The Calibrated Preference–Forecast Mechanism

A Complete Working Dossier

Proposal · Stage-0 Theorems · Stage-1 Simulation

Draft 1 — June 2026

Contents

The through-line across all three parts: a mechanism, like a mind or a government, stays honest only by continuous contact with what it cannot internally generate. What can be proven (Parts II–III, in part) eliminates designs broken on paper; what cannot — legitimacy, value formation, and above all whether real concerns resist measurement (Part III, E3) — requires staged human deployment. The boundary between the two is not a limitation of the analysis; it is the analysis.

Vote Values Cardinally, Weight Beliefs by Calibration Not Cash: A Proposal for a Calibrated Preference–Forecast Governance Mechanism, With Its Own Adversarial Stress-Test and an Empirical Verification Program

A working proposal. Draft 1.

Abstract

This paper proposes a governance mechanism — the **Calibrated Preference–Forecast (CPF)** mechanism — that separates *what a population wants* from *what will actually happen* under each option, aggregates the first through cardinal (intensity-bearing) preference expression and the second through forecasts weighted by demonstrated accuracy rather than money, and uses an AI mediator to make the resulting bookkeeping, elicitation, and anti-collusion enforcement tractable at scale. The design is assembled almost entirely from existing parts. Its only candidate for novelty is a single substitution: replacing the *monetary betting* of Robin Hanson's futarchy with *non-monetary forecasting track record* as the weight on the empirical channel, which is intended to retain futarchy's accuracy-rewarding property while removing its plutocratic failure mode. The paper does three things in deliberately unequal measure. It (1) specifies the mechanism and situates it honestly against close prior art, stating explicitly where novelty is and is not claimed; (2) supplies an adversarial stress-test that separates what can be *proven* about the mechanism from what cannot; and (3) proposes the human, empirical research program required to verify the parts that proof cannot reach. The central methodological claim is that the third part is not optional polish but the load-bearing component: a governance mechanism validated only by its own internal mathematics would exhibit, in institutional form, the exact failure mode — coherence without external correction — that the mechanism is otherwise designed to resist.

---

1. Motivation: Arrow is real, but narrower than its slogan

Arrow's Impossibility Theorem (1951) is frequently summarized as "democracy is mathematically impossible." That summary is false in a specific and useful way. Arrow's theorem constrains the aggregation of *ordinal rankings* — systems in which voters report only the order of their preferences (first, second, third). For that representation, the impossibility is genuine: no rule can jointly satisfy unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives.

The impossibility is therefore an artifact of the *representation*, not of collective choice as such. Cardinal systems that capture preference *intensity* — range voting, approval voting, and quadratic voting — evade Arrow's conditions because they are not rank-aggregation systems. A related result, the Gibbard–Satterthwaite theorem (1973, 1975), shows that essentially all non-

dictatorial systems remain manipulable, but manipulability is a weaker and more manageable property than impossibility. The accurate statement is: *rank-aggregation democracy faces an impossibility that intensity-expression democracy circumvents, at the cost of new manipulability concerns.*

The reason cardinal systems are not already mainstream is practical, not theoretical. Intensity is hard to elicit honestly — everyone claims maximal intensity for everything — and hard to compare across people, since my "10" and your "10" need not denote the same felt magnitude. These are precisely the administrative frictions that an AI mediator, and a quadratic cost on intensity, are positioned to address. That observation is the opening this proposal occupies. It is not the novel part; it is the doorway to it.

---

2. Prior art (and an explicit statement of what is *not* claimed)

A proposal in this area that did not foreground its precedents would be committing exactly the error this document is built to avoid. The mechanism's components are all established:

Futarchy (Robin Hanson, 1999–2013). Hanson's "Decision Markets" (1999) and "Shall We Vote on Values, But Bet on Beliefs?" (Journal of Political Philosophy, 2013) propose that a polity democratically choose a welfare metric (the *values*), then delegate *which policies will best advance it* to prediction markets in which participants bet on conditional outcomes (the *beliefs*). A policy is adopted if its conditional market implies higher expected welfare. Vitalik Buterin's 2014 "Introduction to Futarchy" carried the idea into blockchain governance. This is the closest existing proposal to the present one, and it is the prior art against which novelty must be measured. **The CPF mechanism is, structurally, futarchy with two modifications** (Sections 3–4).

