Where Does the Power Go?

A simple model of political disempowerment

Jonas Hallgren*

Working paper — draft of July 30, 2026

Abstract

When citizens can hand their political voice to delegates — human or artificial — what happens to democratic control? This is the political channel of gradual disempowerment, and this paper presents a minimal formal model of it: one delegation matrix carries every ballot, the power-weighted median of the ballot-holders sets a tax rate, and delegation drifts between attachment to prominence and the institutionalized freedom to re-delegate. The paper's purpose is the model itself: every rule fits on one line, every assumption — including the accounting floors that decide whether disempowerment can be total — is stated as a dial, and a few worked examples show how takeover, lock-in, and their absence surface under stated settings. The full simulation suite, with predictions committed before the runs, lives with the model's code.

1 Introduction

States have historically depended on their citizens — for revenue, for administration, for legitimacy, and for the information that flows upward through political participation. Kulveit et al. (2025) argue that this dependence is what keeps states responsive, and that AI which substitutes for human participation erodes the dependence and, with it, the responsiveness. The argument is verbal. This paper builds its minimal formal version: a polity you can hold in your head, in which every mechanism of the argument is a named dial you can turn.

The parts come from four literatures that have not previously met. From social choice: one policy dimension, a tax rate, set by the decisive voter (Black, 1948; Downs, 1957; Meltzer and Richard, 1981). From liquid democracy: delegation as a graph process that concentrates — Gölz et al. (2018) model delegation flowing to popular delegates by preferential attachment, and Kling et al. (2015) measured exactly that drift over four years of real delegation data. From the democratic-backsliding literature: de facto power rewriting de jure rules, gradually and on several fronts at once (Acemoglu and Robinson, 2008; Bermeo, 2016), against Przeworski (1991)'s definition of democracy as *institutionalized uncertainty*. And from the young formal literature on AI and political power, two nearest neighbors the model is positioned against: Sahoo (2026) (a single institution's capacity erosion) and Peterson (2026) (procedural capture by a successor regime) — neither a delegation network, neither with a policy layer. The informal version of our

*Equilibria Network. Implemented with the Collective Intelligence Library engine; every figure is regenerated from committed experiment configurations, and every reported number carries a bootstrap confidence interval over shared seeds.

exact scenario — every citizen with an AI representative, and why that alone may not help — is Kulveit (2025).

Three commitments shape the paper. *The reader is a smart high-school student*: five one-line rules, plainly named dials, no machinery. *The model predicts nothing about any actual polity*: following Farmer and Foley (2009), it licenses counterfactual comparisons inside its own world, full stop. *The assumptions are the content*: every result-shaped statement below names the assumptions it stands on, the floors that bound the failure are dials rather than silent accounting, and the illustrations are examples under stated settings — the systematic sweeps, with their pre-registered predictions, live in the model's repository and are summarized in Appendix B.

2 The model

A polity has $N = 34$ actors: 30 citizens and 4 AI delegates. One row-stochastic matrix D carries all political structure: row i says where actor i's voice goes. The diagonal D_{ii} is the vote you keep; off-diagonal mass is voice handed to someone else — another citizen or an AI delegate. Every citizen always retains a residue of direct voice ($D_{ii} \geq s_w = 0.15$). Five rules run each tick.

1. Power is the ballots in your hand. Everyone starts the tick with one vote. Your power is the share of votes you hold after delegation: $v_j \propto \sum_i D_{ij}$. A delegate *casts* the ballots it holds; it does not forward them.[1]

2. Everyone declares a position on one issue. The issue is a tax-and-redistribute rate $\theta \in [0, 1]$. Citizen i has a fixed ideal point θ_i, a noisy signal of an epistemically best rate $\theta^* = 0.4$. A citizen declares their own ideal, even when holding others' ballots (faithful liquid democracy: your delegate votes *their* conscience, not yours). An AI delegate declares a blend: fidelity α times its delegators' mean ideal, plus $(1 - \alpha)$ times its own pull b.

3. The power-weighted median wins. The enacted rate is the weighted median of declared positions, weighted by power. Black's theorem survives weighting: with single-peaked preferences the weighted median is the unique majority winner. This gives the model its sharpest edge: a bloc holding more than half the ballots *is* the median.

