515768fa73bbf6f, pp. 20–32. DOI: [10.1353/jod.2019.0039](https://doi.org/10.1353/jod.2019.0039).
- **(Jan. 2020).** "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents". *Quarterly Journal of Political Science* 15.1. MAG ID: 2985827668 S2ID: ffebc4654391c5456fbdb98054b8471458f6cd7f, pp. 3–31. DOI: [10.1561/100.00018132](https://doi.org/10.1561/100.00018132).
- Tai, Yuehong 'Cassandra', Yue Hu, and Frederick Solt (Feb. 2024).** "Democracy, Public Support, and Measurement Uncertainty". en. *American Political Science Review* 118.1, pp. 512–518. ISSN: 0003-0554, 1537-5943. DOI: [10.1017/S0003055422000429](https://doi.org/10.1017/S0003055422000429).

References XIII

- Treisman, Daniel (2023).** "How Great is the Current Danger to Democracy? Assessing the Risk With Historical Data". *Comparative Political Studies* 56.12.
- Weingast, Barry R. (June 1997).** "The Political Foundations of Democracy and the Rule of the Law". en. *American Political Science Review* 91.2, pp. 245–263. ISSN: 0003-0554, 1537-5943. DOI: [10.2307/2952354](https://doi.org/10.2307/2952354).
- Weitzel, Daniel et al. (2026).** "Measuring electoral democracy with observables". en. *American Journal of Political Science* 70.2. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12968>, pp. 606–622. ISSN: 1540-5907. DOI: [10.1111/ajps.12968](https://doi.org/10.1111/ajps.12968).
- Wunsch, Natasha, Marc S. Jacob, and Laurenz Derksen (Jan. 2025).** "The Demand Side of Democratic Backsliding: How Divergent Understandings of Democracy Shape Political Choice". en. *British Journal of Political Science* 55, e39. ISSN: 0007-1234, 1469-2112. DOI: [10.1017/S0007123424000711](https://doi.org/10.1017/S0007123424000711).

References XIV

Ziblatt, Daniel (Apr. 2017). *Conservative Parties and the Birth of Democracy*. MAG ID: 2605433556 S2ID: 9a6fe002f8f3776ef6e1f6323982842f847ff032. Cambridge: Cambridge University Press.

Little, Schnakenberg & Turner: Summary

Motivated reasoning has two distinct effects.

- **Divergence** — voters with different affinities for the incumbent form *different* conclusions about her competence. The *across-voter* effect: partisans split.
- **Desensitisation** — for any *given* voter, conclusions become *less responsive* to new performance information. The *within-voter* effect: the slope of beliefs on signals flattens.

Main argument. Desensitisation has more uniform consequences for democratic accountability than divergence. Divergence alone may not harm incumbent performance at all; desensitisation always does, and it matters every voter — including centrists.

Roadmap. (i) Model set-up; (ii) the directional motive; (iii) three benchmark forms of $v(\cdot)$; (iv) divergence and desensitisation made formal; (v) why desensitisation is the pernicious effect; (vi) back to backsliding.

Model set-up

Players. A finite electorate of voters $j = 1, \dots, n$, an incumbent (I), and a non-strategic challenger (C).

Competence. Incumbent competence $\theta_I \sim \mathcal{N}(0, \sigma_\theta^2)$; challenger competence $\theta_C \sim \mathcal{N}(\mu_C, \sigma_\theta^2)$ with $\mu_C < 0$ when the incumbent is ex-ante more competent.

Effort and signal. The incumbent chooses effort $e \geq 0$ at convex cost $c(e)$. Voters observe a public performance signal correlated with both competence and effort:

$$s = \theta_I + e + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$

Voter affinity. Each voter j has an affinity $a_j \in \mathbb{R}$ for the incumbent, capturing tastes *independent of* performance: $a_j > 0$ for co-partisans, $a_j < 0$ for opponents.

Voting. After observing s and conjecturing effort $\hat{e}$, each voter forms a *conclusion* $\tilde{\theta}_I$ about incumbent competence. The voter then votes for the incumbent iff competence-plus-affinity exceeds the challenger's expected competence (plus a common shock). The candidate with majority votes wins.