Quadratic voting and quadratic funding (Lalley & Weyl 2018; Posner & Weyl, *Radical Markets*, 2018; Buterin, Hitzig & Weyl 2019). Quadratic voting lets a participant buy votes from a budget at quadratic cost (a second vote on an issue costs four credits, a third costs nine), so intensity can be expressed but is expensive, forcing approximately honest revelation. Quadratic funding extends this to public-goods allocation and is used in production by Gitcoin. The Colorado state legislature trialed quadratic voting for budget prioritization in 2019. This is the cardinal-preference component, used here essentially unmodified.

AI-mediated deliberation — the Habermas Machine (Tessler, Bakker, et al., *Science*, 2024). DeepMind's Habermas Machine is an LLM trained to synthesize participants' opinions and critiques into consensus statements. In studies including a demographically representative UK citizens' assembly with over five thousand participants, it produced statements rated clearer, fairer, and less biased than those of human mediators, and left groups less divided. Notably, after processing critiques it tended to *over-weight* minority viewpoints — a finding directly relevant to the marginalization failure mode discussed in Section 5. This is the AI-mediation component, and the empirical demonstration that such mediation is feasible.

Generative social choice (Fish, Gölz, Parkes, Procaccia, et al., EC 2024; Boehmer, Fish & Procaccia 2025). This line pairs the rigor of social-choice theory with LLM capability, using a two-part methodology: first *prove* that a process satisfies representation guarantees given

oracle access to certain queries; then *empirically validate* that an LLM can approximately implement those queries. In a trial with 100 representative US residents, 84 felt "excellently" or "exceptionally" represented by an extracted five-statement slate. This matters here for two reasons: it is prior art for AI-augmented social choice, and its methodology is nearly identical to the verification program proposed in Section 6 — which both strengthens confidence in that program's shape and removes any claim that the methodology is itself novel.

Forecasting track records (Tetlock & Gardner, *Superforecasting*, 2015; IARPA Good Judgment Project). The empirical foundation for weighting influence by calibration rather than money: in the IARPA tournament, identified "superforecasters" beat control groups by large margins and, per Good Judgment's own analysis, beat futures markets by roughly 66% on forecasting Federal Reserve policy inflection points. Calibration is measured with *proper scoring rules* — the Brier (quadratic) score and the logarithmic score — which have the property (Gneiting & Raftery 2007) that expected score is uniquely maximized by reporting one's true probabilities. This is the basis for the empirical channel.

Real-world AI-augmented democracy. Polis (used in Taiwan's vTaiwan process under Audrey Tang) and Remesh demonstrate machine-learning-mediated collective input at civic scale; a 2025 *Nature Human Behaviour* review ("The impact of advanced AI systems on democracy") surveys the rapidly growing field.

What is therefore *not* claimed as novel: prediction-market or forecast-based governance (futarchy); cardinal/intensity voting (QV); AI-mediated deliberation (Habermas Machine); the LLM-plus-social-choice pairing and its prove-then-validate methodology (generative social choice); calibration scoring (proper scoring rules, Good Judgment). Each is someone else's contribution and is cited as such.

---

3. The proposed mechanism (CPF)

The slogan, adapting Hanson, is: **vote values cardinally, weight beliefs by calibration not cash.**

The mechanism runs two distinct channels and joins them.

Channel A — Values (egalitarian, cardinal). The population expresses *what it wants* — the welfare metric(s) and their relative weights, plus direct preferences over outcomes where relevant — using quadratic voting from an equal per-person credit budget. Everyone's budget is identical. This channel is deliberately *not* weighted by expertise or track record: what a society wants is not a question on which anyone has a better or worse "record," and weighting wants by skill is the road to technocracy. Channel A captures intensity (evading Arrow) while quadratic pricing forces approximately honest revelation (mitigating Gibbard–Satterthwaite). An AI mediator (Habermas-Machine-style) assists upstream by synthesizing free-form input into a coherent, proportionally representative set of candidate value-statements before the cardinal vote, so that the menu being priced is not itself a source of distortion.

Channel B — Beliefs (calibration-weighted, non-monetary). For each live policy option, participants forecast its consequences *against the value metric chosen in Channel A* —

probabilistic predictions with defined resolution criteria and dates. Each forecaster carries a **calibration weight** derived from their *resolved* track record under a strictly proper scoring rule (Brier or logarithmic). Influence on the belief aggregate is a function of demonstrated accuracy, continuously re-earned as forecasts resolve. Crucially, **the weight cannot be purchased**: it is not stake, wealth, or credentials, only a public, decaying, outcome-graded record. New participants begin at a baseline weight and earn standing by forecasting accurately on resolved questions.