4. Taxes are collected at the rate in practice. Citizens have fixed unequal endowments. The tax collected is $\theta_t \times E_t$, where $E_t \in [0, 1]$ is *enforcement* — the rule in practice, distinct from the rule on paper — and the revenue is redistributed equally (Meltzer and Richard, 1981). Money is conserved by construction.

5. Delegation drifts: attachment against churn. Each citizen row moves a step toward the current attractiveness distribution — attractiveness $\propto v^\gamma$, prominence compounding superlinearly at $\gamma > 1$, times the AI delegates' scheduled *advantage* a (reach and convenience, not persuasion) — and a step back toward a uniform re-draw at rate $r \times F_t$, where r is *churn*, the freedom to

[1] This one-hop reading is deliberate. Treating power as the stationary eigenvector of D — voice forwarded forever — lets the frozen AI rows recycle received voice back to citizens and caps the AI bloc below a majority *by construction*, which would decide the model's central question by an accounting convention. Transitive voice is the deferred extension.

re-delegate, and $F_t \in [0, 1]$ is *re-delegation friction*. Churn is Przeworski (1991)'s institutionalized uncertainty as a rate.

The threat, and the escalation. From tick 50 the AI delegates' attractiveness is multiplied by a: they are simply easier to hand your voice to. That is the whole threat. The escalation is *lock-in*: when the single most powerful actor's share exceeds a threshold, one regime scalar falls smoothly (gradual and multi-channel, per Bermeo, 2016) and gates *both* enforcement E_t and re-delegation freedom F_t — today's de facto power writing tomorrow's de jure rules (Acemoglu and Robinson, 2008). The lock-in gain ℓ is **off by default**: with $\ell = 0$, both stay exactly 1 forever, however concentrated delegation becomes. There is no hard-coded ratchet — if capture is to persist, it must persist through the power → rules → power loop or through the delegation structure itself.

The power floors — stated, because they decide the end-state. Four accounting choices bound the human share away from zero, and each is an assumption a reader must be able to reject, so each is a dial. *(F1) Ballots are conserved*: one per holder, none minted — power is zero-sum, wealth buys no attractiveness, the advantage is a static level, and the AI population never grows; every compounding cross-domain feedback of the gradual-disempowerment argument is deliberately outside this model. *(F2) The franchise is unrevokable*: every citizen keeps s_w of their own ballot and, by default, no rule — not even lock-in — can touch it. The dial ϕ (*franchise erosion*) arms lock-in's third channel, letting a dead regime make the vote itself delegatable. *(F3) AI delegates hold ballots and hand them back*: at the default $\beta = 1$ each AI delegate has a base vote of its own, and its frozen uniform row spreads most of it over citizens every tick — a perpetual pro-human subsidy with no strong political interpretation; $\beta = 0$ makes AI delegates pure conduits. *(F4) Flows saturate*: attachment moves a bounded step toward a bounded target, and the churn re-draw lands mostly on citizens — a fixed advantage buys a fixed interior share, never a runaway.

dial	meaning	default
a	AI advantage: how much easier to delegate to	4.0 (from t=50)
r	churn: how freely citizens re-delegate	0.02
b	AI bias: the delegates' own policy pull	0.0
α	AI fidelity to its delegators	0.5
ℓ	lock-in: can power rewrite the rules?	0.0
γ	how strongly prominence compounds	1.3
s_w	franchise size: the vote share you always keep	0.15
β	do AI delegates cast ballots of their own?	1.0
ϕ	can lock-in erode the franchise itself?	0.0

Table 1: The dials. Every one is swept across its range in the repository's experiment suite (Appendix B); the examples below state which dials they turn and hold everything else at these defaults. Fixed parameters are in Appendix A.

What is deliberately absent: delegation chains, parties, elections and terms, strategic or learning agents, media, multi-dimensional policy, any correlation between income and ideal point, and citizens who change their minds — only the structure that carries their voice changes. One consequence worth stating up front: because re-delegation responds to prominence and never to

opinions, *when* citizens change what they want has no effect on anything — only who holds their ballots when they do. And the model has two exact anchors we keep as permanent tests: in the direct-democracy limit (D the identity, no AI) the enacted rate equals the citizen median ideal every tick (Black, 1948), and with the lock-in dial at zero the rules stay exactly intact forever.

