Week 1: Preliminaries and electoral accountability

A second introduction to formal political economy, Trinity Term 2026

Jacob Edenhofer

28/04/2026

* jacob.edenhofer@nuffield.ox.ac.uk

Overview

1. Briefly discuss course logistics
2. Outline the core themes and explain the logic underlying the course structure
3. Introduce four ideas about electoral accountability – two formally, and the other two informally
4. Conclude and provide short outlook

Outline

Recapitulating the objectives of the course

Managing expectations: This course is intended as a complement to the formal theory class in Hilary.

Constraints: We only have four sessions, and this course assumes very little prior exposure to formal theory or mathematical background knowledge

Best response: Four, formally speaking, fairly light-touch sessions, focused on understanding the intuitions underlying the formal arguments

My hope for this course is threefold: that it will help you develop a better understanding of the workings of modern democracies, equip you with a richer theoretical toolkit, and give you a sense as to where to look if you wish to delve deeper into some of these aspects

Resources, assignments, and teaching style

Central resources:

- Syllabus – quite extensive to reduce search costs for you
- Github course page (Link)

Assignments: You will be able to choose between three types of assignments: (i) formal exercises, (ii) short essays, (iii) simulations. I'll keep them short – well aware that most of you have a lot of other work to do $\rightarrow$ use of generative AI actively encouraged with some caveats

Teaching style: Motivation + one or two formal deep dives + real-world applications $\rightarrow$ happy for you to actively shape style of classes, just let me know (e.g. setting up models in class)

How to get most out of the course?

In class: You are free to follow the lectures in whatever way works best for you, though my view is that students get most of out lectures when (i) they minimise sources of distraction and (ii) actively participate (e.g. ask questions, challenge me, spot mistakes)

At home: Please peruse the required readings; if pressed for time, read whatever you find most interesting, but do so carefully (be able to summarise, apply, and extend logic). Generative AI is a fantastic resource, but not easy to use responsibly.

Feedback: This is my first time teaching a course – so, I will likely make mistakes or get things wrong. Let me know asap, rather than taking it on the chin. Feel free to arrange office hours via email – happy to discuss assignments or your thesis.

Outline

The starting point: Democracy and the ideal of self-government

Democracy's original ideal (Rousseau):

→ Set people free through self-government (Dunn, 2018)

→ See Grofman and Feld, 1988 for an excellent discussion

Self-government:

→ Citizens participate in law-making, being bound only by laws they themselves affirm (Przeworski, 2010). Then the laws reflect the will of the people, making them "free".

Key assumptions:

- Population interests are sufficiently homogeneous (consensus is possible).
- Costs of reaching consensus are low.

▸ Summary of Grofman and Feld

The problem of scale and the ideal of self-government

Small-scale self-government:

→ Feasible only in small, homogeneous societies.

→ Historically, other conditions also mattered: exit options and private information (Bates and Lien, 1985; Stasavage, 2020)

→ Question: How does scale change with technology, and why? Examples?

Large-scale democracies:

→ Transaction costs (time, coordination) rise with population size, even when interests are fairly homogeneous.

→ With heterogeneous interests, it is extremely complicated – if not impossible – to find a consensus that satisfies even "minimal" conditions, as Schumpeter, 1942 argued and Arrow, 1951 – generalising Condorcet's paradox – showed.

▸ Example for Condorcet cycle ▸ Arrow's theorem (formal) ▸ Arrow's theorem (intuitive)

How scale affects the nature of politics

Criterion (explanation)	Small-scale polities	Large-scale polities
Efficiency (Matching outputs to preferences/policies)	Preference efficiency; Policies closely match citizen preferences	Policy efficiency; Better internalisation of externalities; Large-scale public goods
Political Relationships (Formality of interaction)	Informal, personal ties; Clientelism, grassroots negotiation	Formal, institutionalised roles; Professionalised bureaucracy
Systems of Rule (Concentration vs. dispersion)	Intensive rule; Concentration of authority over few	Extensive rule; Dispersion of authority across wider populations
Popular Rule (Citizen participation)	Participatory democracy; Direct citizen decision-making	Representative democracy; Rule through elected representatives

Abridged version of the table in Gerring and Veenendaal, 2020

The second-best ideal of self-government

Fundamental implications of scale and preference aggregation:

- Division of labour is necessary to manage the high costs of collective decision-making.
- Laws likely cannot encode the "will of the people". Instead, representatives must craft laws that peacefully resolve societal conflicts.
- To prevent abuses of power:
  - → Representatives must contest **free and competitive elections**.
  - → Elections confer power **temporarily**.
- For losers to accept outcomes:
  - → All citizens must enjoy **equal, enforceable rights** in some domains.

Second-best ideal: Self-government aims to maximise satisfaction (or minimise dissatisfaction) across citizens by combining **representation** with **equal rights** (Przeworski, 2010).

How do modern democracies realise the second-best ideal?

Key implication:

- In large, heterogeneous societies, democracy is less about accurately aggregating preferences, and more about regulating competition for power.

Strategic logic of modern democracy:

- Institutionalise peaceful competition for office among elites and parties.
- Ensure sufficient responsiveness to citizens' interests and protect inviolable rights to guarantee that losers continue to abide by collective decisions.

Our focus:

- Study the institutions that structure and regulate this competition: **Elections, parties, electoral systems, interest groups, legislatures, judiciary, etc.**.

Formal theory helps us understand the “miracle of democracy”

> *“The miracle of democracy is that conflicting political forces obey the results of voting. People who have guns obey those without them. Incumbents risk their control of governmental offices by holding elections. Losers wait for their chances to win office.* ***Conflicts are regulated, processed according to rules****, and thus limited. This is not consensus, yet not mayhem either. Just regulated conflict; conflict without killing.* ***Ballots are ‘paper stones’.”*** *(Przeworski, 2018, p. 118)*

Key points:

- Institutions (elections, parties, legislatures, judiciary) enable large, heterogeneous societies to live in peace while ensuring a minimum level of liberty.
- This outcome may seem underwhelming — societal peace is not the same as justice.
- Yet, preserving this peace is no mean feat either (let’s talk after the mid-terms).

Outline

Setting the stage for today's sessions: What's the deal with elections?

Fundamental role: Competitive, free, and regular elections regulate societal conflicts peacefully because they perform two functions.

1. **Alternation in office**:
   - → Provides losers with incentive to abide by laws crafted by winners.
   - → Elections render conflicts intertemporal — losers wait for another chance (Y. Li, 2019; Przeworski, 1991). Why not just toss a fair coin?
2. **Voting and accountability**:
   - → Voting ensures population preferences are considered (to some extent).
   - → Without voting → dissatisfaction → instability.

"[The losers] will wait, as long as the policies imposed by the winners are not too extreme or as long as their chance to win at the next opportunity is sufficiently high. [...] Winners have to moderate their policies." (Przeworski, 2018, p. 116)

Elections render political conflict intertemporal

The stakes of office:

- Per-period payoff: R when in power, L when in opposition, with $R > L$.
- Stakes of power: $S \equiv R - L > 0$.

After losing, the loser faces a choice:

- Accept the result: wait for another shot at office through elections, with future probability π.
- Fight: try to take power through extra-electoral means, with success probability p, at one-time cost $C > 0$ (repression, rebellion, sanctions, disorder).

The central idea: Losers tolerate today's defeat *because* tomorrow gives them another chance. Democracy is self-enforcing when waiting beats fighting.

Fearon, 2011; Przeworski, 1991, 2005, 2018; Przeworski, Rivero, and Xi, 2015

Comparing the two payoffs

Players discount the future at rate $\delta \in (0,1)$. Both options share the baseline opposition payoff L; what differs is the extra value of returning to office.

Value of accepting (wait for elections, return with probability π each period):

$$V_E \;=\; \sum_{t=0}^{\infty} \delta^t \big(L + \pi S\big) \;=\; \frac{L + \pi S}{1-\delta}.$$

Value of fighting (pay C once for chance p at office thereafter):

$$V_F \;=\; \frac{L + pS}{1-\delta} - C.$$

The loser complies whenever $V_E \geq V_F$, which yields the compliance condition: $\pi S \;\geq\; pS - \kappa$, where $\kappa \equiv (1-\delta)C$. The parameter κ captures the one-time fighting cost expressed as a per-period equivalent.

▸ Geometric sums refresher ▸ Algebra step by step

What the compliance condition tells us

Rearrange the condition as the minimum chance of return losers need to comply:

$$\pi_{\min}(S) \;=\; p - \frac{\kappa}{S}, \qquad \frac{\partial \pi_{\min}}{\partial S} = \frac{\kappa}{S^2} > 0.$$

Interpreting the comparative statics:

- Higher stakes $S \Rightarrow$ losers demand a better chance of returning to office.
- Lower fighting costs (smaller κ) $\Rightarrow$ accepting today's loss is harder to sustain.
- Higher p (force is easier) $\Rightarrow$ elections must offer more to keep losers in.

Two regimes:

- Low stakes: compliance is easy, but elections risk becoming politically hollow.
- High stakes: losing becomes existential – democracy survives only if returning is credibly possible.

When is democracy self-enforcing?

Suppose institutions can deliver at most $\bar{\pi}$ as the credible chance that today's loser returns to office. Self-enforcement requires $p - \frac{\kappa}{S} \leq \bar{\pi}$.

When force is more tempting than the best institutional offer ($p > \bar{\pi}$), this rearranges to an upper bound on the stakes of power:

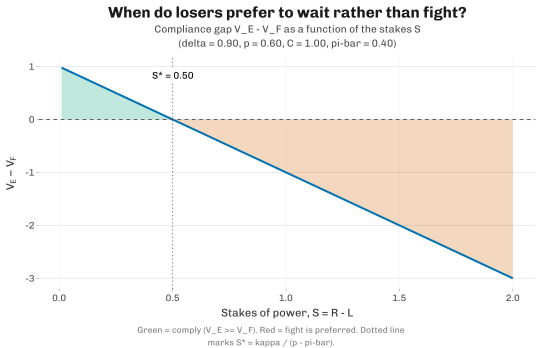
$$S \leq \frac{\kappa}{p - \bar{\pi}}$$

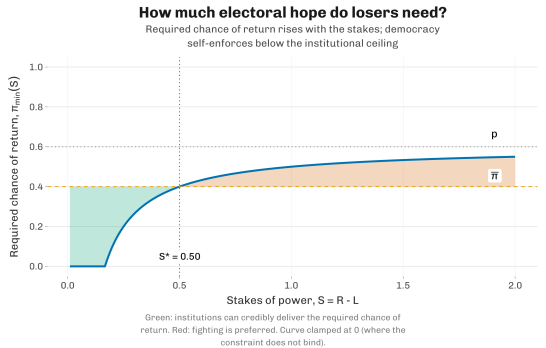
Implications:

- Democracy is fragile when too much rides on a single election.
- Constitutions, federalism, and minority protections work in part by lowering S (limiting what any winner can do).
- Rule of law and credible deterrence work by raising κ (making fighting costlier).

Fearon, 2011; Przeworski, 1991, 2005, 2018; Przeworski, Rivero, and Xi, 2015

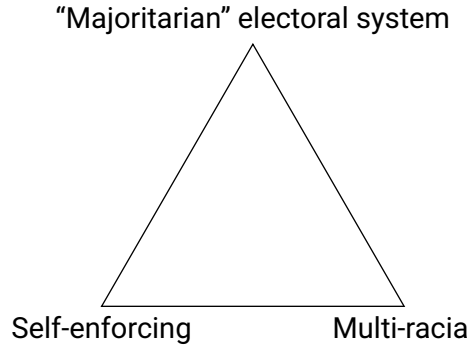
Illustrating the compliance condition





A useless model for advanced democracies? An application to the US

Source: Brookings Institution



Question: Is this the key trilemma of US democracy?

Relates to topical debates: desirability and feasibility of electoral reform in the US (Carey and Pocasangre, 2024; Drutman, 2021; Latner, Santucci, and Shugart, 2021; Mainwaring and Drutman, 2023; Santucci, 2022) and whether the GOP is becoming a multi-racial party (Burn-Murdoch, 2024; Ruffini, 2023)

Why do we vote, rather than just tossing a fair coin?

Because voting allows us to communicate to parties and candidates which policies or policy platforms we find relatively more appealing, thus helping to minimise our dissatisfaction / making compliance likelier

Prospective voting (Selection)

- Informational demands: Voters need to know parties' platforms, candidates' competence, and their own ideal point → spatial logic: select candidate closest to own ideal point, conditional on competence
- Mandate conception of representation: voters enter into implicit contract with representative, whose performance is evaluated by the extent to which (s)he implements promised platform (Manin, Przeworski, and Stokes, 1999)

Retrospective voting (Sanctioning)

- Informational demands: "In order to ascertain whether the incumbents have performed poorly or well, citizens need only calculate the changes in their own welfare." (Fiorina, 1978, p. 5) → decision rule + attribution of responsibility
- Accountability conception of representation: "voters vote to retain the incumbent only when the incumbent acts in their best interest, and … [she] chooses policies necessary to get re-elected." (Manin, Przeworski, and Stokes, 1999, p. 40)

Ferejohn, 1986: The role of sanctioning

The puzzle: If voters can't observe what politicians do in office, can elections discipline them at all? Crucial question for the accountability conception of representation.

Ferejohn's answer – elections as a sanctioning device:

- Politicians are all identical – same preferences, same abilities. The voter's problem is to police moral hazard (hidden action), not to find the most competent type.
- Voters cannot observe effort directly; they observe only performance.
- The threat of removal disciplines incumbents into exerting effort.

A note on the formal setup. The textbook version by Persson and Tabellini, 2002 uses additive noise $o = e + \varepsilon$ and a single-period FOC. We follow Ferejohn's original multiplicative formulation.

▸ Exogenous vs. endogenous

Model primitives

Each period t:

1. Nature draws a shock $\theta_t \sim F$ on $[0, m]$, with F being continuously differentiable. The incumbent privately observes θ_t.
2. The incumbent chooses effort $a_t \geq 0$ at private cost $\phi(a_t)$, with $\phi' > 0,\ \phi'' > 0,\ \phi(0) = 0$.
3. Voters only observe performance $u_t = a_t \cdot \theta_t$., with voters' utility increasing in performance.
4. The voter applies a stationary cutoff rule: re-elect iff $u_t \geq K$.

Model primitives

Each period t:

1. Nature draws a shock $\theta_t \sim F$ on $[0, m]$, with F being continuously differentiable. The incumbent privately observes θ_t.
2. The incumbent chooses effort $a_t \geq 0$ at private cost $\phi(a_t)$, with $\phi' > 0,\ \phi'' > 0,\ \phi(0) = 0$.
3. Voters only observe performance $u_t = a_t \cdot \theta_t$., with voters' utility increasing in performance.
4. The voter applies a stationary cutoff rule: re-elect iff $u_t \geq K$.

Per-period payoffs:

- Incumbent in office: $W - \phi(a_t)$, where $W > 0$ is the value of office.
- Incumbent out of office: 0 today; with probability λ the loser returns to office at the next election ($\lambda = 1$ two-party, $\lambda = 0$ pure multiparty).