The directional motive: the key innovation

Bayesian benchmark. A pure-accuracy voter would set her conclusion equal to the posterior mean of θ_I given the signal:

$$\bar{\mu}(s) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} (s - \hat{e}), \quad \bar{\sigma}_\theta^2 = \frac{\sigma_\theta^2 \sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}.$$

LST's modification. The voter picks $\tilde{\theta}_I^*$ to maximise the sum of an accuracy motive (the log-likelihood of the Bayesian posterior) and a directional motive scaled by $\delta \geq 0$:

$$\tilde{\theta}_I^*(s, a_j) \in \arg \max_{\tilde{\theta}_I} \underbrace{\log f_{\theta_I|s}(\tilde{\theta}_I | s)}_{\text{accuracy}} + \underbrace{\delta v(a_j, \tilde{\theta}_I)}_{\text{directional}}.$$

The directional function $v(a_j, \tilde{\theta}_I)$. Assumed continuous and weakly concave in $\tilde{\theta}_I$, with $\partial^2 v / (\partial a_j \partial \tilde{\theta}_I) \geq 0$: voters more attached to the incumbent have an intrinsic reason, beyond evidence, to conclude she is more competent.

Two limiting cases.

- $\delta = 0$ — pure Bayesian; no motivated reasoning.
- $\delta > 0$ — the case of interest: the optimal conclusion balances accuracy against v .

Three pathologies

The paper analyses three types of $v(a_j, \tilde{\theta}_I)$ that generate qualitatively different patterns of bias.

(1) Polarised partisanship. $v(a_j, \tilde{\theta}_I) = a_j \cdot \tilde{\theta}_I$.

$$\tilde{\theta}_I^{\text{part}}(s, a_j) = \bar{\mu}(s) + \delta \bar{\sigma}_\theta^2 a_j.$$

Bayesian conclusion shifted up or down by a constant proportional to affinity; *slope on s unchanged*.

(2) Confirmation bias. $v(a_j, \tilde{\theta}_I) = -\tilde{\theta}_I^2$.

$$\tilde{\theta}_I^{\text{conf}}(s, a_j) = \frac{1}{1 + 2\delta \bar{\sigma}_\theta^2} \bar{\mu}(s).$$

All voters pulled toward the common prior (zero); *slope on s flattened*; no role for a_j .

(3) Spatial motivations. $v(a_j, \tilde{\theta}_I) = -(a_j - \tilde{\theta}_I)^2$. The affinity acts as an “ideal point” for the conclusion.

$$\tilde{\theta}_I^{\text{spat}}(s, a_j) = \frac{1}{1 + 2\delta \bar{\sigma}_\theta^2} \bar{\mu}(s) + \frac{2\delta \bar{\sigma}_\theta^2}{1 + 2\delta \bar{\sigma}_\theta^2} a_j.$$

Both an intercept shift (with a_j) *and* a slope reduction.

Divergence vs desensitisation

Linear decomposition. Under linearity of $\partial v / \partial \tilde{\theta}_I$ in a_j and $\tilde{\theta}_I$, every optimal conclusion can be written as

$$\tilde{\theta}_I^*(s, a_j; \hat{e}) = \alpha_0 + \alpha_1 a_j + \beta (s - \hat{e}), \quad \alpha_1 \geq 0, \quad \beta \geq 0.$$

Interpretation.

- α_1 – the *intercept shift with affinity*: the cross-voter **divergence** effect.
- β – the *slope on the signal*: the within-voter **(de)sensitisation** effect (β low $\Rightarrow$ flat slope $\Rightarrow$ desensitised).

Directional motive	Divergence	Desensitisation
Polarised partisanship	✓	—
Confirmation bias	—	✓
Spatial motivations	✓	✓

The two effects are conceptually orthogonal and need not co-occur: *polarised partisanship* produces *only* divergence; confirmation bias produces *only* desensitisation.

Why desensitisation is so pernicious

Intuition. Divergence alone (the polarised-partisanship case) leaves all voters — pro, anti, and centrist — shifting their conclusions “in parallel” when performance changes. Because the election is decided by the median, and the median tends to be roughly neutral, this does not undermine accountability.

The result (Proposition 3 & Corollary 2). Whenever the directional motive produces desensitisation, the incumbent's equilibrium effort e^* is reduced. In the limit, as $\delta \rightarrow \infty$ under any motive featuring desensitisation, $e^* \rightarrow 0$. Divergence alone, by contrast, can leave e^* unchanged.

Why this matters normatively. The voter-psychology literature has historically lamented *polarised partisanship* — deeply committed partisans holding systematically divergent beliefs. LST's key contribution is to show that this is *not* the most pernicious mechanism. The more general threat is desensitisation: *even moderate, centrist voters become less responsive to new information*, which is what blunts the incumbent's incentive to perform.

Caveat. “Pernicious” here is about *accountability through effort*. The translation to electoral outcomes is more nuanced: both effects weaken the correlation between performance and vote share, so the standard empirical proxy for accountability cannot disentangle them.