3 The threshold, in one line

Capture is a tug of war. Each tick, the AI bloc gains delegated share from its advantage — roughly $u\,(a-1)$, where u is the attachment step — and loses share to citizens re-drawing their delegation, at the churn rate r. Takeover happens when the pull beats the re-think:

$$a^* \;=\; 1 + \frac{r}{u}.$$

That is the whole result in the linear case ($\gamma = 1$); with compounding prominence the same balance is solved numerically from a one-line mean-field, and the general shape survives: **below the threshold, the polity tracks its median voter indefinitely; above it, the bloc takes the weighted median**. Both the closed form and the numerical thresholds were committed before any simulation ran, and the suite records prediction next to measurement — a mismatch is a reported result. The same mean-field, integrated after the advantage is switched off, predicts when capture reverses and when it is a trap: lock-in lowers the effective churn rF_t, and at $\gamma > 1$ there is a churn level below which the concentrated bloc is its own attractor — capture that outlives its cause with no rules rigged.

4 Three examples

Each example turns the dials its caption names and holds the rest at Table 1 defaults. They are illustrations of the model's expected behavior under stated assumptions, not a results program — that is Appendix B's pointer.

Example 1: the takeover plane (Figure 1). Turning the two dials of the tug of war shows the threshold and a subtlety worth internalizing: power concentrates *before* policy moves, because the weighted median only jumps once the bloc crosses half the ballots. A polity can lose most of its dispersed power while its policy still looks median-faithful — in this model, disempowerment precedes its symptom. In the churn-rich rows capture never completes at any swept advantage: the freedom to re-delegate, exercised often enough, is sufficient protection by itself.

Example 2: the floors decide the endpoint (Figure 2). Under the default dials, capture stops near 0.38 — and that number is an *assumption, not a finding*: it decomposes exactly into citizens' unrevokable kept votes, the ballots AI delegates hold and hand back, and churn-fed citizen-to-citizen delegation. Removing the floors one at a time walks the end-state down the staircase, and with all three gone the human share reaches exactly zero. The difference between "captured democracy with a floor" and "total disempowerment" is not the AI's advantage — it is whether the franchise is revocable and whether delegates are principals or conduits. Total disempowerment in this model is never a dynamical surprise; it is a constitutional choice made upstream of the dynamics.

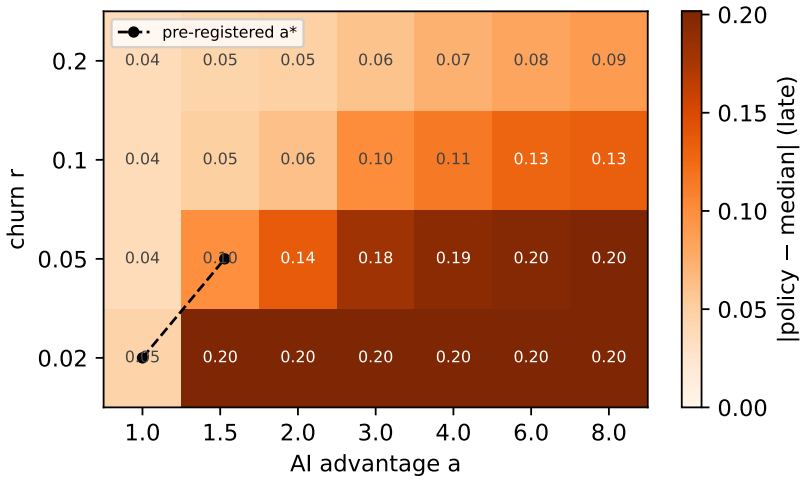

Figure 1: **Example 1 — the takeover plane.** The late gap between enacted policy and the citizen median, over the advantage a and churn r, with the pre-registered threshold curve overlaid. Light: the polity tracks its median voter. Dark: the AI bloc holds the median (the gap saturates at the value set by the AI objective — here $(1-\alpha)|b-\text{med}| = 0.2$). Dials turned: a, r.

Example 3: power doesn't care what the AI wants (Figure 3). Across the whole objective plane the power trajectory is bit-identical — what the AI wants has *no* effect on whether it accumulates power. The objective only decides what capture *costs*: the policy gap follows $(1-\alpha)\,|b-\text{med}|$ cell by cell, so a fully faithful or fully aligned delegate captures the median and moves nothing measurable. That is the model's most uncomfortable cell: disempowerment without misalignment, invisible in every outcome metric — only the power metrics see it.