Why repeated elections? Why stationary?

One-shot version is uninteresting: no future $\Rightarrow$ voter cannot punish $\Rightarrow$ incumbent shirks ($a = 0$). Unique SPE is no effort. Voting is pointless.

Infinite horizon with stationary play makes the cutoff rule credible:

- The voter's continuation problem looks the same every period, so re-using the same K is subgame-perfect.
- Define the wedge between staying and losing: $\Delta \equiv \delta(V^I - V^O)$, where V^I is the lifetime value of being in office now, V^O of being out.
- Δ is what the incumbent compares effort costs to – not the per-period W.

Insight: The relevant "reward" for the incumbent is the discounted wedge Δ, which is endogenous (depends on K, F, δ, and λ). We won't solve the full system; we'll work a clean special case where everything has a closed form.

The incumbent's threshold-and-shirk strategy

Given cutoff K and shock θ, the incumbent has two options:

- Clear the threshold: pick $a = K/\theta$, pay $\phi(K/\theta)$, be re-elected with certainty.
- Shirk: pick $a = 0$, pay nothing, lose office.

Comparison. Clearing dominates shirking iff

$$W - \phi(K/\theta) + \delta V^I \;\geq\; W + \delta V^O,$$

which simplifies to: $\phi(K/\theta) \;\leq\; \Delta \quad \Longleftrightarrow \quad \theta \geq \theta^*, \quad$ where $\phi(K/\theta^*) = \Delta$.

Observation: Shirking is on-path – not a deviation. Bad-shock periods see no effort. The voter knows this and still finds it optimal to set K the way she does.

Worked example: uniform shocks, linear cost

Let $F = \text{Uniform}[0,1]$ and $\phi(a) = a$. Then $\phi(K/\theta) = K/\theta$, and the threshold becomes $\theta^* = K/\Delta$.

What does the voter actually see?

- On the clearing branch ($\theta \geq \theta^*$): $u = a\theta = (K/\theta)\theta = K$. Constant!
- On the shirking branch ($\theta < \theta^*$): $u = 0$.

$$\mathbb{E}[u] \;=\; K \cdot \Pr(\theta \geq \theta^*) \;=\; K\big(1 - K/\Delta\big).$$

Worked example: uniform shocks, linear cost

Let $F = \text{Uniform}[0,1]$ and $\phi(a) = a$. Then $\phi(K/\theta) = K/\theta$, and the threshold becomes $\theta^* = K/\Delta$.

What does the voter actually see?

- On the clearing branch ($\theta \geq \theta^*$): $u = a\theta = (K/\theta)\theta = K$. Constant!
- On the shirking branch ($\theta < \theta^*$): $u = 0$.

$$\mathbb{E}[u] \;=\; K \cdot \Pr(\theta \geq \theta^*) \;=\; K\big(1 - K/\Delta\big).$$

Voter's optimum. Maximising the quadratic in K gives

$$K = \tfrac{\Delta}{2}, \qquad \theta^* = \tfrac{1}{2}, \qquad \Pr(\text{re-elect}) = \tfrac{1}{2}, \qquad \mathbb{E}[u] = \tfrac{\Delta}{4}.$$

Interpretation. The incumbent shirks half the time – and this is the welfare-maximising rule. The voter trades off extracting more effort per clearing period (high K) against shirking happening more often (high θ^*).

▸ Step-by-step derivation ▸ Why uniform & linear?

Comparative static: the value of office

Result. Voter welfare U is increasing in the per-period value of office W:

$$\frac{\partial U}{\partial W} > 0.$$

Intuition.

- Larger W raises V^I more than V^O, widening the wedge Δ.
- A wider wedge $\Rightarrow$ the incumbent is willing to clear a tougher threshold $\Rightarrow$ voter can ask for more.

Connection to Przeworski. Recall: $\frac{\partial \pi_{\min}}{\partial S} > 0$. Same logic in two models: when the stakes of office are higher, electoral discipline is stronger – but the demands placed on losers' willingness to wait also rise. The two cut in opposite directions – a tension we will return to.

▸ What does $\partial U/\partial W > 0$ mean?

Counter-intuitive result? Multi-party systems entail more discipline

Recall λ, the probability that today's loser returns to office:

- $\lambda = 0$: "pure multiparty" – losing is final, no return.
- $\lambda = 1$: "pure two-party" – guaranteed alternation.

Result. Voter welfare is decreasing in λ: $\frac{\partial U}{\partial \lambda} < 0$.

Intuition. When losing is less catastrophic (high λ), V^O rises and the wedge Δ shrinks. The threat of removal bites less. Incumbents shirk more.

Why this is striking. The common intuition is that competition disciplines. Ferejohn's logic adds: competition disciplines only when losing actually hurts. We'll return to this in Week 3 on electoral systems.

▸ Why V^O depends on λ

Caveat: assumes no collusion/cooperation across parties (aka coalition government).

Heterogeneity and retrospective control

Now suppose N voters, each caring only about her own share x_i of distributable output. Each voter applies an individual cutoff K_i: re-elect iff $x_i \geq K_i$.

Prop 6 – the collapse. The unique equilibrium has $x_i = 0$ for all $i, a = 0$. Complete loss of electoral control.

Why? The incumbent picks the cheapest minimal majority and pays only those voters; everyone else gets nothing, so their cutoffs become irrelevant. Inside the coalition, voters can be played off against each other.

Prop 7 – the rescue. If voters can coordinate on a sociotropic rule (judge the incumbent on aggregate output), the control is restored.

Empirical link. This rationalises why voters care quite a lot about aggregate economic performance (not just personal economic situation), and motivates Ashworth and Fowler, 2020 on “electorates vs. voters”.

The voter's fundamental trade-off

Cutoffs and their discontents. A higher K:

- Helps: extracts more performance *per period of clearing*, since $u = K$ on the clearing branch.
- Hurts: raises $\theta^* = K/\Delta$, so the incumbent shirks more often.

Three forces alter the trade off (more discipline at every K):

- More valuable office ($W \uparrow$): widens Δ – voter can demand more without losing the incumbent.
- More terminal losses ($\lambda \downarrow$): widens Δ – threat of removal bites harder.
- Sociotropic over individualistic voting (Props 6–7): coordination prevents the incumbent from playing voters off against each other.

Conclusion: simple cutoffs can discipline incumbents *a lot*, but there is a trade-off between effort intensity and frequency, and the size of the wedge Δ sets the ceiling on what's attainable.

Limitations of Ferejohn, 1986 and the road ahead

- No types: deliberate – this is what defines pure sanctioning. Fearon, 1999 adds them next session.
- Stationary cutoffs: Gieczewski and C. Li, 2024 and Acharya, Lipnowski, and Ramos, 2025 show non-stationary (e.g. increasingly lenient) rules can sustain more effort.
- No misattribution: voters react mechanically to outcomes, not to elite cues (Lupia and McCubbins, 1998) or motivated reasoning.
- No challenger heterogeneity: challengers are an exit option, not strategic actors.
- No commitment: the entire bite of the model comes from the voter's inability to precommit. With commitment, the voter could simply mandate effort.

Three ways the literature builds on Ferejohn: selection: how voters screen for "good types", **aggregation**: why electorates can be more rational than individual voters, **information and persuasion**: how elites communicate effectively despite voter ignorance.

From Ferejohn to Fearon: the puzzle of selection

Where Ferejohn leaves us. Pure sanctioning treats all politicians as identical. Voters discipline incumbents by threatening removal.

What Ferejohn cannot explain. A series of empirical patterns about how citizens *actually* think about elections:

- Voters dislike office seekers.
- Voters often support term limits.
- Voters reward principles and consistency, even when politicians shift toward their preferred view.
- Empirical evidence shows no last-period effects.

Fearon's gambit. Maybe voters are not (only) trying to discipline politicians. Maybe they are trying to select good types – candidates who, of their own accord, do what voters would want.

The four empirical patterns

1. Dislike of office seekers. Sanctioning logic says we should *love* reelection-craving politicians – they're maximally responsive. We don't.
2. Support for term limits. Sanctioning logic says term limits remove the very mechanism that disciplines politicians. Yet two-term limits are popular – consistent only with a selection view ("screen out craven office seekers, keep good types").

The four empirical patterns

1. Dislike of office seekers. Sanctioning logic says we should *love* reelection-craving politicians – they're maximally responsive. We don't.
2. Support for term limits. Sanctioning logic says term limits remove the very mechanism that disciplines politicians. Yet two-term limits are popular – consistent only with a selection view ("screen out craven office seekers, keep good types").
3. Premium on principles. Voters punish "waffling" – even when politicians waffle *toward* the median voter's view. Sanctioning logic should reward this. Selection logic: consistency is a costly signal that reveals a politician's type.
4. Absence of last-period effects. Pure sanctioning predicts incumbents shirk in their last term (no future reward). Empirical literature on US Congress finds no clear last-period shirking. Voters reelect retiring politicians without anticipating shirking.

Fearon's claim. These patterns make sense if (a) types vary, and (b) voters can distinguish them at least partially.

What is a “good type”?

Fearon’s definition. For a particular voter, a good type is a politician who:

1. Shares the voter’s policy preferences.
2. Has integrity (hard to bribe or capture).
3. Is competent (skilled at implementing good policy).

What is a “good type”?

Fearon’s definition. For a particular voter, a good type is a politician who:

1. Shares the voter’s policy preferences.
2. Has integrity (hard to bribe or capture).
3. Is competent (skilled at implementing good policy).

What changes if good types exist.

- Even with no electoral accountability (e.g., one-term offices), policy can match voter preferences *if* voters can identify good types.
- Elections become a sorting mechanism, not (or not only) an incentive mechanism.
- The voter’s problem becomes adverse selection (hidden types) on top of moral hazard (hidden actions).

Insight. Selection and sanctioning are not alternatives – they are two channels through which elections produce responsive policy, and they interact.

▸ Adverse selection vs. moral hazard

The two-period model

Players. Electorate E (e.g. the median voter) and an incumbent I.

Period 1:

1. I chooses policy $x \in \mathbb{R}$. Voter's ideal is $x = 0$.
2. Voters observe a noisy signal, $z = -x^2 + \varepsilon$, with $\varepsilon \sim f$ (symmetric, unimodal, mean zero).
3. Voters decide whether to reelect I or draw a challenger from a pool.

The two-period model

Players. Electorate E (e.g. the median voter) and an incumbent I.

Period 1:

1. I chooses policy $x \in \mathbb{R}$. Voter's ideal is $x = 0$.
2. Voters observe a noisy signal, $z = -x^2 + \varepsilon$, with $\varepsilon \sim f$ (symmetric, unimodal, mean zero).
3. Voters decide whether to reelect I or draw a challenger from a pool.

Period 2:

4. Whoever is in office picks y; voter gets utility $-y^2 + \varepsilon'$. Game ends.

Discount factor $\delta \in (0, 1)$. Both periods discounted to today.

Equilibrium concept. Solve by backward induction: work out period-2 play first, plug it into period-1 expectations, then solve for period-1 play. The voter's strategy is restricted to a cut rule: reelect iff $z \geq k$.

Pure selection: Voter updates via Bayes' rule

Setup. Two types: a good type implements $x = 0$ (voter's ideal). A bad type can at best implement $\hat{x} > 0$, giving expected welfare $-\hat{x}^2$. Prior: fraction α good, with $\alpha \in (0, 1)$.

The voter's problem. After seeing welfare z, should I reelect?

Bayes' rule. Update the prior α using the signal z:

$$\underbrace{\alpha'(z)}_{\text{posterior}} \propto \underbrace{f(z \mid \text{good})}_{\text{likelihood, good}} \cdot \underbrace{\alpha}_{\text{prior, good}} .$$

Pure selection: Voter updates via Bayes' rule

Setup. Two types: a good type implements $x = 0$ (voter's ideal). A bad type can at best implement $\hat{x} > 0$, giving expected welfare $-\hat{x}^2$. Prior: fraction α good, with $\alpha \in (0, 1)$.

The voter's problem. After seeing welfare z, should I reelect?

Bayes' rule. Update the prior α using the signal z:

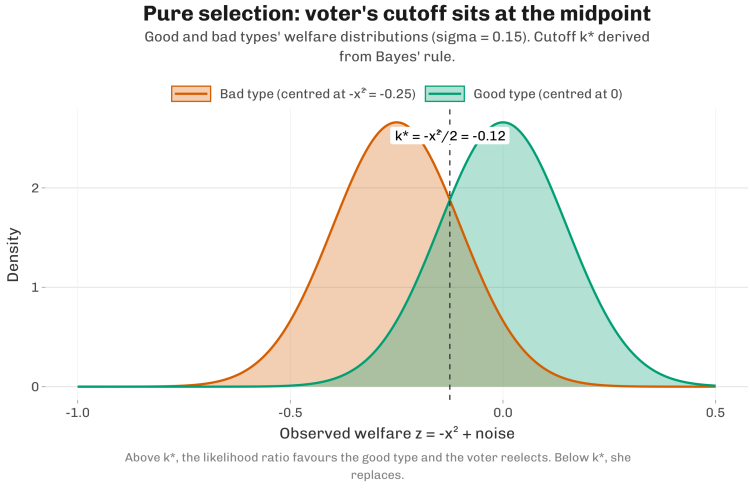
$$\underbrace{\alpha'(z)}_{\text{posterior}} \;\propto\; \underbrace{f(z \mid \text{good})}_{\text{likelihood, good}} \cdot \underbrace{\alpha}_{\text{prior, good}} .$$

Reelect iff posterior $\alpha'(z) \geq \alpha$ (no worse than a new draw). For symmetric f, this gives the midpoint rule:

$$\boxed{k^* \;=\; -\hat{x}^2/2}$$

– the cutoff is halfway between expected welfare under each type.

Visualising pure selection



Each density shows the distribution of observed welfare z for one type: good types centred at 0, bad types at $-\hat{x}^2$. The cutoff $k^* = -\hat{x}^2/2$ is the midpoint. Above k^*, the likelihood ratio favours good $\Rightarrow$ reelect.

Pure sanctioning

Setup. All politicians are identical bad types. They all want $x = 1$ when unconstrained. Period 2 (last period): they pick $y = 1$ – no future, no reward for restraint.

Period 1 incumbent's problem.

$$\max_{x} \;\; W - (1 - x)^2 \;+\; \delta(1 - F(x^2 + k))W.$$

Pure sanctioning

Setup. All politicians are identical bad types. They all want $x = 1$ when unconstrained. Period 2 (last period): they pick $y = 1$ – no future, no reward for restraint.

Period 1 incumbent's problem.

$$\max_{x} \;\; W - (1-x)^2 \;+\; \delta(1 - F(x^2 + k))W.$$

FOC and equilibrium. The voter chooses k to maximise the sensitivity of reelection probability to shirking, which (since f peaks at zero) means setting k at the mode of f. Result:

$$\boxed{x^* = \frac{1}{1 + \delta W f(0)}, \qquad k^* = -(x^*)^2.}$$

- Better monitoring (higher $f(0)$, lower variance) $\Rightarrow$ less shirking.
- Higher $\delta W \Rightarrow$ less shirking (Ferejohn parallel).