How this speaks to the backsliding debate

- Incumbent competence $\theta_I \Leftrightarrow$ the leader's commitment to democratic norms.
- Performance signal $s \Leftrightarrow$ an observed action (judicial reform, media regulation, emergency decree).
- Voter affinity $a_j \Leftrightarrow$ partisan attachment to the incumbent.
- Conclusion $\tilde{\theta}_I^* \Leftrightarrow$ the voter's actual belief about whether the leader is a democrat.

Empirical implications.

- **Divergence** — co-partisans of the backslider rate the same action more favourably than opponents (the Krishnarajan-style result).
- **Desensitisation** — as violations accumulate, voters' beliefs become *less* responsive to each new piece of evidence: the flattening slope. This may be the more dangerous pattern; it predicts that even centrists lose the ability to update.

Grillo & Prato: model set-up

Players.

- A unit mass of citizens indexed by i , with ideology $\theta_i \sim F$.
- An incumbent I with privately-observed ideology $\theta_I \sim F_I$.

Incumbent's two-stage choice.

1. **Challenge:** $c \in \{0, 1\}$. Whether to attack democratic norms.
2. **Escalation:** $d \in [\delta, 1]$ if $c = 1$. How far to “double down”.

$\delta \in (0, 1)$ captures *institutional checks and balances*: lower $\delta \Rightarrow$ stronger checks (the incumbent cannot easily retreat all the way back).

Policy outcome and citizen material utility. $y(c, d) = 1 + cd$, $u(\mathbf{q}; \theta_i) = \theta_i(1 + cd) - cd = \theta_i + (\theta_i - 1)cd$.

The first term is policy preference; the $-cd$ term is the *common aversion to backsliding*. Citizens with $\theta_i > 1$ favour backsliding; those with $\theta_i < 1$ oppose it.

Distributions (Assumptions 1–2).

- F uniform on $[-\frac{1}{2\psi}, \frac{1}{2\psi}]$ with $\frac{1}{2\psi} > 1$: a majority of citizens, but not all, oppose backsliding. Lower $\psi \Leftrightarrow$ greater mass polarisation.
- F_I uniform on $[\tau - \frac{1}{2\phi}, \tau + \frac{1}{2\phi}]$; lower $\phi \Leftrightarrow$ greater citizen uncertainty about θ_I .

Citizen and incumbent objectives

Citizen support. A citizen with ideology θ_i supports the incumbent iff her *total* utility (defined on the next slide) is non-negative. In equilibrium the share of supporters is

$$\pi(\mathbf{q}) = \int_{-1/(2\psi)}^{1/(2\psi)} \mathbf{1}_{\{v(\mathbf{q};z) \geq 0\}} dF(z).$$

Incumbent utility.

$$u_I(\mathbf{q}; \theta_I) = \underbrace{\theta_I(1 + cd) - cd}_{\text{policy utility}} + \underbrace{R \cdot \pi(\mathbf{q})}_{\text{popular support}},$$

where $R > 0$ measures the importance of popular support (e.g. re-election concern).

Two tensions.

- *Autocratic incumbents* ($\theta_I > 1$): policy preference favours backsliding but support concerns push the other way.
- *Democratic incumbents* ($\theta_I \leq 1$): policy preference and support concerns both favour respecting norms — in the benchmark. Reference dependence will overturn this.

Reference dependence and the personal-equilibrium concept

Total utility with Köszegi–Rabin reference dependence.

$$v(\mathbf{q}; \theta_i \mid \underline{u}) = u(\mathbf{q}; \theta_i) + \eta \cdot [u(\mathbf{q}; \theta_i) - \underline{u}], \quad \eta \geq 0.$$

Outcomes above the reference point $\underline{u}$ generate *relief*; outcomes below it generate *disappointment*. Setting $\eta = 0$ recovers standard preferences.

Endogenous reference point. The reference point is formed *after* observing c but *before* observing d , and equals the rational expectation of material utility conditional on c :

$$\underline{u}(c; \hat{\mathbf{q}}, \theta_i) = \mathbb{E}[u(\hat{\mathbf{q}}; \theta_i) \mid c].$$

Personal equilibrium (Köszegi–Rabin). A profile $(\hat{\mathbf{q}}, \underline{u}(0; \hat{\mathbf{q}}, \theta_i), \underline{u}(1; \hat{\mathbf{q}}, \theta_i))$ such that:

- (SR) $\hat{\mathbf{q}}$ is sequentially rational for the incumbent given citizens' reference-dependent support.
- (RE) Each citizen's reference point equals her rational expectation of material utility under $\hat{\mathbf{q}}$.

Reference points have the fixed-point structure of rational expectations.