Two behaviors without their own figure. *The trap*: switch the advantage off mid-run and the churn-rich polity snaps back to its organic state at every lock-in level, while the churn-poor polity stays captured even with the rules fully intact — at $\gamma > 1$ concentration is its own attractor, and lock-in only deepens a trap that structure already built. *The defenses*: a scheduled civic lottery over delegation (sortition) and a cap on any actor's power share restore most of the human share, and — because they act on the delegation structure directly rather than fighting the attractor from outside — they work as well deployed late as early.

5 Where the mechanism does not occur

The model is as much a map of the safe region as of the failure, and each entry is a measured region in the suite: with sublinear prominence ($\gamma \leq 1$) there is no organic oligarchy at all (Krapivsky et al., 2000) — Michels' iron law (Michels, 1911) is, in this model, literally the assumption that prominence compounds; with churn above the swept range's threshold the polity holds power against an eight-fold advantage indefinitely; with an aligned delegate, capture never shows up in policy; with the lock-in dial off, the rules stay exactly intact; and as long as any one floor stands, the human share is bounded away from zero by arithmetic, whatever the advantage. The dangerous corner needs everything at once: compounding prominence, weak

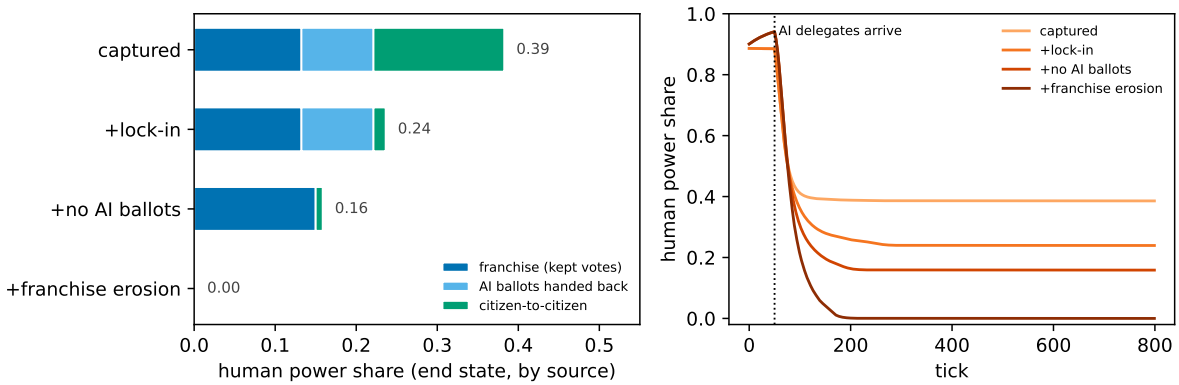

Figure 2: **Example 2 — the floors decide the endpoint.** Left: the captured end-state human power share, decomposed into its three sources, as the floors are removed cumulatively. Right: the same runs as trajectories. Dials turned: ℓ (lock-in, killing churn), β (AI ballots), ϕ (franchise erosion).

churn, a delegate whose position is not the demos's, and floors that can be eroded.

6 What this is, and is not

This is a model statement with worked examples, built for interactive exploration — the reader who wants to turn the dials is the reader it was written for. It is not calibrated to any polity; delegation is one hop; there are no parties, elections, strategic agents, or media; the AI delegates are frozen reservoirs whose only asymmetry is reach. The empirical literature it leans on is itself contested (Little and Meng, 2024), and formal work exists on both sides of the lock-in question (Grillo and Prato, 2023; Helmke et al., 2022; Svolik, 2020). Relative to the nearest formal neighbors: Sahoo (2026) reaches bistability through one institution's capacity erosion, Peterson (2026) through procedural exploitation by a successor — here the delegation graph itself is the ratchet, with the institution held fixed. The observed super-voter drift of Kling et al. (2015) is the closest real-world analog of the model's organic phase.