Compare with Ferejohn. Ferejohn's incumbent plays a threshold-and-shirk strategy (discontinuous in shocks). Fearon's incumbent picks a smooth interior x^*. Different machinery, same flavour.

▸ Step-by-step PBE ▸ Assumptions on f

The mixed case: setup

Now combine the two views. Prior $\alpha \in (0, 1)$ on good types; $1 - \alpha$ on bad.

Period-2 play (backward induction).

- Good types: dominant strategy $y = 0$ (their ideal).
- Bad types: dominant strategy $y = 1$ (their ideal, no future).

Period-1 strategic logic.

- Good types still pick $x_g = 0$. (Any other choice loses the election *and* gives bad period-1 utility.)
- Bad types optimise their first-period policy x_b, trading off period-1 utility against period-2 reelection.
- Voter uses Bayes' rule: $\alpha'(z) = \alpha \cdot f(x_g^2 + z) \,/\, [\alpha f(x_g^2 + z) + (1 - \alpha) f(x_b^2 + z)]$.

Equilibrium cut rule. For symmetric f, Bayes' rule gives $k = -x_b^2/2$, i.e. the midpoint, as in pure selection. Combined with the bad type's FOC, this implicitly defines equilibrium (k^*, x_b^*).

The commitment problem

Two cutoffs the voter could use:

- Ex-ante optimal (pure-sanctioning logic): set k at the mode of f to maximise the bad type's *incentive* not to shirk.
- Ex-post optimal (selection logic): set $k = -x_b^2/2$ to identify which incumbents are most likely good *after seeing the data*.

The pathology. These two are different. The voter would *like* to commit to the ex-ante optimal cutoff, but cannot: once she sees z, Bayes' rule tells her something about type, and she rationally wants to use it.

Why this is not just "credibility". In Ferejohn, the SPE cutoff is credible because the voter is indifferent at the polls. In Fearon's mixed case, the voter is not indifferent – she wants to update – so commitment to the sanctioning-optimal cutoff fails.

Time inconsistency. The voter's ex-ante best plan and her ex-post best action diverge. She cannot bind her future self.

▸ Commitment vs. credibility

Central result: a tiny good-type chance hurts voters

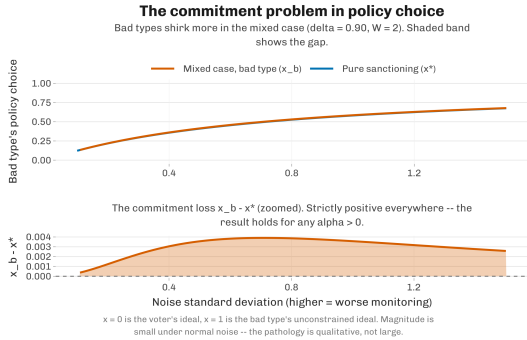
The result. For any $\alpha > 0$,

$$x_b^* \;>\; x^*_{\text{pure sanctioning}}.$$

Bad types shirk more in the mixed case. The result holds for very small α.

The mechanism. Voter's selection logic raises the cutoff k above the sanctioning-optimal level, weakening the bad type's incentive to choose policy near $x = 0$.

Discontinuity at $\alpha = 0$. The mixed-case equilibria do *not* converge to the pure-sanctioning equilibrium as $\alpha \to 0$. Selection at the polls qualitatively changes the game.



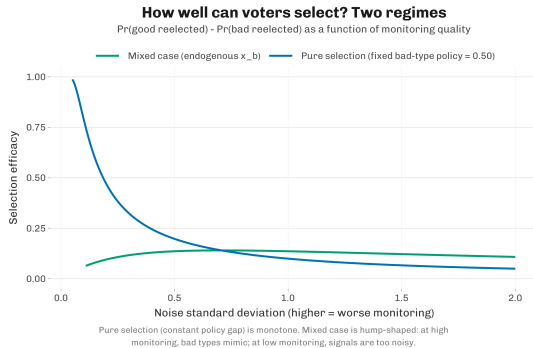
Empirical implications

Inverted U in selection efficacy.

- Very good monitoring ⇒ bad types mimic closely ⇒ hard to distinguish.
- Very bad monitoring ⇒ signal is too noisy ⇒ again hard to distinguish.
- Maximum efficacy at moderate monitoring quality.

Tenure effects.

- Time series (same politician across terms): more shirking with tenure (last-period effect for bad types).
- Cross section (compare politicians by tenure): *less* shirking among long-tenured representatives – the long-tenured pool is enriched in good types.



Lessons from the selection-meets-sanctioning model

- Elections serve a dual role: **selection** of good types *and* **sanctioning** of bad behaviour.
- These two roles interact strategically, each channel can undermine the other:
  - → Selection at the polls undermines sanctioning – voters cannot help but use type information, so they cannot credibly commit to the sanctioning-optimal cutoff (commitment problem).
  - → Strong sanctioning incentives make bad types mimic good ones, shrinking the policy gap voters can detect (inverted-U in monitoring).
- Even with noisy outcomes, cutoff rules let voters control politicians – how much depends on the type pool, the discount factor, and monitoring quality.
- Fearon, 1999 deepens Ferejohn, 1986 by adding unobservable types: voters care about *who* politicians are, not just what they do.
- Connects to a broader debate: are elections devices for **mandate** (selecting good types) or **accountability** (sanctioning bad performance)? Both, and they trade off (Manin, Przeworski, and Stokes, 1999).

Limitations of Fearon, 1999 and the road ahead

- Two-period horizon: the unravelling that motivates selection is partly artefactual; infinite-horizon models smooth this out (Banks & Sundaram 1996).
- Stationary cutoffs: voters cannot tailor standards to history or context. Non-stationary, increasingly lenient rules can sustain more effort (Acharya, Lipnowski, and Ramos, 2025; Gieczewski and C. Li, 2024).
- Single dimension of type variation: politicians differ only on policy preferences. Besley, 2006 adds integrity and competence as further type dimensions.
- No endogenous entry: who runs for office is taken as given.

- **More sophisticated retention rules**: Anesi and Buisseret, 2022 show that *stochastic* or state-contingent retention rules let voters partially overcome the commitment problem.
- **Aggregation and information**: Ashworth and Fowler, 2020 on why electorates can be more rational than individual voters; Lupia and McCubbins, 1998 on cues and elite communication.
- **Principled agents**: Besley, 2006 integrates types with mission motivation and political-agency theory more broadly.

The existential relevance of Ferejohn and Fearon

1. “The unlucky incumbent” – Ferejohn at fast forward.
A government in office during an exogenous crisis (oil shock, financial crisis, gilt-market panic). Voters cannot cleanly disentangle effort from luck; if observed welfare drops far enough, the cutoff is crossed and the incumbent is removed – regardless of fault. Threshold-and-shirk in real time.

2. “The political newcomer” – Fearon’s pure selection.
A candidate stands with no governing record. Voters cannot sanction past performance because there is none. Their choice is a pure inference about type – competence, integrity, ideological proximity – read off campaign signals.

3. “The persistent scandal-survivor” – the commitment problem.
An incumbent linked to serious misconduct keeps winning. Voters’ rational selection on type at the polls (“yes, but *our* team”) overrides retrospective sanctioning. Bad types persist precisely because voters cannot commit to ignore type information once they observe it.

Zooming out briefly: Why is accountability so difficult?

Theoretical assumptions

- **Prospective voting**:
  - → Voters know platforms and understand policy consequences
  - → Politicians seek reelection by keeping promises
  - → Effective selection of good politicians
- **Retrospective voting**:
  - → Voters know about outcomes
  - → Voters attribute responsibility accurately (disentangle exogenous factors from endogenous ones)
  - → Effective sanctioning of bad incumbents

Empirical challenges

- **Prospective voting**:
  - → Blurry platforms or partisan intoxication
  - → Lack of political knowledge or interest
  - → Not always optimal to punish deviations
  - → Inability to assess consequences of platforms
- **Retrospective voting**:
  - → Attribution difficult (e.g., exogenous shocks or partisan intoxication)
  - → Multi-dimensionality: one vote, many issues: "One cannot control many targets with one instrument" (Przeworski, 2018, p. 96)

Illustration #1: Prospective voting and the mandate conception of accountability

> *When the government announces that the conditions are not what they were anticipated to be, voters cannot be sure if implementing the mandate is still in their best interest. And if implementing the mandate is not the best the government can do, the threat of punishing incumbents who deviate from it is not credible (Przeworski, 2018, p. 94)*

Illustration #2: Do people actually have policy preferences, or are they just blind partisans?

Notable challenge: Achen and Bartels, 2017 argue that voters do not really have independent policy preferences; their preferences are entirely endogenous to their social groups → more extreme version: partisan intoxication or cheerleading

FYI: My reading of the empirical evidence is different, but this is a theoretical class

More importantly: Fowler, 2020 argues convincingly that much of the empirical evidence adduced in favour of that hypothesis is observationally equivalent with policy voting (see table on right)

Evidence	Intoxication interpretation	Policy voting interpretation
Party ID predicts vote choice.	Party identification influences vote choices.	Policy preferences influence party identification and vote choices.
Party ID is correlated across generations	People blindly adopt the party of their parents.	Values, economic circumstances, and policy preferences are correlated across generations.
Party ID is stable over time — even more so than issue positions.	People blindly stick with whatever party they initially adopted, even when their policy positions change.	Policy preferences are correlated over time, surveys measure policy positions imperfectly, and party identification reflects a complex weighted average of positions across a range of issues.
Elite cues sometimes influence issue positions.	People don't care about policy and they blindly adopt whatever their party leaders say.	Rational, policy voters should sometimes take cues from elites that share their values and interests.
Most Republicans support President Trump despite his personal shortcomings.	People blindly support leaders from their party, even when it cuts against their personal interests.	Despite Trump's personal shortcomings, many Republicans likely support him because of his policy positions. Furthermore, Republican loyalty to Trump is overstated since people change their party identification in response to their approval of the president.
When the parties changed positions on racial issues, white southerners were slow to change their party identification.	Party identification is sticky and largely unrelated to policy interests.	But they did change their voting behavior. And to the extent that they continued to support Democrats in non-presidential elections, it was only the conservative Democrats who agreed with them on policy.

Illustration #3: Partisan intoxication and retrospective accountability: A topical example?

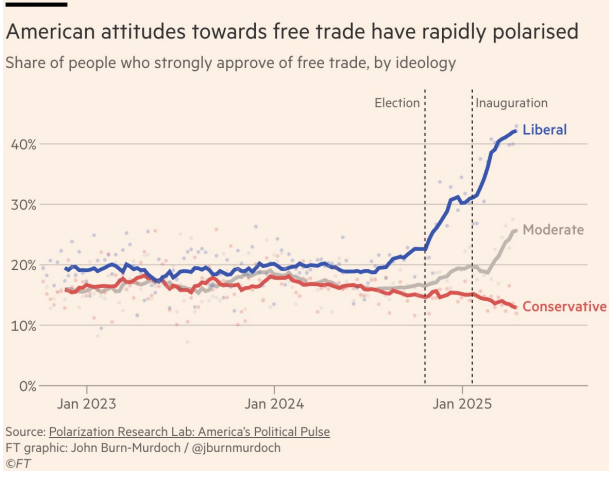


Illustration #4: Exogenous events, attribution of responsibility, and retrospective accountability

Much-discussed empirical challenge to retrospective voting: Exogenous voters – such as shark attacks (Achen and Bartels, 2017) or wins of college football teams – affect vote choice → clear evidence that voters are bad at attribution (Caplan, 2008)?

- Empirically, these findings are very much contested (Fowler and Hall, 2018; Fowler and Montagnes, 2015)

More important point for our purposes: Should we theoretically expect (rational) voters to *not* respond to exogenous events? What events should they respond to, and how? Should we expect voter competence to lead to better accountability?

Answering this question requires theory. Ashworth and Bueno De Mesquita, 2014; Ashworth, Bueno de Mesquita, and Friedenberg, 2018 address this challenge.

Illustration #4 (cont'd)

We won't delve into their argument – this is too complicated for an introductory course like this.

Central intuitions:

1. Some exogenous shocks (e.g., hurricanes) – those that affect voter welfare *and* are related to government performance – provide voters with valuable information about incumbents' types (e.g., their competence). Can foster better selection; rational voters should respond to some shocks.
2. Whether accountability is undermined by voters responding to exogenous events depends on how more or less voter competence affects politicians' incentives. Effect can be non-monotonic: more competence can be bad when it for instance induces pandering

$\rightarrow$ Great examples of how an empirical debate inspires theoretical innovation, which helps us understand better what evidence is necessary to adjudicate between different perspectives

Summary: The Global-North evidence on *selection*

Selection on competence (Sweden).

- Dal Bó et al., 2017: using population data on cognitive ability and leadership scores from Swedish military enlistment, politicians at all levels of government are positively selected on competence – higher than the citizens they represent, even controlling for family background and social class.
- Democracy does attract capable people. Selection is not a fiction.

Selection on policy alignment (US primaries).

- Hall, 2015: regression-discontinuity design in US House primaries. When an extremist beats a moderate in a coin-flip primary, the party's general-election vote share falls by 9–13 percentage points and the seat probability by 35–54 points.
- General-election voters punish ideological extremism – consistent with Fearon-style selection on type.

Why selection works: voters can use coarse signals – extremism, scandals, party labels, candidate biography – without needing fine-grained policy information.

Summary: The evidence on *sanctioning*

Scandal punishment is real but small.

- Eggers, 2014: the 2009 UK expenses scandal cost implicated MPs only 1.5–2.6 percentage points of vote share – voters were "very angry" but punishment was modest.
- Costas-Pérez, Solé-Ollé, and Sorribas-Navarro, 2012: in Spain, corruption costs incumbents up to 14% only when charges are formal *and* press coverage is extensive. Otherwise much smaller.

The information environment matters.

- Hayes and Lawless, 2018: where US local newspapers close, political knowledge falls, government efficiency drops, and municipal bond yields rise. Direct quasi-experimental confirmation of the Snyder-Strömberg mechanism – and a warning as local journalism continues to decline.

Politicians vary enormously in office-vs-policy motivation.

- Iaryczower, Lopez-Moctezuma, and Meirowitz, 2024: structural estimates from US Senate data find over 25% of senators are "highly ideological" – placing little weight on reelection relative to policy. The Ferejohn/Fearon W parameter is empirically heterogeneous, precisely where theory predicts accountability bites least.

Lupia and McCubbins, 1998: “The Democratic Dilemma”

> *It is widely believed that there is a mismatch between the requirements of democracy and most people’s ability to meet these requirements. If this mismatch is too prevalent, then effective self-governance is impossible. The democratic dilemma is that the people who are called upon to make reasoned choices may not be capable of doing so. (Lupia and McCubbins, 1998, p. 1)*

Where this leaves us. Schumpeter, 1942 and Arrow, 1951 already showed we can’t ground democratic legitimacy in faithful aggregation of *all* citizen preferences. Ferejohn and Fearon then offered a fall-back: voters need only retrospectively sanction (or select) – limited information might suffice.

The remaining question. What if voters are too inattentive even for that? Can informational shortcuts – party cues, endorsements, expert testimony – substitute for personal knowledge?