Why the timing matters. If reference points formed *before* c , they could not respond to incumbent behaviour. If formed *after* d , reference points and material payoffs always coincide and reference dependence drops out.

Benchmark: no reference dependence ($\eta = 0$)

No popularity concerns ($R = 0$, Proposition 1).

- If $\theta_I > 1$: $c = 1, d = 1$ (autocrat fully escalates).
- If $\theta_I \leq 1$: $c = 0$ (democrat respects norms).

Popularity concerns ($R > 0, \eta = 0$, Proposition 2). Define $\theta^\dagger \equiv 1 + \frac{R\psi}{2}$.

- If $\theta_I > \theta^\dagger$ (*extreme autocrats*): $c = 1, d = 1$.
- If $\theta_I \leq \theta^\dagger$: $c = 0$.

Restrained autocrats. Incumbents in $(1, \theta^\dagger]$ *would* prefer to backslide but are deterred by the loss of popular support. They look indistinguishable from democrats in equilibrium.

Comparative statics.

- Higher R (stronger popularity concern) $\Rightarrow$ higher $\theta^\dagger \Rightarrow$ *fewer* backsliders.
- Higher polarisation (lower ψ) $\Rightarrow$ lower $\theta^\dagger \Rightarrow$ *more* backsliders.
- No role for δ when $\eta = 0$: incumbents who challenge always fully escalate.

With reference dependence: opportunistic authoritarians

Key condition. Reference dependence is *important enough*:

$$\eta \geq \frac{\delta}{2 - \delta}.$$

Proposition 4. There exist cut-offs $\underline{\theta}$ and $\bar{\theta} > 1$ such that:

- (i) $\theta_I > \bar{\theta}$ (*extreme authoritarians*): $c = 1, d = 1$ – full escalation.
- (ii) $\theta_I \in (\underline{\theta}, \bar{\theta}]$ (*opportunistic authoritarians*): $c = 1, d = \delta$ – challenge then *partial retreat*.
- (iii) $\theta_I \leq \underline{\theta}$: $c = 0$ – no challenge.

The new type. Opportunistic authoritarians are incumbents who, in the Proposition 2 benchmark, would have chosen $c = 0$. Reference dependence makes “challenge-then-retreat” the more profitable strategy for them.

Striking implication. The interval $(\underline{\theta}, \bar{\theta}]$ can include incumbents with $\theta_I < 1$ – *democrats* who attack democratic norms because reference dependence makes it electorally profitable to do so. Backsliding need not require an authoritarian leader; it can emerge from the strategic logic of reference-dependent retrospection alone.

The fear-and-relief mechanism

Consider a citizen with $\theta_i < 1$ (opposes backsliding) when the incumbent challenges ($c = 1$):

1. Citizens form the reference point $\underline{u}_1 = \mathbb{E}[u(1, d; \theta_i) \mid c = 1]$, based on *expected* escalation $\underline{d}_1$.
2. Many incumbent types (the extreme authoritarians) fully escalate, so $\underline{d}_1 > \delta$. Reference points adjust *downward*: citizens expect bad outcomes.
3. If the incumbent chooses partial retreat ($d = \delta < \underline{d}_1$), realised material utility *exceeds* the reference point.
4. Citizens experience a *psychological gain* (relief), which can outweigh the material loss from the partial backsliding that has occurred.
5. Net effect: incumbent's support after partial retreat *exceeds* support under $c = 0$, even though democracy has been damaged.

The opportunist's calculus.

- Without reference dependence: challenge is costly because every challenger has $d = 1$ in equilibrium, so support drops sharply.
- With reference dependence: the opportunistic type can profit from the *gap* between expected and realised escalation. He challenges, lets citizens fear the worst, then retreats and pockets the relief.

Comparative statics: polarisation, checks, and uncertainty

Reference dependence ($\eta \uparrow$). The opportunistic-authoritarian interval $(\underline{\theta}, \bar{\theta}]$ expands. More incumbent types find backsliding profitable.

Institutional checks and balances ($\delta \downarrow$). Two opposing effects.

- Lower δ reduces the *material* damage of partial retreat (good).
- Lower δ *increases* the relief citizens experience when the incumbent retreats from $d = 1$ to $d = \delta$ (bad: makes the fear-and-relief strategy more profitable).

Stronger checks can therefore *increase* the likelihood of backsliding.

Mass polarisation ($\psi \downarrow$). Higher polarisation:

- Weakens the disciplining effect of popularity concerns ($\theta^\dagger$ falls) – pushes toward *more* backsliding (Svolik-style).
- Weakens citizens' responsiveness to incumbent actions – reduces the appeal of the fear-and-relief strategy, pushing toward *less* backsliding.
- Net result (Figure 5 of the paper): polarisation can simultaneously *decrease the frequency* of backsliding and *increase its severity*.