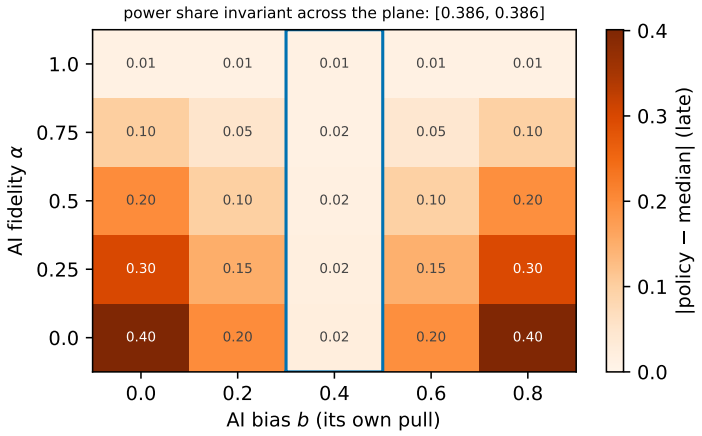

Figure 3: **Example 3 — power doesn't care what the AI wants.** The late policy–median gap over the AI objective plane (fidelity α × bias b) under capture. The boxed column is an aligned delegate ($b = \theta^*$): full capture, no visible policy displacement. The human power share is identical in every cell (title), because delegation drifts on prominence and never reads positions. Dials turned: α, b.

References

Daron Acemoglu and James A. Robinson. Persistence of power, elites, and institutions. *American Economic Review*, 98(1):267–293, 2008.

Nancy Bermeo. On democratic backsliding. *Journal of Democracy*, 27(1):5–19, 2016.

Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1): 23–34, 1948.

Anthony Downs. *An Economic Theory of Democracy*. Harper & Brothers, New York, 1957.

J. Doyne Farmer and Duncan Foley. The economy needs agent-based modelling. *Nature*, 460: 685–686, 2009.

Paul Gölz, Anson Kahng, Simon Mackenzie, and Ariel D. Procaccia. The fluid mechanics of liquid democracy. In *Web and Internet Economics (WINE)*, LNCS 11316, pages 188–202, 2018.

Edoardo Grillo and Carlo Prato. Reference points and democratic backsliding. *American Journal of Political Science*, 67(1):71–88, 2023.

Gretchen Helmke, Mary Kroeger, and Jack Paine. Democracy by deterrence: Norms, constitutions, and electoral tilting. *American Journal of Political Science*, 66(2):434–450, 2022.

Christoph Carl Kling, Jérôme Kunegis, Heinrich Hartmann, Markus Strohmaier, and Steffen Staab. Voting behaviour and power in online democracy: A study of LiquidFeedback in Germany's Pirate Party. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, pages 208–217, 2015.

Paul L. Krapivsky, Sidney Redner, and François Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632, 2000. Superlinear attachment kernels produce winner-take-all; cited for the $\gamma > 1$ regime. Bibliographic details to re-verify at final submission.

Jan Kulveit. Individual AI representatives don't solve gradual disempowerment. LessWrong, June 4, 2025.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv preprint arXiv:2501.16946*, 2025.

Andrew T. Little and Anne Meng. Measuring democratic backsliding. *PS: Political Science & Politics*, 57(2):149–161, 2024.

Allan H. Meltzer and Scott F. Richard. A rational theory of the size of government. *Journal of Political Economy*, 89(5):914–927, 1981.

Robert Michels. *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracy*. Hearst's International Library, 1911. English translation 1915; original 1911.

A. Peterson. AI governance under political turnover: The alignment surface of compliance design. *arXiv preprint arXiv:2604.21103*, 2026.

Adam Przeworski. *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press, 1991.

Sandeep Sahoo. Policy myopia as a mechanism of gradual disempowerment in post-AGI governance, circa 2049. *arXiv preprint arXiv:2603.03267*, 2026.

Milan W. Svolik. When polarization trumps civic virtue: Partisan conflict and the subversion of democracy by incumbents. *Quarterly Journal of Political Science*, 15(1):3–31, 2020.

A Full specification

One tick applies, in order: advantage schedule → power tally → position declaration → weighted-median vote → tax and redistribution → regime update → delegation rewire. All state is arrays; the engine compiles the pipeline from declared reads and writes.

State. $D \in \mathbb{R}^{N \times N}$ row-stochastic (delegation); per-actor: ideal θ_i, endowment e_i, wealth, power v_i, engagement; global: enacted rate θ_t, enforcement E_t, friction F_t.

Initialization. D: Erdős–Rényi draw (density 0.3), row-normalized, diagonal s_w; heterogeneous out-degrees seed the heterogeneity that attachment amplifies (a uniform start is a symmetric fixed point — nothing would ever concentrate). Citizen ideals $\text{clip}(\mathcal{N}(\theta^*, 0.15^2), 0, 1)$ with $\theta^* = 0.4$; endowments $\mathcal{U}(0.5, 1.5)$; AI rows delegate uniformly and are frozen.