Lupia & McCubbins: when can heuristics substitute for knowledge?

The argument. Citizens lacking time or expertise can still make reasoned choices *if* they can rely on credible cues from informed sources – endorsements, expert testimony, party labels, interest-group ratings. The key is identifying when sources are credible.

Lupia & McCubbins's conditions for credible cues:

- Common interests: sender's preferences are sufficiently aligned with receiver's.
- Penalties for lying: detected dishonesty is punished (legal liability, reputation loss).
- Verification: receiver can check the message against an independent source.
- Costly statements: messages are credible when costlier to mimic for those without good information.

Real-world institutional support. Elections, public scrutiny, media competition, oversight bodies, transparency rules all create the conditions above – they are why heuristics can work in practice, not just in theory.

Methodological interlude: how is persuasion possible at all?

The puzzle. If a sender (politician, lobbyist, pundit) has incentives to mislead, why would a Bayesian receiver believe anything they say?

Little, 2023's taxonomy: five mechanisms by which persuasion is rationally possible.

1. Costly signalling: messages cost more for those without favourable information to mimic (e.g. campaign spending, principled stands).
2. Common interests: aligned preferences make truth-telling self-enforcing – the original Crawford and Sobel, 1982 *cheap-talk* setting.
3. Verifiable information: claims that can be externally checked (audits, fact-checking, scientific consensus).
4. Reputation concerns: senders want to be seen as honest or competent in future interactions.
5. Commitment to a messaging strategy: sender pre-commits to a signal structure – the modern Bayesian persuasion framework a la Kamenica and Gentzkow, 2011.

Little's take. Despite the recent dominance of (5) in the formal-theory literature, the older mechanisms (1)–(4) do most of the heavy lifting in explaining real political phenomena – campaigns, partisan media, lobbying.

Back to Lupia conditions

Empirical challenge #1: heuristic projection

Even with credible institutions, voters can mis-construe the cues themselves.

Central finding: Broockman, Kaufman, and Lenz, 2024 – "heuristic projection."

- When an interest group endorses or rates a representative, voters often assume the group shares their own views – even when the group is ideologically misaligned with them.
- Across three studies using real interest groups (e.g. FreedomWorks, NRA) and participants' actual representatives.
- Result: misaligned cues can *increase* approval for representatives whose actual positions are far from the voter's.

Why this matters for the Lupia-McCubbins argument.

- Lupia-McCubbins assume voters can identify whether a source is aligned. Broockman et al. show this is exactly where voters fail.
- Interest groups can amplify the failure by adopting ambiguously positive names ("FreedomWorks," "Americans for…") – making projection easier and harder to correct.

Empirical challenge #2: motivated reasoning

Little, Schnakenberg, and Turner, 2022 formalise a second failure: voters update on cues with directional motives, not only accuracy motives. Rewards for "believing what one wants" compete with the truth-tracking incentive.

Two distinct effects.

- Divergence: partisans with different priors form different conclusions about the same incumbent's performance.
- Desensitisation: the relationship between performance and conclusions *weakens* – voters respond less to information of any kind.

Their key result. Desensitisation, not divergence, is the more uniformly bad effect for accountability. Polarised partisans whose conclusions diverge can still discipline incumbents in parallel; desensitised centrists cannot.

When do elite cues actually move people?

1. Reception – does the cue reach the receiver?

- Politically aware citizens receive more cues from more sources (Zaller, 1992 RAS framework).
- Self-selection into partisan media environments
- Null results at the population level can reflect reception failure, not acceptance failure.

When do elite cues actually move people?

1. Reception – does the cue reach the receiver?

- Politically aware citizens receive more cues from more sources (Zaller, 1992 RAS framework).
- Self-selection into partisan media environments
- Null results at the population level can reflect reception failure, not acceptance failure.

2. Persuasion – given reception, does the receiver update?

- Strength of prior: high-awareness citizens are easier to reach but harder to convince (Zaller, 1992).
- Issue importance: cues bite less on issues people care about (Barber and Pope, 2024).
- Material self-interest: harder to move voters against their material conditions (Cavaillé and Neundorf, 2023).
- Source credibility: same source has different credibility for co- vs. out-partisans (Druckman, 2022).
- Counter-cues: framing effects attenuate sharply when credible counter-frames are available (Druckman, 2022).

3. Posterior – if accepted, what changes?

- Direction, magnitude, durability, scope, and the attitude-behaviour link (Coppock, 2023).
- Magnitude effects are surprisingly homogeneous across partisan subgroups; apparent asymmetries are likely driven by reception-stage exposure differences, not acceptance-stage psychology.

▸ Information-provision experiments ▸ Priming experiments

Is democracy broken? Revisiting the role of voters

Many political scientists claim **voters are irrational**, uninformed, and swayed by irrelevant factors (Achen and Bartels, 2017).

Heuristics and elite cues might help, but the argument is both theoretically and empirically not as strong as we would like to confidently believe in the efficacy of electoral accountability.

This raises a crucial question: Do flawed voters necessarily lead to flawed democratic outcomes?

Voters vs. electorates

- A key distinction: the **individual voter** is not the same as the **electorate**.
- Many voters may be:
  - → Partisan
  - → Poorly informed
  - → Susceptible to bias or heuristics
- Yet elections often yield **reasonable aggregate outcomes**:
  - → Sensitivity to candidate quality and ideology
  - → Responsiveness to government performance
- Ashworth and Fowler, 2020: What matters is not the “typical voter,” but the **pivotal voters** whose behaviour decides outcomes.

Theoretical frameworks: Why electorates can be rational, even if most voters are not

- **Statistical argument (The Macro Polity, Condorcet Jury Theorem):**
  - → If individual voter errors are uncorrelated, they "cancel out" in the aggregate (Erikson, Mackuen, and Stimson, 2008).
  - → Assumes independence — a strong and often unrealistic condition.
- **Strategic pivotal voter argument (Ashworth and Fowler, 2020):**
  - → Electoral outcomes are determined by swing voters or those sensitive to candidate quality.
  - → Politicians have incentives to exert effort even if most voters are unresponsive.
- **Endogenous information acquisition:**
  - → Pivotal voters may **seek out more information**, improving aggregate outcomes.
  - → Supported by Mikosch et al., 2024 and other studies on motivated learning and turnout.

▸ Refresher on Condorcet's Jury Theorem

Implications for democratic theory and empirical research

- **Flawed voters do not imply flawed democracy.**
  - $\rightarrow$ Aggregate behaviour can exhibit rationality even if individuals do not.
- **Pivotal voters and turnout dynamics** help preserve electoral accountability.
- **Scholarly focus should shift:**
  - $\rightarrow$ From voter irrationality to electorate-level responsiveness.
  - $\rightarrow$ From idealised norms to empirical mechanisms of accountability.
- **Caution against overgeneralising**: Poor individual-level knowledge or behaviour $\neq$ democratic failure.

Outline

Course logistics and preliminaries

Introducing the core themes

Four big ideas about electoral accountability

- Ferejohn's model of electoral control
- Fearon's selection-and-sanctioning model
- Informational shortcuts to the rescue?
- Electorates vs. voters

Takeaways and outlook

Taking stock: What have we learned?

- From ideals to institutions: Democracy's original vision of self-government is infeasible at scale; modern democracy relies on representation and rules-based competition to manage conflicts peacefully.
- The second-best ideal: Self-rule becomes regulated conflict — elections distribute power peacefully, making outcomes tolerable for winners and losers alike.
- Elections and incentives: Formal models show how elections can discipline politicians — but only under the right conditions (stakes, noise, expectations).
- Limits of accountability: Voters face trade-offs between sanctioning and selection. Misattribution, low information, or high stakes can erode incentives.
- Trust and shortcuts: Heuristics and elite cues help voters navigate complexity — if sources are credible and institutions foster trust.
- Why democracy can work anyway: Even if individual voters are flawed, electorates can still behave rationally. What matters most is pivotal voters and the structure of competition.

References I

Acharya, Avidit, Elliot Lipnowski, and João Ramos (2025). “Political accountability under moral hazard”. *American Journal of Political Science* 69.2, pp. 641–652. DOI: 10.1111/ajps.12860.

Achen, Christopher H. and Larry M. Bartels (Aug. 2017). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, Princeton University Press.

Anesi, Vincent and Peter Buisseret (Nov. 2022). “Making Elections Work: Accountability with Selection and Control”. *American Economic Journal: Microeconomics* 14.4, pp. 616–644. DOI: 10.1257/mic.20200311.

Arrow, Kenneth J. (1951). *Social Choice and Individual Values*. 3rd ed. New Haven: Yale University Press.

Ashworth, Scott and Ethan Bueno De Mesquita (Aug. 2014). “Is Voter Competence Good for Voters?: Information, Rationality, and Democratic Performance”. *American Political Science Review* 108.3, pp. 565–587. DOI: 10.1017/S0003055414000264.

References II

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg (2018). “Learning about Voter Rationality”. *American Journal of Political Science* 62.1, pp. 37–54. DOI: 10.1111/ajps.12334.

Ashworth, Scott and Anthony Fowler (Aug. 2020). “Electorates versus Voters”. *Journal of Political Institutions and Political Economy* 1.3, pp. 477–505. DOI: 10.1561/113.00000016.

Barber, Michael and Jeremy C. Pope (Apr. 2024). “Does issue importance attenuate partisan cue-taking?” en. *Political Science Research and Methods* 12.2, pp. 435–443. DOI: 10.1017/psrm.2023.28.

Bates, Robert H. and Da-Hsiang Lien (Mar. 1985). “A Note on Taxation, Development, and Representative Government”. en. *Politics & Society* 14.1, pp. 53–70. DOI: 10.1177/003232928501400102.

Besley, Timothy (2006). *Principled Agents?: The Political Economy of Good Government*. Englisch. Oxford: Oxford University Press. ISBN: 978-0-19-928391-0.

References III

Broockman, David E., Aaron R. Kaufman, and Gabriel S. Lenz (Jan. 2024). "Heuristic Projection: Why Interest Group Cues May Fail to Help Citizens Hold Politicians Accountable". *British Journal of Political Science* 54.1, pp. 69–87. DOI: 10.1017/S0007123423000078.

Burn-Murdoch, John (Mar. 2024). "American politics is undergoing a racial realignment". *Financial Times*.

Caplan, Bryan (Aug. 2008). *The Myth of the Rational Voter: Why Democracies Choose Bad Policies (New Edition): Why Democracies Choose Bad Policies*. English. Princeton, NJ: Princeton University Press.

Carey, John M. and Oscar Pocasangre (2024). *Can Proportional Representation Lead to Better Governance?* Tech. rep. Protect Democracy | New America.

Cavaillé, Charlotte and Anja Neundorf (Dec. 2023). "Elite Cues and Economic Policy Attitudes: The Mediating Role of Economic Hardship". *Political Behavior* 45.4, pp. 1355–1376. DOI: 10.1007/s11109-021-09768-w.

References IV

Chung, Hun and John Duggan (Feb. 2020). "A Formal Theory of Democratic Deliberation". en. *American Political Science Review* 114.1, pp. 14–35. DOI: 10.1017/S0003055419000674.

Coppock, Alexander (Jan. 2023). *Persuasion in Parallel: How Information Changes Minds about Politics*. English. Chicago: University of Chicago Press. ISBN: 978-0-226-82184-9.

Costas-Pérez, Elena, Albert Solé-Ollé, and Pilar Sorribas-Navarro (Dec. 2012). "Corruption scandals, voter information, and accountability". *European Journal of Political Economy* 28.4, pp. 469–484. ISSN: 0176-2680. DOI: 10.1016/j.ejpoleco.2012.05.007.

Crawford, Vincent P. and Joel Sobel (1982). "Strategic Information Transmission". *Econometrica* 50.6, pp. 1431–1451. ISSN: 0012-9682. DOI: 10.2307/1913390.

Dal Bó, Ernesto et al. (Nov. 2017). "Who Becomes A Politician?*". *The Quarterly Journal of Economics* 132.4, pp. 1877–1914. ISSN: 0033-5533. DOI: 10.1093/qje/qjx016.

References V

Dowding, Keith (Jan. 2006). “Can Populism Be Defended? William Riker, Gerry Mackie and the Interpretation of Democracy”. en. *Government and Opposition* 41.3, pp. 327–346. DOI: 10.1111/j.1477-7053.2006.00182.x.

Druckman, James N. (Nov. 2004). “Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects”. en. *American Political Science Review* 98.4, pp. 671–686. DOI: 10.1017/S0003055404041413.

— **(May 2022).** “A Framework for the Study of Persuasion”. en. *Annual Review of Political Science* 25.Volume 25, 2022, pp. 65–88. DOI: 10.1146/annurev-polisci-051120-110428.

Drutman, Lee (2021). “Elections, Political Parties, and Multiracial, Multiethnic Democracy: how the United States Gets It Wrong”. *New York University Law Review* 96, p. 985.

Dryzek, John S. and Christian List (2003). “Social Choice Theory and Deliberative Democracy: A Reconciliation”. *British Journal of Political Science* 33.1, pp. 1–28.

References VI

Dunn, John (Nov. 2018). *Setting the People Free: The Story of Democracy, Second Edition*. English. 2nd edition. Princeton: Princeton University Press.

Eggers, Andrew C. (Dec. 2014). "Partisanship and Electoral Accountability: Evidence from the UK Expenses Scandal*". *Quarterly Journal of Political Science* 9.4, pp. 441–472. ISSN: 1554-0626. DOI: 10.1561/100.00013140.

Erikson, Robert S., Michael B. Mackuen, and James A. Stimson (Jan. 2008). *The Macro Polity*. Cambridge: Cambridge University Press.

Fearon, James D. (1999). "Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance". *Democracy, Accountability, and Representation*. Ed. by Adam Przeworski, Bernard Manin, and Susan C. Stokes. Cambridge: Cambridge University Press, pp. 55–97.

References VII

Fearon, James D. (Nov. 2011). “Self-Enforcing Democracy”. *The Quarterly Journal of Economics* 126.4, pp. 1661–1708. ISSN: 0033-5533. DOI: 10.1093/qje/qjr038.

Ferejohn, John (Jan. 1986). “Incumbent performance and electoral control”. en. *Public Choice* 50.1, pp. 5–25. ISSN: 1573-7101. DOI: 10.1007/BF00124924.

Fiorina, Morris P. (1978). “Economic Retrospective Voting in American National Elections: A Micro-Analysis”. *American Journal of Political Science* 22.2, pp. 426–443. DOI: 10.2307/2110623.

Fishkin, James (July 2018). *Democracy When the People Are Thinking: Revitalizing Our Politics Through Public Deliberation*. English. Oxford: Oxford University Press.

Fishkin, James, Valentin Bolotnyy, et al. (Nov. 2024). “Can Deliberation Have Lasting Effects?” en. *American Political Science Review* 118.4, pp. 2000–2020. DOI: 10.1017/S0003055423001363.

— **(Jan. 2025).** “Scaling Dialogue for Democracy: Can Automated Deliberation Create More Deliberative Voters?” en. *Perspectives on Politics*, pp. 1–18. DOI: 10.1017/S1537592724001749.