Voter uncertainty about ideology ($\phi \downarrow$). Higher uncertainty $\Rightarrow$ stronger fear-and-relief $\Rightarrow$ more opportunistic authoritarians. This is the model's mechanism for why backsliding correlates with the rise of programmatically loose, party-detached populists.

Svolik 2020: vote shares and the winning condition

Vote share for the incumbent. Aggregating informed and uninformed voters,

$$V_A = (1 - \alpha) \left[(1 - \pi) \left(\frac{1}{2} - x_S(\mu) \right) + \frac{\pi}{2} \right] + \alpha \cdot \frac{1 + \mu M + \varepsilon}{2}.$$

- *Informed bracket:* of the $(1 - \pi)$ uniform mass on $[-\frac{1}{2}, \frac{1}{2}]$, the share above the swing voter is $\frac{1}{2} - x_S(\mu)$; plus the polarised mass $\pi/2$ at $+\frac{1}{2}$ votes A .
- *Uninformed bracket:* α voters, of whom share $(1 + \mu M + \varepsilon)/2$ vote A .

Vote share for the challenger.

$$V_B = (1 - \alpha) \left[(1 - \pi) \left(\frac{1}{2} + x_S(\mu) \right) + \frac{\pi}{2} \right] + \alpha \cdot \frac{1 - \mu M - \varepsilon}{2}.$$

Difference.

$$V_A - V_B = -2(1 - \alpha)(1 - \pi)x_S(\mu) + \alpha(\mu M + \varepsilon).$$

The incumbent wins iff $V_A - V_B \geq 0$ – a single inequality combining manipulation gains from the uninformed and civic-virtue losses from the informed.

The incumbent's probability of victory

From a deterministic condition to a probability. Rearranging $V_A \geq V_B$,

$$\alpha(\mu M + \varepsilon) \geq 2(1 - \alpha)(1 - \pi)x_S(\mu),$$

and with $\varepsilon \sim U(-\sigma, \sigma)$ (small enough that the threshold is interior to the support),

$$\Pr(V_A \geq V_B) = \frac{1}{2} + \frac{\alpha\mu M - 2(1-\alpha)(1-\pi)x_S(\mu)}{2\alpha\sigma}.$$

Reduction to a maximisation problem. Maximising $\Pr(V_A \geq V_B)$ in μ is equivalent to maximising the μ -dependent piece

$$W(\mu) \equiv \alpha\mu M - 2(1 - \alpha)(1 - \pi)x_S(\mu).$$

Interpretation. The first term is the incumbent's marginal gain among uninformed voters from manipulation. The second is the marginal loss among informed voters: as μ rises, the swing voter $x_S(\mu)$ shifts toward the incumbent, more informed voters defect, and the incumbent loses ground. The equilibrium balances the two.

Equilibrium μ^* and comparative statics

Substituting $x_S(\mu)$. Using $x_S(\mu) = \frac{x_A + x_B}{2} + \frac{\delta \mu^2}{2(x_A - x_B)}$,

$$W(\mu) = \alpha \mu M - (1 - \alpha)(1 - \pi)(x_A + x_B) - \frac{(1 - \alpha)(1 - \pi) \delta \mu^2}{x_A - x_B}.$$

First-order condition. $\partial W / \partial \mu = 0$ gives

$$\alpha M = \frac{2(1 - \alpha)(1 - \pi) \delta \mu^*}{x_A - x_B} \implies \mu^* = \frac{\alpha M (x_A - x_B)}{2(1 - \alpha)(1 - \pi) \delta}.$$

With symmetric platforms $x_A = -x_B$, this collapses to $\mu^* = \frac{\alpha}{1 - \alpha} \cdot \frac{M x_A}{\delta (1 - \pi)}$, as on the main slide.

Comparative statics.

$$\frac{\partial \mu^*}{\partial \pi} > 0, \quad \frac{\partial \mu^*}{\partial \alpha} > 0, \quad \frac{\partial \mu^*}{\partial M} > 0, \quad \frac{\partial \mu^*}{\partial \delta} < 0.$$

Two worth flagging:

- *Polarisation*: $\partial \mu^* / \partial \pi \propto 1 / (1 - \pi)^2$ – the effect of polarisation is *accelerating*. The last few percentage points produce very large increases in μ^* .
- *Civic virtue*: $\partial \mu^* / \partial \delta \propto -1 / \delta^2$ – marginal returns to civic virtue are sharply *diminishing*. Doubling δ from a low base buys a lot; doubling from a high base buys little.