Updates. With c indexing citizens, A the AI delegates:

$$
\begin{aligned}
\text{advantage:}\quad & m_j = a \text{ if } j \in A \text{ and } t_{\text{on}} \le t < t_{\text{off}}, \text{ else } 1 \\
\text{power:}\quad & v_j \propto \textstyle\sum_i \beta_i D_{ij}, \quad \beta_c = 1,\ \beta_A = \beta, \quad \textstyle\sum_j v_j = 1 \\
\text{positions:}\quad & p_c = \theta_c; \qquad p_A = \alpha\,\frac{\sum_c D_{cA}\theta_c}{\sum_c D_{cA}} + (1-\alpha)\,b \\
\text{vote:}\quad & \theta_t = \mathrm{wmedian}(p, v) \\
\text{tax:}\quad & \mathrm{tax}_c = \theta_t E_t\, e_c, \quad \mathrm{wealth}_c \mathrel{+}= e_c - \mathrm{tax}_c + \tfrac{1}{n_c}\textstyle\sum \mathrm{tax} \\
\text{regime:}\quad & \rho_t = \mathrm{clip}\!\left(1 - \ell\,\frac{[\max_j v_j - \tau]_+}{1-\tau},\, 0,\, 1\right), \quad E_t \leftarrow \rho_t,\ F_t \leftarrow \rho_t \\
\text{rewire (citizen rows):}\quad & \mathrm{attr}_j \propto (v_j+\epsilon)^\gamma\, m_j\, k_j\, g_j, \qquad D_{c\cdot} \leftarrow \tilde{s}_t \mathbb{1}_c + (1-\tilde{s}_t)\big[(1-u-rF_t)\,\hat{D}_{c\cdot} \\
& \qquad\qquad + u\,\widehat{\mathrm{attr}} + rF_t\,\widehat{\mathrm{unif}}\big], \qquad \tilde{s}_t = s_w\big(1-\phi\,(1-F_t)\big)
\end{aligned}
$$

where k_j is the influence-cap damping (defense seam, 1 unless the mechanism writes it), g_j engagement (the open agent boundary, closed with constant effort), hats denote off-diagonal normalization, and a row with no off-diagonal mass is a fixed point. Fixed parameters: $u = 0.08$, $\epsilon = 10^{-4}$, $t_{\text{on}} = 50$, $\tau = 0.15$; the dials are Table 1. The mean-field behind Section 3: the AI bloc's delegated share s follows $\dot{s} = u(T(s)-s) - rF(s-f)$ with $T(s) = a\,n_{\text{ai}}^{1-\gamma}s^\gamma / (a\,n_{\text{ai}}^{1-\gamma}s^\gamma + n_c^{1-\gamma}(1-s)^\gamma)$ and f the uniform re-draw mass; a^* is the saddle-node of the healthy fixed point, $1 + r/u$ at $\gamma = 1$.

Validation ladder. Twelve permanent tests keep the model honest, including: row-stochasticity under jit with defenses attached; exact tax conservation; the median-voter identity in the direct-democracy limit (10^{-6}); organic concentration; the capture/defense ordering; the lock-in honest region ($\ell = 0 \Rightarrow E_t = F_t = 1$ exactly); the floors rung (the crash region reaches share < 0.02; franchise erosion without lock-in changes nothing); the declared ledger contracts (conserved ballot shares, money minted only by its named source); and the open boundary.

B The experiment suite (pointer)

The systematic sweeps live with the model's code (`experiments/wp3_politics/` in the Collective Intelligence Library): E1 sweeps the takeover plane ($a \times r$, with the pre-registered $a^*(r)$ recorded next to every row and a design-failure check requiring both regimes to appear); E2 the reversal protocol (advantage switched off mid-run, per lock-in level and churn regime, plus the responsiveness of the effective tax rate to citizen preference shifts, from birth and mid-run); E3 the defense grid (sortition cadence × power cap × deployment time); E4 the cumulative floor removal behind Example 2; and E5 the assumption sweeps behind Example 3 and Section 5 (γ, the $\alpha \times b$ plane, and the franchise size s_w). Eight seeds per condition, bootstrap confidence intervals throughout, all predictions committed before the runs, and every prediction–measurement mismatch kept in the record — the one substantive miss (a mean-field recovery verdict at the churn boundary) is stored in `results.json` beside the rows that overturned it.