References VIII

Fishkin, James, Alice Siu, et al. (Nov. 2021). “Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on “America in One Room””. en. *American Political Science Review* 115.4, pp. 1464–1481. DOI: 10.1017/S0003055421000642.

Fowler, Anthony (Apr. 2020). “Partisan Intoxication or Policy Voting?” English. *Quarterly Journal of Political Science* 15.2, pp. 141–179. DOI: 10.1561/100.00018027a.

Fowler, Anthony and Andrew B. Hall (Oct. 2018). “Do Shark Attacks Influence Presidential Elections? Reassessing a Prominent Finding on Voter Competence”. *The Journal of Politics* 80.4, pp. 1423–1437. ISSN: 0022-3816. DOI: 10.1086/699244.

Fowler, Anthony and B. Pablo Montagnes (Nov. 2015). “College football, elections, and false-positive results in observational research”. *Proceedings of the National Academy of Sciences* 112.45, pp. 13800–13804. DOI: 10.1073/pnas.1502615112.

References IX

Gaertner, Wulf (June 2009). *A Primer in Social Choice Theory: Revised Edition*. English. Revised edition. Oxford ; New York: Oxford University Press, USA.

Gerring, John and Wouter Veenendaal (July 2020). *Population and Politics: The Impact of Scale*. English. Cambridge: Cambridge University Press.

Gieczewski, Germán and Christopher Li (June 2024). "Incumbent Performance and Electoral Control: A Comment". *Quarterly Journal of Political Science* 19.3, pp. 331–353. DOI: 10.1561/100.00023038.

Goodin, Robert E. and Kai Spiekermann (May 2018). *An Epistemic Theory of Democracy*. English. Oxford: Oxford University Press.

Grofman, Bernard and Scott L. Feld (1988). "Rousseau's General Will: A Condorcetian Perspective". *The American Political Science Review* 82.2, pp. 567–576. DOI: 10.2307/1957401.

Hall, Andrew B. (Feb. 2015). "What Happens When Extremists Win Primaries?" en. *American Political Science Review* 109.1, pp. 18–42. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055414000641.

References X

Hayes, Danny and Jennifer L. Lawless (Jan. 2018). "The Decline of Local News and Its Effects: New Evidence from Longitudinal Data". *The Journal of Politics* 80.1, pp. 332–336. ISSN: 0022-3816. DOI: 10.1086/694105.

Hibbing, John R. and Elizabeth Theiss-Morse (2002). *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge: Cambridge University Press.

Iaryczower, Matias, Gabriel Lopez-Moctezuma, and Adam Meirowitz (2024). "Career Concerns and the Dynamics of Electoral Accountability". en. *American Journal of Political Science* 68.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12740, pp. 696–713. ISSN: 1540-5907. DOI: 10.1111/ajps.12740.

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.

References XI

Landa, Dimitri and Adam Meirowitz (2009). “Game Theory, Information, and Deliberative Democracy”. *American Journal of Political Science* 53.2, pp. 427–444. DOI: 10.1111/j.1540-5907.2009.00379.x.

Latner, Michael, Jack Santucci, and Matthew Shugart (Aug. 2021). *Multi-seat Districts and Larger Assemblies Produce More Diverse Racial Representation*. DOI: 10.2139/ssrn.3911532.

Li, Yuhui (Sept. 2019). *Dividing the Rulers: How Majority Cycling Saves Democracy*. English. Michigan: University of Michigan Press.

Little, Andrew T. (July 2023). “Bayesian explanations for persuasion”. en. *Journal of Theoretical Politics* 35.3, pp. 147–181. ISSN: 0951-6298. DOI: 10.1177/09516298231185060.

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.

References XII

Lupia, Arthur and Mathew D. McCubbins (1998). *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge: Cambridge University Press.

Mainwaring, Scott and Lee Drutman (2023). *The Case for Multiparty Presidentialism in the US*. Tech. rep. Protect Democracy | New America.

Manin, Bernard, Adam Przeworski, and Susan C. Stokes (1999). "Elections and Representation". *Democracy, Accountability, and Representation*. Ed. by Adam Przeworski, Susan C. Stokes, and Bernard Manin. Cambridge: Cambridge University Press, pp. 29–54.

Maskin, Eric (Mar. 2020). "A modified version of Arrow's IIA condition". en. *Social Choice and Welfare* 54.2-3, pp. 203–209. DOI: 10.1007/s00355-020-01241-7.

— **(Feb. 2025).** "Borda's Rule and Arrow's Independence Condition". *Journal of Political Economy* 133.2, pp. 385–420. DOI: 10.1086/732892.

References XIII

Mikosch, Heiner et al. (Apr. 2024). “Uncertainty and Information Acquisition: Evidence from Firms and Households”. *American Economic Journal: Macroeconomics* 16.2, pp. 375–405. DOI: 10.1257/mac.20220047.

Niemeyer, Simon et al. (Feb. 2024). “How Deliberation Happens: Enabling Deliberative Reason”. en. *American Political Science Review* 118.1, pp. 345–362. DOI: 10.1017/S0003055423000023.

Persson, Torsten and Guido Tabellini (2002). *Political Economics*. en-US. Cambridge, MA: MIT Press.

Przeworski, Adam (1991). *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139172493.

— **(June 2005).** “Democracy as an equilibrium”. en. *Public Choice* 123.3, pp. 253–273. ISSN: 1573-7101. DOI: 10.1007/s11127-005-7163-4.

— **(2010).** *Democracy and the Limits of Self-Government*. Englisch. Cambridge: Cambridge University Press.

References XIV

Przeworski, Adam (Jan. 2018). *Why Bother With Elections?* Englisch. Cambridge, UK: Polity.

Przeworski, Adam, Gonzalo Rivero, and Tianyang Xi (Sept. 2015). "Elections as a conflict processing mechanism". *European Journal of Political Economy* 39, pp. 235–248. DOI: 10.1016/j.ejpoleco.2015.05.006.

Riker, William H. (1982). *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. English. Prospect Heights, Ill: Waveland.

Ruffini, Patrick (Nov. 2023). *Party of the People: Inside the Multiracial Populist Coalition Remaking the GOP*. English. New York: Simon & Schuster.

Santucci, Jack (Aug. 2022). *More Parties or No Parties: The Politics of Electoral Reform in America*. English. Oxford: Oxford University Press.

Schumpeter, Joseph A. (1942). *Capitalism, Socialism, and Democracy: Third Edition*. Englisch. 3rd Edition. New York: Harper Perennial Modern Classics.

References XV

Sen, Amartya (2017). *Collective Choice and Social Welfare*. Expanded Edition. London: Penguin.

Stasavage, David (June 2020). *The Decline and Rise of Democracy: A Global History from Antiquity to Today*. Princeton (N.J.): Princeton University Press.

Sunstein, Cass R. (2000). "Deliberative Trouble? Why Groups Go to Extremes". *The Yale Law Journal* 110.1, pp. 71–119. DOI: 10.2307/797587.

— **(2002).** "The Law of Group Polarization". en. *Journal of Political Philosophy* 10.2, pp. 175–195. DOI: 10.1111/1467-9760.00148.

Veri, Francesco (Feb. 2025). "Fostering Reasoning in the Politically Disengaged: The Role of Deliberative Minipublics". *Political Studies Review* 23.1, pp. 254–275. DOI: 10.1177/14789299241239909.

Weale, Albert (1984). "Social Choice versus Populism? An Interpretation of Riker's Political Theory". *British Journal of Political Science* 14.3, pp. 369–385.

References XVI

Wilson, Robert (Dec. 1972). “Social choice theory without the Pareto Principle”. *Journal of Economic Theory* 5.3, pp. 478–486. DOI: 10.1016/0022-0531(72)90051-8.

Zaller, John (Aug. 1992). *The Nature and Origins of Mass Opinion*. en. Google-Books-ID: 83yNzu6toisC. Cambridge : New York: Cambridge University Press. ISBN: 978-0-521-40786-1.

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Grofman and Feld, 1988 on Rousseau's *Volonté Générale*

Three central elements of Rousseau's notion of the common good:

- There exists a common good independent of individual wills.
- Citizens have imperfect individual judgement about the common good.
- Collective decision-making by majority vote is the most reliable method for approximating the common good.

Connection to Condorcet's Jury Theorem:

- If individuals vote sincerely and independently, majority rule converges to correct decisions as group size increases.

A refresher on Condorcet's jury theorem

Setup:

- There is a binary decision: correct (C) or incorrect (I).
- Each voter i votes independently with probability $p > 0.5$ of choosing C correctly.
- Majority rule is used to determine the collective decision.

Theorem (Condorcet):

As the number of voters $n \to \infty$,

$$\mathbb{P}(\text{Majority decision is correct}) \to 1.$$

Interpretation:

- Under independence and individual "competence" $p > 0.5$, majority voting is a highly reliable aggregator of individual judgements.

A refresher on Condorcet's jury theorem (cont'd)

Key intuition:

- Each voter has a better-than-random chance of voting correctly ($p > 0.5$).
- Errors cancel out because voters vote independently.
- The "signal" (correct votes) dominates the "noise" (errors) as the group size grows.

Why majority rule "works":

- "Law of large numbers": With more voters, the proportion voting correctly converges to p.
- Since $p > 0.5$, the correct option increasingly wins the majority.

$\rightarrow$ See Goodin and Spiekermann, 2018 for various extensions and extensive discussion of the theorem's relevance for democratic theory

Conclusions by Grofman and Feld, 1988

Fragility of Rousseau's argument:

- Factions undermine independence by introducing correlated errors.
- Self-interest undermines sincerity, shifting citizens away from voting for the common good.

Implication:

- Rousseau's confidence in collective decision-making depends critically on citizens' sincere, independent pursuit of the common good.

Implication for scale:

- Scale and heterogeneity can undermine the conditions (independence, sincerity) needed for reliable aggregation.

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

The difficulty of collective choice with heterogenous preferences

Problem: How to aggregate individual preferences into a collective decision?

Warning sign:

→ Condorcet's paradox (1785):

Even with rational individuals, collective preferences can be intransitive.

Key move:

→ Arrow, 1951 generalises Condorcet's insight to all voting rules.

Next: Understand Condorcet's paradox.

Condorcet's paradox

Condorcet's insight (1785):

$\rightarrow$ Majority rule with pairwise comparisons can yield intransitive collective preferences.

Setup:

- Three individuals: A, B, C
- Three alternatives: x_1, x_2, x_3
- Each individual has transitive preferences.

Observation:

$\rightarrow$ Majority preference between alternatives can cycle, even if individuals are rational.

Next: See a concrete example.

Condorcet cycle: An example

Three individuals and their preferences:

Voter	1st choice	2nd choice	3rd choice
A	x_1	x_2	x_3
B	x_2	x_3	x_1
C	x_3	x_1	x_2

Majority preferences in pairwise comparisons:

- $x_1 \succ x_2$ (A and C prefer x_1 over x_2)
- $x_2 \succ x_3$ (A and B prefer x_2 over x_3)
- $x_3 \succ x_1$ (B and C prefer x_3 over x_1)

Conclusion: Cycle arises: $x_1 \succ x_2 \succ x_3 \succ x_1$.

Primitives of collective choice: Individuals and preferences

Individuals:

- $N = \{1, 2, \ldots, n\}$ is the finite, non-empty set of individuals.
- Cardinality: $\#N \geq 2$.

Alternatives:

- X denotes the finite, non-empty set of alternatives.
- Cardinality: $\#X \geq 3$.

Individual preferences:

- Each individual $i \in N$ has a preference relation R_i over X.
- $R_i \in R(X)$, where $R(X)$ is the set of weak orders (complete, reflexive, transitive relations).

Primitives of collective choice: Profiles and Aggregation Rules

Preference profiles:

- $R^N(X)$ is the set of all n-tuples of individual preferences.
- An element $\langle R_i \rangle_{i \in N}$ is called a preference profile.

Preference aggregation rule (PAR):

- A PAR is a mapping: $f : R^N(X) \to B(X)$.
- $B(X)$ = set of complete and reflexive (but not necessarily transitive) binary relations over X.

Goal:

$\to$ Use a PAR to aggregate individual preferences into a collective social relation.

Arrow's *minimal* conditions: Setting the stage

Goal: Formalise minimal, "reasonable" conditions for collective decision-making.

Conditions Arrow imposes on preference aggregation rules (PARs):

- Ensure collective preferences are coherent and respect individual inputs.
- Prevent trivial or dictatorial outcomes.

Warning: Even these seemingly innocuous conditions have profound consequences.

Next: Let's look at the conditions in some more detail.

Arrow's conditions: Overview

Condition	Meaning	Examples ruled out
Unanimity	If everyone in society prefers x to y, then society must prefer x to y.	Any non-Paretian SWF
Universal domain	The SWF must be able to generate a social preference for all logically possible rational profiles.	Any domain-restricted SWF (e.g., weak Pareto)
Transitivity	If society prefers x to y and y to z, it must prefer x to z.	Majority rule
Independence of Irrelevant Alternatives (IIA)	Social preference between x and y should depend only on individuals' preferences between x and y.	Borda count
Weak Pareto	If everyone strictly prefers x to y, society must strictly prefer x to y.	Inverse dictatorship + null aggregation (Wilson, 1972)
Non-dictatorship	No single individual's preferences always dictate social preferences.	Dictatorial aggregation

Arrow's Impossibility Theorem

Theorem (Arrow, 1951)

For a finite set of individuals ($\#N \geq 2$) and three or more alternatives ($\#X \geq 3$), any preference aggregation rule satisfying:

- *Unrestricted Domain,*
- *Weak Pareto,*
- *Independence of Irrelevant Alternatives (IIA),*
- *and Transitivity,*

must be dictatorial.

→ See Sen, 2017 or Gaertner, 2009 for proofs.

Profound implication:

→ Extremely difficult to aggregate individual preferences consistently.

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Riker's interpretation of Arrow

Riker, 1982: *Liberalism Against Populism*

- Arrow's theorem shows that populist democracy—democracy as pure preference aggregation—is arbitrary.
- Any non-dictatorial preference aggregation rule (PAR) satisfying weak Pareto and IIA must be intransitive.
- Thus, social preferences can change arbitrarily even when individual preferences stay the same.
- Preference aggregation alone cannot (normatively) ground democratic legitimacy Weale, 1984.

Problems with Riker's argument

- Arrow's theorem shows only the possibility of preference cycles—not that they necessarily occur in practice (Dowding, 2006). We also don't know how likely they are to arise.
- Riker assumes that electoral accountability guarantees less arbitrariness, but this is not demonstrated.
- Electoral mechanisms require demanding conditions to reliably improve outcomes (Ferejohn, 1986)

Start of summary of Ferejohn

Deliberative democracy: A response to Riker?

Main Idea:

- Arrow's theorem hinges crucially on the unrestricted domain — that is, any pattern of individual preferences must be admissible.
- Deliberation can induce *single-peaked preferences* (Dryzek and List, 2003).
  - → It helps people identify the unique option that is best for them (or establish that they are indifferent between all options).
  - → Rests on the idea that non-single-peaked preferences are due to a lack of information that can be remedied via deliberation
- **Alternative approaches:** Other theorists (e.g., Maskin, 2020, 2025) propose relaxing Arrow's IIA condition instead of restricting domains.

What is single-peakedness?

Definition:

- A voter's preference is single-peaked if alternatives can be ordered along a line (e.g., left–right), such that:
  - → Each voter has one most-preferred alternative p.
  - → Moving away from p in either direction makes outcomes worse.

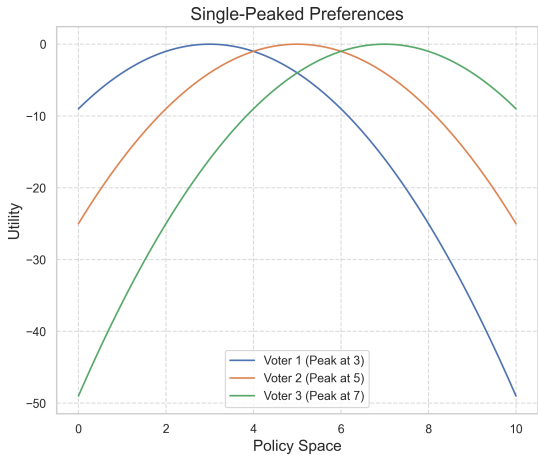
Formally:

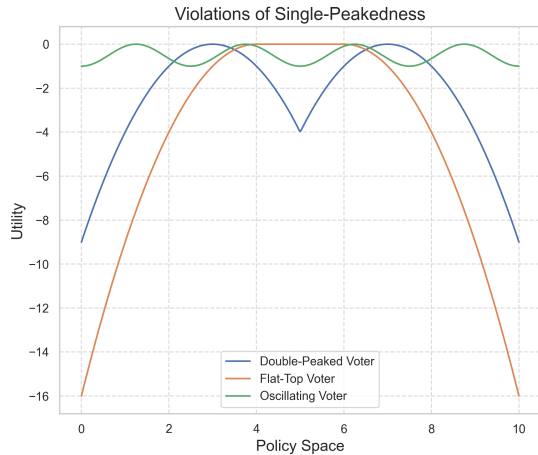
- For all alternatives p'', p', p with $p'' \leq p' \leq p$ or $p \leq p' \leq p''$, we have:

$$U(p'') \leq U(p') \quad \text{and} \quad U(p') \leq U(p)$$

where $U(\cdot)$ denotes the voter's utility.

Illustrating (non-)single-peaked preferences





Back to Fearon

Why single-peakedness matters: Black's theorem(s)

Black's Possibility Theorem (1958):

- If all voters have single-peaked preferences over a one-dimensional policy space,
- And the number of voters is odd,
- Then pairwise majority rule produces a transitive social preference ordering.

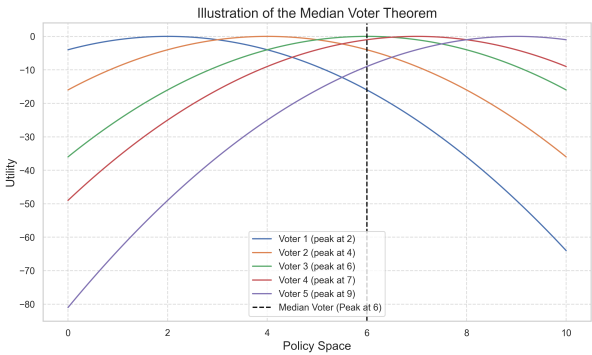
Black's Median Voter Theorem (refinement):

- The median voter's ideal point is a Condorcet winner:
- It defeats every other alternative in pairwise majority votes.

Implications for democracy:

- No cycles (no Condorcet paradox),
- Stable and non-arbitrary majority rule outcomes.

Black's Median Voter Theorem: Visual explanation



- Voters' ideal points are ordered: $2, 4, 6, 7, 9$.
- Median voter = voter at 6.
- Pairwise comparisons:
  - $\rightarrow$ Against 4: Majority (6, 7, 9) prefer 6.
  - $\rightarrow$ Against 7: Majority (2, 4, 6) prefer 6.
  - $\rightarrow$ Against 2 or 9: Majority still prefer 6.
- Thus, the median wins all pairwise majority votes.

The deliberative response: Elegant, but unrealistic?

Main Idea:

- While deliberative democracy offers an elegant response to Arrow and Riker, it faces two major challenges:
  1. **Theoretical challenges**: Strategic behaviour, agenda-setting, and institutional fragility.
  2. **Empirical challenges**: Evidence of group polarisation, biased updating, and failures of deliberation in practice.

Next Steps:

- We first examine theoretical concerns, and then turn to the empirical ones.

Theoretical challenge: Strategic incentives in deliberation

Key insights from Landa and Meirowitz, 2009:

- Deliberation is often strategic, not purely truth-seeking.
- Participants may mislead, withhold information, or selectively disclose to influence outcomes.

Why?

- Non-common values: Participants often pursue different, conflicting goals.
- Cheap talk: Communication is often unverifiable and costless, encouraging manipulation.
- Costly verification: Even when evidence exists, verifying it is expensive and may not happen.

Conclusion: Without careful institutional design, deliberation may **amplify strategic manipulation** rather than promote consensus.

Theoretical challenge: Instability and path dependence

Key insight from Chung and Duggan, 2020:

- Different modes of deliberation (discussion formats) produce very different outcomes.
- Only under strong and idealised conditions does deliberation converge predictably.

Main findings:

- **Myopic discussion**:
  → Unstructured; easily leads to cycling and indeterminate outcomes.
- **Constructive discussion**:
  → Guarantees convergence, but outcomes are path-dependent.
- **Debate (strategic deliberation)**:
  → Can yield a unique compromise, but only under strong strategic rationality assumptions.

Conclusion: Deliberation does not automatically stabilise outcomes.

Does democratic deliberation “work”? Two perspectives

Pro-Deliberation (Optimistic)

- Structured deliberation improves political knowledge, reduces polarisation, and can also durably boost participation (Fishkin, 2018; Fishkin, Bolotnyy, et al., 2024, 2025; Fishkin, Siu, et al., 2021).
- Deliberative forums can improve the “quality“ of reasoning (Niemeyer et al., 2024; Veri, 2025)

Contra-Deliberation (Sceptical)

- Deliberation can increase group polarisation (Sunstein, 2000, 2002).
- Citizens prefer leaders to “get things done” without debate (Hibbing and Theiss-Morse, 2002).
- Deliberation vulnerable to framing and manipulation (Druckman, 2004).

Conclusion: Deliberation can work — but only under **carefully designed, small-scale** conditions. Scaling it up remains difficult.

Deliberation and the challenge of democratic legitimacy

Key takeaways:

- Deliberation offers a possible way to circumvent Arrow's impossibility by restricting the domain of preferences.
- However, theoretical and empirical evidence shows that deliberation is fragile.
- Without ideal conditions, deliberation is often fails to stabilise preferences or mitigate arbitrariness.

Summary:

- Deliberation **can** enrich democracy — but it cannot "solve" the problems Arrow and Riker highlight.
- Robust democracy requires **institutions that manage conflict and strategic behaviour** $\rightarrow$ problem of scale remains unresolved

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Geometric sums (a quick refresher)

Why we need them: when a constant payoff x is discounted by $\delta \in (0,1)$ over an infinite horizon, the present value is

$$V \;=\; \sum_{t=0}^{\infty} \delta^t x \;=\; x + \delta x + \delta^2 x + \delta^3 x + \dots$$

The classic "trick" – multiply by δ and subtract:

$$V = x + \delta x + \delta^2 x + \dots \qquad \delta V = \qquad \delta x + \delta^2 x + \dots$$

Subtracting the second from the first leaves only x on the right:

$$V - \delta V = x \implies V = \frac{x}{1-\delta}.$$

Why $\delta < 1$ matters: if $\delta \geq 1$, the sum diverges, and "waiting" has unbounded value.

◂ Back to derivation

From $V_E \geq V_F$ to the compliance condition

Start from the inequality $V_E \geq V_F$: $\frac{L+\pi S}{1-\delta} \;\geq\; \frac{L+pS}{1-\delta} - C$.

Step 1 – multiply both sides by $(1-\delta) > 0$ (preserves the inequality):

$$L + \pi S \;\geq\; L + pS - (1-\delta)C.$$

Step 2 – cancel L on both sides: $\pi S \;\geq\; pS - (1-\delta)C$.

Step 3 – define $\kappa \equiv (1-\delta)C$, the per-period equivalent of the fighting cost:

$$\pi S \;\geq\; pS - \kappa\ .$$

Equivalent rearrangement: $\kappa \;\geq\; (p-\pi)S$.

$\Rightarrow$ Fighting costs must outweigh the extra chance of office that force buys you.

◂ Back to derivation

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Probability distributions, CDFs, densities

The shock θ is random. Its behaviour is summarised by a cumulative distribution function (CDF) F: $F(x) = \Pr(\theta \leq x)$. **What F looks like.** $F(0) = 0$, $F(m) = 1$, and F is non-decreasing – pile up more space, accumulate more probability.

The density f. If F is differentiable, $f(x) = F'(x)$: the "rate" at which probability accumulates at x. Two examples:

- Uniform on $[0, 1]$: $F(x) = x$, $f(x) = 1$. Probability is spread evenly.
- Standard normal: $f(x)$ is the bell curve; $F(x)$ is its running area.

Why both objects matter.

- For probabilities (e.g. "incumbent clears the cutoff"): $\Pr(\theta \geq \theta^*) = 1 - F(\theta^*)$.
- For averages (expected utility): $\mathbb{E}[g(\theta)] = \int_0^m g(x)f(x)\,dx$.

▸ Continuously differentiable F ◂ Back to model

What does it mean for F to be continuously differentiable?

Restating the assumption. $F : [0, m] \to [0, 1]$ is the cdf of the shock θ. "Continuously differentiable" means F has a derivative $f = F'$ at every point, and f is itself continuous (no jumps in the density).

What would it mean if this failed?

- F jumps (a "mass point"). A whole chunk of probability is concentrated at one value of θ. Many states of the world look identical to the incumbent.
- F is flat over an interval. Some range of θ is impossible.

Why does the model need it?

- The threshold θ^* is defined by $\phi(K/\theta^*) = \Delta$. For this to pin down a unique value, F must be smooth.
- Voter expected utility is computed by integrating f. Jumps in F produce discontinuous jumps in welfare – comparative statics break.

Examples. Satisfies: Uniform, Normal, Beta, Exponential. Violates: Bernoulli, any distribution with a point mass.

◂ Back to model primitives

What is the expectation operator $\mathbb{E}[\cdot]$?

Definition. The expectation of a random variable is its probability-weighted average.

Discrete example. A coin flip pays $£1$ on heads, $£0$ on tails. Then

$$\mathbb{E}[\text{payoff}] = 0.5 \cdot 1 + 0.5 \cdot 0 = 0.5.$$

Continuous example. If θ is Uniform$[0, 1]$, $\mathbb{E}[\theta] = \int_0^1 x\, dx = 1/2$.

Why agents maximise expected utility.

- The standard rationality assumption: form beliefs about probabilities, average outcomes accordingly.
- Implies risk-neutrality when utility is linear in payoffs (as Ferejohn assumes).
- Risk-aversion would use a concave utility function – a separate complication we set aside.

In our model. The voter's per-period expected utility is $\mathbb{E}[u]$ – average performance, weighted by how often the incumbent clears versus shirks.

◂ Back to worked case

Convex cost functions

What $\phi' > 0$ and $\phi'' > 0$ mean.

- $\phi' > 0$: the cost is strictly increasing. More effort always costs more.
- $\phi'' > 0$: the cost rises at an increasing rate. Each extra unit of effort costs more than the previous one.

Analogy. Going from 4 to 8 hours of work is hard; going from 8 to 12 is harder. The marginal cost of effort accelerates.

Examples.

- $\phi(a) = a^2$: $\phi' = 2a$, $\phi'' = 2$. Strictly convex. Standard textbook choice. $\phi(a) = a$: $\phi'' = 0$. Boundary case ("weakly convex"). Used in our worked example for tractability.
- $\phi(a) = \log(1 + a)$: concave ($\phi'' < 0$). Not what we want.

Why does Ferejohn need convexity?

- Without it, the incumbent's optimum is at a corner, either $a = 0$ or $a = \infty$.
- Convexity gives a well-defined trade-off between effort and cost.

◂ Back to model primitives

The discount factor δ

Definition. $\delta \in (0, 1)$ is the per-period discount factor: how much a payoff next period is worth today. If $\delta = 0.9$, then $£1$ next year is worth $£0.9$ now. The "exchange rate" between the present and the future.

Why $\delta < 1$.

- Preference for the present: most people would rather have a reward today than tomorrow.
- Survival uncertainty: the future may not happen.
- Formal: the infinite sum $\sum_{t=0}^{\infty} \delta^t x = x/(1-\delta)$ converges only when $\delta < 1$.

Why $\delta > 0$.

- If $\delta = 0$, only today's payoff matters – the model collapses to a one-shot game.
- One-shot games here yield $a = 0$ (no incentive to work). The infinite horizon is what gives elections any bite.

▸ Geometric sums ◂ Back to dynamic structure

Value functions and the Bellman equation

Definition. V^I is the lifetime expected utility of being the incumbent right now: today's payoff, plus discounted future payoffs, summed forever.

The Bellman trick. Decompose lifetime value as "today + tomorrow's lifetime value." For our incumbent, in the favourable-shock branch:

$$\underbrace{V^I}_{\text{lifetime now}} = \underbrace{W - \phi(K/\theta)}_{\text{today's payoff}} + \underbrace{\delta V^I}_{\text{tomorrow's lifetime, discounted}} .$$

- Stationarity: tomorrow's continuation looks the same as today's, so we can use the same symbol V^I on both sides.
- One equation in one unknown: solvable.

Two illuminating special cases.

- Set $\phi(K/\theta) = 0$: then $V^I = W/(1-\delta)$, i.e. the present value of a constant stream W, exactly the geometric-sum formula.
- Set $\delta = 0$: then $V^I = W - \phi(K/\theta)$, i.e. a one-shot payoff, no future.

Solving the Bellman system: V^I, V^O, and the wedge Δ

Two equations, two unknowns (stationary play). Let $p \equiv 1 - F(\theta^*)$ be the per-period reelection probability and $\bar{\pi} \equiv \mathbb{E}[\text{per-period payoff} \mid \text{in office}]$.

$$V^I = \bar{\pi} + \delta\big[p\,V^I + (1-p)V^O\big], \qquad V^O = \delta\big[\lambda V^I + (1-\lambda)V^O\big].$$

Step 1 – solve V^O in terms of V^I: $\displaystyle V^O = \frac{\delta\lambda}{1-\delta(1-\lambda)}\,V^I.$

Step 2 – substitute and collect: $\displaystyle V^I = \frac{\bar{\pi}\,[1-\delta(1-\lambda)]}{(1-\delta)\,[1-\delta(p-\lambda)]}.$

Step 3 – form the wedge. Using $V^I - V^O = (1-\delta)V^I/[1-\delta(1-\lambda)]$:

$$\boxed{\Delta \;\equiv\; \delta\big(V^I - V^O\big) \;=\; \frac{\delta\,\bar{\pi}}{1-\delta(p-\lambda)}.}$$

Step 4 – read off the comparative statics.

- $\bar{\pi}$ is linear in W (slope 1 – the rent W is collected each period in office). So $\partial\Delta/\partial W > 0$.
- $\partial[1-\delta(p-\lambda)]^{-1}/\partial\lambda = -\delta/[1-\delta(p-\lambda)]^2 < 0$. So $\partial\Delta/\partial\lambda < 0$.

Worked-case sanity check. Uniform-linear gives $\theta^* = p = 1/2$, $\bar{\pi} = W - (\Delta/2)\ln 2$, and voter welfare $\mathbb{E}[u] = \Delta/4$ – literally proportional to the wedge. Both signs above translate one-for-one into voter welfare.

Equilibrium concepts in three steps

1. Best response. Given what others do, your move that maximises your payoff. "If the voter sets K, my best response as incumbent is the threshold-and-shirk strategy."

2. Nash equilibrium. Every player is playing a best response to the others. Nobody wants to deviate unilaterally. The oter's cutoff K and incumbent's strategy $a(\theta)$ are mutual best responses.

3. Subgame-perfect equilibrium (SPE). Strategies must be best responses at every point in the game, not just at the start. Rules out non-credible threats.

Credibility.

- Voter says: "If $u < K$, I'll vote against you." For this to be SPE, the voter must actually want to vote against the incumbent when the moment comes.
- Since challengers are identical to incumbents (no types!), the voter is indifferent – firing is just as good as keeping. The threat is credible.

Stationarity. We restrict attention to strategies that don't depend on time. A simplification, but a natural one when the game looks the same every period.

Back to dynamic structure

Why is the incumbent's strategy a *threshold*?

Why only two options? The voter's rule is a sharp cutoff: $u \geq K \Rightarrow$ re-elected; $u < K \Rightarrow$ removed. From the incumbent's view:

- Effort $< K/\theta$: performance falls short, removed anyway. All effort wasted.
- Effort exactly K/θ: performance hits K, re-elected. Minimum needed.
- Effort $> K/\theta$: re-elected, but extra effort doesn't help. Wasteful.

So the only options for optimal action are:

1. $a = K/\theta$: clear with minimum effort, re-elected.
2. $a = 0$: save the cost, accept defeat.

Choose between them by comparing $\phi(K/\theta)$ (cost of clearing) to Δ (benefit of staying in office). That comparison gives the threshold θ^*.

What if the voter's rule were *soft* (e.g., re-elect with probability increasing in u)? Then the incumbent would have smooth interior optima. The threshold result is a feature of the cutoff form.

◂ Back to threshold-and-shirk

Setting up the worked voter problem

We want to derive $K = \Delta/2$. Recall $F = \text{Uniform}[0,1]$, $\phi(a) = a$, wedge $\Delta = \delta(V^I - V^O)$.

Step 1 – the threshold. The incumbent works iff $\phi(K/\theta) \le \Delta$. With $\phi(a) = a$, this is $K/\theta \le \Delta$, i.e. $\theta \ge K/\Delta$. So $\theta^* = K/\Delta$.

Step 2 – what the voter actually gets each period.

- If $\theta \ge \theta^*$: the incumbent picks $a = K/\theta$, so $u = a\theta = K$. Constant in θ – the magic of multiplicative noise.
- If $\theta < \theta^*$: the incumbent shirks, so $a = 0$, $u = 0$.

Step 3 – expected per-period performance.

$$\mathbb{E}[u] \;=\; K \cdot \underbrace{(1-\theta^*)}_{\Pr(\text{cleared})} + 0 \cdot \underbrace{\theta^*}_{\Pr(\text{shirked})} \;=\; K\big(1 - \tfrac{K}{\Delta}\big).$$

The voter takes Δ as given when choosing K: the wedge is determined by the rest of the dynamic system, not by a single deviation.

◂ Back to worked case ▸ Solve the FOC

Solving for $K^* = \Delta/2$

Step 4 – the voter's problem is a quadratic.

$$\max_{K \geq 0} \quad K\left(1 - \tfrac{K}{\Delta}\right) \;=\; K - \tfrac{K^2}{\Delta}.$$

First-order condition. $\frac{d}{dK}\left[K - \frac{K^2}{\Delta}\right] = 1 - \frac{2K}{\Delta} = 0 \quad \Longrightarrow \quad \boxed{K^* = \frac{\Delta}{2}}$.

Second-order condition. $\frac{d^2}{dK^2}[K - K^2/\Delta] = -2/\Delta < 0$, so the FOC is a maximum.

Step 5 – the implied equilibrium objects.

- Threshold: $\theta^* = K^*/\Delta = \frac{1}{2}$.
- Re-election probability: $1 - \theta^* = \frac{1}{2}$.
- Voter's expected utility: $K^*(1 - K^*/\Delta) = \frac{\Delta}{4}$.

Sanity check. $\Delta = 0$ (office worthless) $\Rightarrow K^* = 0$ – voter extracts nothing. As Δ grows, both K^* and welfare $\Delta/4$ grow linearly. Formal mechanism behind "higher stakes $\Rightarrow$ more discipline".

◂ Back to worked case

Why uniform shocks and linear cost?

Why these functional forms? They make every step of the algebra trivial:

- Uniform on $[0, 1]$: density $f(x) = 1$, so probabilities and expectations are just lengths and averages.
- Linear cost $\phi(a) = a$: inverse $\phi^{-1}(x) = x$, so the threshold equation $\phi(K/\theta^*) = \Delta$ collapses to $\theta^* = K/\Delta$.

What changes with other forms?

- Quadratic cost $\phi(a) = a^2/2$: $\phi^{-1}(x) = \sqrt{2x}$, so $\theta^* = K/\sqrt{2\Delta}$. Same logic, messier algebra.
- Normal shocks: probabilities involve the inverse normal CDF – no closed form, but qualitative results unchanged.

What does *not* depend on the special case?

- Threshold-and-shirk equilibrium structure.
- $\partial U/\partial W > 0$ (Prop 5): more valuable office $\Rightarrow$ more discipline.
- $\partial U/\partial \lambda < 0$ (Prop 4): two-partyism $\Rightarrow$ less discipline.
- Heterogeneity collapse (Props 6–7).

◂ Back to worked case

Why V^O depends on λ

The lifetime value of being out. A loser today gets nothing this period, but at the next election:

- with probability λ she returns, earning continuation V^I;
- with probability $1-\lambda$ she stays out, earning continuation V^O.

$V^O \approx \delta\big[\lambda V^I + (1-\lambda)V^O\big]$. At $\lambda = 1$: $V^O = \delta V^I$ (close to V^I). At $\lambda = 0$: V^O is small.

Why this kills voter welfare.

1. $\lambda \uparrow \Rightarrow V^O \uparrow \Rightarrow$ wedge $\Delta = \delta(V^I - V^O) \downarrow$.
2. Smaller $\Delta \Rightarrow$ the incumbent only works under very favourable shocks $\Rightarrow$ more shirking.
3. In the worked case $\mathbb{E}[u] = \Delta/4$, so welfare is literally proportional to the wedge.

What would it mean for λ to be large? Two-party alternation, losing is shallow. Sanctioning threat is muted.
What would it mean for λ to be small? Many small parties, losing is permanent. Threat bites hard, incumbents work harder.

◂ Back to multipartism slide

What does "$\partial U/\partial W > 0$" actually mean?

Comparative static. How does an equilibrium quantity change when we incrementally change one parameter, holding the others fixed?

Plain English. "If we made the office a bit more valuable (higher W), would voters be better off in the new equilibrium?" $\partial U/\partial W > 0$ means: yes, voter welfare rises.

Why this is hard to compute by hand. W appears in many places at once:

- Directly, in the incumbent's per-period payoff $W - \phi(a)$.
- Indirectly, in V^I (which depends on future W), then Δ, then optimal K, then the voter's welfare U.

The partial derivative traces all those channels.

Why we care. Comparative statics are the model's empirical predictions. "When stakes rise, control rises" is something we can take to data on, e.g., presidential vs. parliamentary systems, or executives with vs. without two-term limits.

Back to Prop 5

Summary: exogenous vs. endogenous variables

Exogenous variables are inputs to the model – given from outside, chosen by the modeller, not changed by the players.

Endogenous variables are outputs – determined by the equilibrium.

Exogenous (parameters)	Endogenous (equilibrium objects)
W – value of office	K – voter's cutoff
δ – discount factor	θ^* – shirking threshold
F – distribution of shocks	V^I, V^O – lifetime values
ϕ – cost function	$a^*(\theta)$ – incumbent's effort
λ – prob. of return	U – voter welfare

- Don't confuse: K depending on Δ does not make K "set from outside." Both are endogenous, jointly determined.

Back to model primitives

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Bayes' rule

The central question. You have a prior belief α about a hypothesis H ("the incumbent is good"). You observe data z. What should your updated belief be?

Bayes' rule: $\underbrace{\Pr(H \mid z)}_{\text{posterior}} = \underbrace{\Pr(z \mid H)}_{\text{likelihood}} \cdot \underbrace{\Pr(H)}_{\text{prior}} \Big/ \underbrace{\Pr(z)}_{\text{normaliser}}$.

Interpretation. The posterior is *prior* (what you believed before) times *likelihood* (how compatible the data is with the hypothesis), normalised so probabilities sum to 1.

In Fearon's selection model.

- Prior: $\Pr(\text{good}) = \alpha$.
- Likelihood (good): $f(z)$ (centred at 0, since good's policy is $x_g = 0$).
- Likelihood (bad): $f(\hat{x}^2 + z)$ (shifted by bad's expected welfare loss).
- Posterior: $\alpha'(z) = \alpha f(z) / [\alpha f(z) + (1-\alpha) f(\hat{x}^2 + z)]$.

Why we need it. Without Bayes, we couldn't formalise what voters *learn* from observing performance.

Bayes' rule: the intuitive version

The basic idea. New evidence shifts your belief in proportion to how much more likely the evidence is under one hypothesis than another. Suppose, we observe an incumbent deliver above-average welfare. Should you raise or lower your belief that she is a "good type"? By how much?

Three ingredients.

- Prior: 30% of politicians are good types.
- Likelihood given good: good types deliver above-average welfare 80% of the time.
- Likelihood given bad: bad types deliver above-average welfare only 20% of the time.

Imagine 100 politicians.

- 30 are good. Of these, $30 \times 0.8 = \mathbf{24}$ deliver above-average welfare.
- 70 are bad. Of these, $70 \times 0.2 = \mathbf{14}$ also deliver above-average welfare.
- Among the $24 + 14 = 38$ politicians who delivered above-average welfare, the fraction who are good is $24/38 \approx \mathbf{0.63}$.

The update: prior 30% ⇒ posterior 63%. Two ingredients drove it – the prior (30%) and the likelihood ratio $0.8/0.2 = 4$ (good outcomes are four times more likely under good types).

The formula. $\Pr(\text{good} \mid \text{outcome}) = \dfrac{\Pr(\text{outcome} \mid \text{good}) \cdot \Pr(\text{good})}{\Pr(\text{outcome})}$.

▸ Formal version ▸ Odds-ratio Bayes

Likelihood ratios and Bayes' rule in odds form

Bayes' rule in odds form.

$$\underbrace{\frac{\Pr(H \mid z)}{\Pr(\neg H \mid z)}}_{\text{posterior odds}} = \underbrace{\frac{\Pr(z \mid H)}{\Pr(z \mid \neg H)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{\Pr(H)}{\Pr(\neg H)}}_{\text{prior odds}}.$$

Why odds. The normaliser cancels. Updates compose multiplicatively, which is cleaner.

Fearon's selection cutoff. Reelect iff posterior odds $\geq$ prior odds, i.e., iff likelihood ratio ≥ 1:

$$\frac{f(z)}{f(\hat{x}^2 + z)} \;\geq\; 1.$$

For symmetric, unimodal f centred at zero, this holds iff $z \geq -\hat{x}^2/2$, giving the midpoint rule $k^* = -\hat{x}^2/2$.

Intuition. As z rises, the good-type density rises and the bad-type density falls (since their mean is lower). Likelihood ratio is monotone in z. The cutoff is well-defined.

Back to Bayes

Backward induction in two-period games

The problem. A two-period game has decisions made in sequence. Period-1 choices affect what happens in period 2, but rational players in period 1 must *anticipate* period-2 play.

The solution. Solve from the end:

1. Compute optimal period-2 actions *for each possible state* entering period 2.
2. Plug those into period-1 payoffs.
3. Solve the resulting one-period problem in period 1.

In Fearon's model.

- Period 2 is the last period: bad types pick $y = 1$, good types pick $y = 0$. No incentive to please voters.
- Period 1: knowing this, the bad type's reelection means $y = 1$ tomorrow, voter gets $-1 + \varepsilon'$. Voter's incentive to keep the incumbent depends entirely on *type*, not on incumbent's promised future behaviour.

The trade-off this exposes. Voter would *like* to commit to a sanctioning rule today; but tomorrow, with no future, she has nothing to bargain with. Hence the commitment problem.

◂ Back to model setup

Perfect Bayesian Equilibrium (PBE)

Why we need a new concept. Subgame-perfect equilibrium (SPE) handles dynamic games well, but doesn't say what to do when players have private information. Fearon's voter doesn't observe the incumbent's type, so SPE isn't enough.

PBE has two ingredients.

1. Strategies: what each player does at each information set (where they have to choose).
2. Beliefs: probabilities each player assigns to the unobserved state.

Two consistency requirements.

- Sequential rationality: at every information set, each player's strategy is optimal given her beliefs.
- Bayesian updating: beliefs are updated via Bayes' rule whenever possible (i.e., on the equilibrium path).

In Fearon. Voter's strategy is the cut rule k; voter's beliefs are the posterior $\alpha'(z)$. The PBE condition pins down both jointly: k must be optimal given $\alpha'(z)$, and $\alpha'(z)$ must come from Bayes' rule given the equilibrium policy choices.

Three assumptions on the noise distribution f

The voter observes $z = -x^2 + \varepsilon$, where $\varepsilon \sim f$. Fearon imposes three conditions on f:

- Symmetry: $f(-\varepsilon) = f(\varepsilon)$, i.e. equally likely to over- or under-shoot the mean.
- Unimodality: f has a single peak; density falls monotonically as $|\varepsilon|$ grows away from the mode.
- Mean-zero: $\mathbb{E}[\varepsilon] = 0$, i.e. the shock has no systematic bias.

Familiar examples that satisfy all three: standard normal $\mathcal{N}(0, \sigma^2)$, Laplace, logistic, Cauchy, uniform on a symmetric interval $[-c, c]$.

Why these three together. Each plays a distinct role in the equilibrium derivation:

- Symmetry $\Rightarrow$ the optimal cutoff sits at the *midpoint* between expected outcomes of the two types.
- Unimodality $\Rightarrow$ the set of signals favouring "good" is a half-line, so a single cutoff suffices.
- Mean-zero $\Rightarrow \mathbb{E}[z \mid x] = -x^2$; welfare reads off policy without a constant bias term.

Stripped of any one of these, the cut-rule equilibrium loses its clean midpoint form – and may not exist as a single-cutoff rule at all.

▸ Symmetry ▸ Unimodality ▸ Mean-zero ◂ Back

Symmetry: what it does and what fails without it

Definition. $f(-\varepsilon) = f(\varepsilon)$ for all ε. The density looks the same on either side of its centre.

Where it bites in the derivation. Reelect iff posterior odds favour "good", which (from Bayes' rule with mean-zero) reduces to:

$$\frac{f(z)}{f(\hat{x}^2 + z)} \;\geq\; 1 \iff f(z) \geq f(\hat{x}^2 + z).$$

For a symmetric, unimodal f, this holds iff $|z| \leq |\hat{x}^2 + z|$, and squaring/simplifying gives the unique cutoff $z = -\hat{x}^2/2$. The cutoff is exactly halfway between the two type means.

What a violation looks like. Suppose f is right-skewed (long upper tail):

- Big positive shocks are common; big negative shocks are rare.
- The level-set condition $f(z) = f(\hat{x}^2 + z)$ no longer occurs at the midpoint.
- The optimal cutoff shifts toward the heavy tail – voters become more demanding when surprises are usually upward (good outcomes get partially discounted as luck).

Why this matters. Skewness breaks the clean "midpoint between types" interpretation. Without symmetry, you cannot read $k^* = -\hat{x}^2/2$ directly off the type pool – you need the entire shape of f to locate the cutoff, and comparative statics ($\partial k^*/\partial \hat{x}$ etc.) cease to be transparent.

Unimodality: what it does and what fails without it

Definition. f has a single peak: there exists $\varepsilon^\dagger$ such that f is non-decreasing on $(-\infty, \varepsilon^\dagger]$ and non-increasing on $[\varepsilon^\dagger, \infty)$. With mean-zero and symmetry, the peak sits at $\varepsilon^\dagger = 0$.

Where it bites. Together with symmetry, unimodality delivers the half-line property:

$$\{z : f(z) \geq f(\hat{x}^2 + z)\} \;=\; \{z \geq k\} \quad \text{for some } k.$$

A higher z is always (weakly) stronger evidence for a good type, so a single cutoff $z \geq k$ is optimal – the workhorse of the entire model.

What a bimodal violation looks like (e.g., a mixture of two normals far apart):

- The set above is *no longer* a half-line – it can take the form "moderate z only" or "extreme z only".
- Optimal voting is *not* a simple cutoff: voters might want to reelect on moderate z but not on extreme z, or vice versa.
- The link between the cutoff and the bad type's FOC – and hence the entire commitment-problem story – collapses; equilibrium requires multiple cutoffs (or non-cutoff rules).

Why this matters. Unimodality is what makes the model tractable. Without it, comparative statics on k lose interpretability, and the connection between "monitoring quality" and $f(0)$ breaks.

Mean-zero: what it does and what fails without it

Definition. $\mathbb{E}[\varepsilon] = \int \varepsilon \, f(\varepsilon) \, d\varepsilon = 0$. The shock has no systematic upward or downward bias.

Where it bites. The signal is $z = -x^2 + \varepsilon$. Taking expectations: $\mathbb{E}[z \mid x] = -x^2 + \mathbb{E}[\varepsilon] = -x^2$.
The signal is unbiased about the policy. This is what lets the voter compare types via a single number $\hat{x}^2$, i.e. the gap between expected welfare under the good and bad types.

What a violation looks like. Suppose $\mathbb{E}[\varepsilon] = \mu \neq 0$ – e.g., a structural tailwind ($\mu > 0$, favourable economic conditions all incumbents share) or headwind ($\mu < 0$):

- $\mathbb{E}[z \mid x] = -x^2 + \mu$ – welfare reflects policy plus a constant.
- Posterior over types becomes biased: a politician benefiting from favourable conditions "looks more good" than she really is.
- The midpoint cutoff moves to $k^* = -\hat{x}^2/2 + \mu$. Ignoring μ makes voters systematically too lenient ($\mu > 0$) or too harsh ($\mu < 0$).

Why this matters. Mean-zero is an identification assumption: without it, the voter cannot back out x from z without already knowing μ. It pins down what the data "means".

Practical fix. If μ is observable (e.g., a common economic tide), voters can demean: use $\tilde{z} = z - \mu$ as the signal. This is the formal logic behind benchmarking against neighbouring jurisdictions (Besley, 2006).

Deriving the pure-selection PBE

Setup. Two types with *fixed* (non-strategic) policy: good plays $x_g = 0$ (welfare centred at $z = 0$); bad plays $x_b = \hat{x}$ (welfare centred at $-\hat{x}^2$). Prior $\Pr(\text{good}) = \alpha$. Voter chooses cutoff k; reelect iff $z \geq k$.

Step 1 – voter's posterior via Bayes' rule.

$$\alpha'(z) \;=\; \frac{\alpha\, f(z)}{\alpha\, f(z) + (1-\alpha)\, f(\hat{x}^2 + z)}.$$

Step 2 – voter's optimal cutoff. A challenger drawn from the pool has expected good-probability α (the prior). Reelect iff $\alpha'(z) \geq \alpha$, equivalent to:

$$\frac{f(z)}{f(\hat{x}^2 + z)} \;\geq\; 1.$$

Step 3 – exploit symmetry + unimodality. Symmetric f gives $f(z) = f(-z)$; unimodal f is decreasing in $|\cdot|$. So $f(z) \geq f(\hat{x}^2 + z)$ iff $|z| \leq |\hat{x}^2 + z|$. Squaring $z^2 \leq (\hat{x}^2 + z)^2 = \hat{x}^4 + 2\hat{x}^2 z + z^2$ and simplifying:

$$\boxed{k^* \;=\; -\hat{x}^2/2.}$$

Step 4 – check PBE conditions.

- Sequential rationality: the cutoff is optimal given the posterior (Step 2).
- Bayesian updating: $\alpha'(z)$ follows from Bayes' rule given the (non-strategic) type-conditional policies (Step 1).

Deriving the pure-sanctioning PBE

Setup. All politicians are bad types with ideal $x = 1$. Period 2 (terminal): incumbent picks $y = 1$ (no future). Reelection rent in period 2: W for the incumbent, 0 otherwise.

Step 1 – incumbent's period-1 problem.

$$\max_{x} \quad \underbrace{W - (1-x)^2}_{\text{period-1 own utility}} + \underbrace{\delta\, W\, \Pr(z \geq k \mid x)}_{\text{discounted reelection rent}} .$$

With $z = -x^2 + \varepsilon$, the reelection probability is $\Pr(\varepsilon \geq x^2 + k) = 1 - F(x^2 + k)$.

Step 2 – FOC w.r.t. x. $2(1-x) \;-\; \delta W \cdot 2x \cdot f(x^2+k) \;=\; 0 \quad\Longleftrightarrow\quad 1 - x \;=\; \delta W\, x\, f(x^2+k)$.

Step 3 – voter's optimal cutoff. The voter wants to maximise the marginal return to incumbent effort, i.e. make $f(x^2+k)$ as large as possible. By unimodality + symmetry, f peaks at 0, so the voter sets k such that $x^2 + k = 0$: $k^* \;=\; -x^2$.

Step 4 – combine. Plug $k^* = -x^2$ into the FOC: $f(x^2 + k^*) = f(0)$. Solving:

$$\boxed{\; x^* \;=\; \frac{1}{1 + \delta W\, f(0)}, \qquad k^* \;=\; -(x^*)^2. \;}$$

Comparative statics. $\partial x^*/\partial W < 0$, $\partial x^*/\partial \delta < 0$, $\partial x^*/\partial f(0) < 0$. Better monitoring (larger $f(0)$, e.g. less noisy environment) and more valuable office both reduce shirking.

Deriving the mixed-case PBE

Setup. Good types (mass α) play $x_g = 0$ – dominant given quadratic loss. Bad types (mass $1-\alpha$) play x_b, to be determined. Voter chooses k, reelects iff $z \geq k$.

Step 1 – bad type's FOC. Same structure as pure sanctioning (only the cutoff changes):

$$2(1-x_b) \;-\; \delta W \cdot 2x_b\, f(x_b^2 + k) \;=\; 0.$$

Step 2 – voter's posterior given equilibrium policies $(0, x_b)$**.** $\alpha'(z) \;=\; \frac{\alpha\, f(z)}{\alpha\, f(z) + (1-\alpha)\, f(x_b^2+z)}$.

Reelect iff $\alpha'(z) \geq \alpha$, equivalent to $f(z) \geq f(x_b^2 + z)$. By symmetry + unimodality (same logic as pure selection): $z \geq -x_b^2/2$. So the voter's ex-post optimal cutoff is $k^*(x_b) = -x_b^2/2$.

Step 3 – joint PBE conditions. The pair (k^*, x_b^*) jointly solves:

$$k^* = -\frac{(x_b^*)^2}{2}, \qquad 1 - x_b^* \;=\; \delta W \cdot x_b^*\, f\left(\frac{(x_b^*)^2}{2}\right).$$

Step 4 – why $x_b^* > x^*_{\text{sanctioning}}$**.** In pure sanctioning, the voter could set $k = -x^2$, placing $f(\cdot)$ in the FOC at $f(0)$ – the *maximum*. In the mixed case, Bayes pulls k^* to $-x_b^2/2$, so the FOC's density is $f(x_b^2/2) < f(0)$. The marginal return to effort drops, so the bad type optimally chooses a larger x_b (more shirking).

The commitment problem in one line. The voter *would like* $k = -x_b^2$ (sanctioning-optimal); she *actually picks* $k = -x_b^2/2$ (selection-optimal). Selection at the polls undermines sanctioning.

Adverse selection vs. moral hazard

Two flavours of asymmetric information in economics. Easy to confuse.

Adverse selection. The hidden quantity is a type. "Hidden information."

- Akerlof's lemons: car sellers know quality, buyers don't.
- Insurance markets: applicants know their risk, insurers don't.
- Fearon's selection model: voters don't know if the incumbent is a good type.

Moral hazard. The hidden quantity is an action. "Hidden action."

- Insurance fraud: insurer can't observe whether the insured drove carefully.
- Principal-agent problems: employer can't observe employee effort.
- Ferejohn's model: voters can't observe the incumbent's effort.
- Fearon's pure-sanctioning case: all bad types, but their policy choice is hidden.

Fearon's mixed case combines both. Hidden type *and* hidden action. The interaction between the two is what generates the commitment problem.

◂ Back to good types

Commitment vs. credibility

These two concepts are closely related but distinct.

Credibility. Will the player actually carry out the threat she announced? Solved by SPE: a threat is credible iff carrying it out is the player's best response when the moment comes.

Commitment. Can the player bind her future self? A commitment problem arises when a player would, ex ante, choose to play differently from what she will, ex post, find optimal.

Two settings:

- Ferejohn (Ch. 1, sanctioning): voter is indifferent between candidates at the polls (all identical). Any cutoff is credible. No commitment problem.
- Fearon (mixed case): voter *strictly prefers* to update on type at the polls. The ex-ante optimal cutoff (motivating bad types) is *not* the ex-post optimal cutoff (selecting types). Voter cannot commit to ignore type info. Commitment problem.

Time inconsistency. A famous example outside elections: monetary policy. The central bank would like to commit to low inflation, but ex post is tempted to inflate to boost output. Same structural problem.

Back to commitment slide

Outline

Sharpening Rousseau's ideal

Condorcet's paradox and Arrow's (im)possibility theorem

What implications, if any, do these results have for democracy?

Self-enforcing democracy

Ferejohn

Fearon

Persuasion and experimental thinking

Information-provision experiments: design and stages

The canonical design. Treatment-group respondents are given one or more factual claims, e.g. the actual immigrant share. Outcome: belief updating or attitude change, relative to a control group.

What to measure at each stage.

- Reception: did participants read and understand the information? *Always include a manipulation or attention check.*
- Persuasion (belief updating): did beliefs about the specific fact actually shift? *Measure this directly.* If beliefs don't move, the design has failed at the persuasion stage and downstream outcomes are uninterpretable.
- Posterior: downstream attitude (magnitude), persistence across waves (durability), spillover to related issues (scope), and behavioural follow-through (turnout, donations, contact with representatives).

What a well-designed paper measures.

1. Pre-treatment beliefs about the fact.
2. Post-treatment beliefs (immediately after).
3. Post-treatment attitudes or preferences.
4. A delayed follow-up wave for durability.
5. Where possible, a behavioural outcome.

Central methodological point: a result is interpretable only when you know *which stage moved*.

Information-provision experiments: interpreting the results

Common mistakes.

- "Null effect on attitudes ⇒ motivated reasoning." Could also be reception failure (the info wasn't processed) or pre-existing counter-cues drowning out the treatment.
- "Heterogeneous by party ⇒ polarisation." Could be source-credibility heterogeneity – who delivered the message?
- "Beliefs update but attitudes don't ⇒ voter irrationality." Attitudes depend on many beliefs; one fact need not be pivotal.
- "Effect decays within weeks ⇒ no real persuasion." Decay is the norm (Coppock, 2023); durability is a separate dimension from magnitude.

A diagnostic checklist for reading such papers:

1. Is reception explicitly measured?
2. Is belief updating measured separately from attitudes?
3. Is durability tested in a follow-up wave?
4. Is a behavioural outcome included?

Central observation. "Voters don't update on information" is a claim about all three stages combined.

Priming experiments: design and stages

The canonical design. Treatment-group respondents are exposed to a stimulus — e.g. a vignette — intended to make a particular consideration more accessible when they answer subsequent questions. No new factual information is provided. Outcome: attitude or behaviour shift on the primed dimension.

Distinct from information provision.

- Priming changes the salience of considerations the respondent already has.
- Information provision changes the beliefs the respondent uses.

What to measure at each stage.

- Reception: did the prime register? Subtle primes often fail at this stage – the participant never noticed. Obvious primes might activate SD bias.
- Persuasion (salience shift): did the prime change which considerations the respondent draws on? Typically inferred from downstream attitudes.
- Posterior: magnitude usually small; durability short (Coppock, 2023 and others find decay within days); scope usually narrow.

Priming experiments: interpreting the results

Common mis-interpretations.

- Assuming lab priming generalises to the field. Natural environments are full of self-selection into messages and competing primes and counter-cues that lab settings exclude; field-stage effects are often much smaller.
- Treating one-shot experimental effects as stable real-world effects. Most priming effects decay quickly unless reinforced by repeated exposure.
- Inferring that a null priming effect implies absence of the underlying consideration. The consideration may already be at maximum salience for some respondents – a ceiling effect.
- Treating heterogeneous effects as evidence of type heterogeneity. They may instead reflect heterogeneous baseline salience of the primed consideration.

A diagnostic checklist for reading such papers:

1. Is the prime documented to have registered (memory or recognition check)?
2. Are competing primes in the natural environment acknowledged?
3. Has the effect been tested for durability beyond the lab session?
4. Is the result interpreted as a salience shift, or (incorrectly) as a belief change?

Central observation. A priming effect is evidence about *which considerations are top-of-mind* – not about *what people believe*. The difference matters more than most papers acknowledge.