The join. A policy's score is its calibration-weighted forecasted effect on the democratically-chosen welfare metric. Adoption favors the option that the best-calibrated aggregate predicts will most advance what the population (equally) said it wants. This is futarchy's values/beliefs separation preserved intact — but with the belief channel *de-monetized*.

The AI mediator's role is administrative and is the reason the mechanism is feasible now rather than in 1999: it (i) synthesizes free-form input into representative value menus (generative-social-choice function); (ii) maintains the calibration ledger and computes proper-scoring weights as questions resolve; (iii) performs the collusion- and Sybil-detection bookkeeping that overwhelmed human institutions attempting cardinal or market mechanisms; (iv) surfaces, to every participant, the strongest opposing forecasts and the prior-art objections to each policy, as a structural check against one-sided information. The mediator does *not* set values, choose welfare metrics, or adjudicate outcomes; those are reserved to Channel A and to reality, respectively.

---

4. What is novel, and with what confidence

Stated precisely, the candidate novelty is **one substitution and its consequences**: replacing futarchy's *monetary prediction market* with a *non-monetary, calibration-weighted forecast aggregate* on the belief channel, while retaining a *cardinal* (rather than ordinal or implicit) values channel and adding AI mediation to make both administrable.

The intended payoff of the substitution:

- It **cannot be plutocratic.** Money buys votes in a market; it cannot buy a resolved forecasting record. Influence over *beliefs* accrues only to demonstrated accuracy. This addresses the most common objection to futarchy — that it hands policy to whoever has the most capital to move markets.
- It **remains meritocratic only on the one axis where meritocracy is defensible** — accuracy about consequences — while remaining strictly egalitarian on values. It does not create a permanent technocratic class, because the record is continuously re-earned and decays, and anyone can build one from the baseline.
- It is **self-grading against reality.** Because the belief channel's weights are set by *resolved* outcomes, the system has a built-in, continuous contact with the external world. This is the property the mechanism most wants, for reasons developed in Section 7.

Novelty confidence: low-to-moderate, and explicitly uncertain. I was unable to locate this exact synthesis — calibration-weighted (rather than money-weighted) futarchy with a

cardinal values channel and an AI administrative layer — in the literature searched. But absence of a found precedent is not evidence of novelty; it is the precise epistemic situation in which confident novelty claims have repeatedly proven wrong (the 2025 episode in which an AI was claimed to have "solved" Erdős problems it had merely retrieved from the literature is the cautionary case). The substitution is *adjacent enough* to futarchy, to proxy/liquid-democracy weighting schemes, and to "epistocracy" debates that a dedicated social-choice scholar may well find a direct precedent. **This proposal should be read as: here is a synthesis that falls out of combining known parts under a stated principle, here is the closest prior art I found, and here is exactly where I believe it differs — verification of true novelty requires a domain expert and a more exhaustive search than was performed.**

---

5. Adversarial stress-test

This section attacks the mechanism. It separates failure modes that can be analyzed *mathematically* (and therefore caught before any human participates) from those that are *irreducibly empirical* (and can only be found by staged contact with real people). Conflating the two is itself one of the failure modes.

5.1 What yields to proof or simulation

- **Strategic misreporting / gaming.** Whether truthful participation is a Nash equilibrium is a mechanism-design question. Quadratic voting's honesty property is provable under stated assumptions; the residual manipulability guaranteed by Gibbard–Satterthwaite can be *characterized* (how much, what coalition size, what information is required). On the belief channel, the use of a strictly proper scoring rule (Brier/log) provably makes truthful probability reporting the unique expected-score maximizer for an individual forecaster — a guarantee against one class of gaming, available before deployment.
- **Aggregation pathologies.** One can prove which fairness axioms the joined mechanism satisfies and which it must violate, and identify the preference/forecast configurations that produce perverse outcomes (the analog of Condorcet cycles), bounding their frequency under distributional assumptions.
- **Collusion and Sybil resistance.** Whether actors can gain disproportionate influence via fake identities or secret coalitions is partly analyzable with cryptographic and game-theoretic tools — but only *relative to identity-verification assumptions that themselves require real-world enforcement*.
- **Dynamic stability of the calibration ledger.** Does belief-channel influence concentrate over time (a rich-get-richer dynamic in which a small set of early-accurate forecasters accrue unbounded standing), or does it circulate? This is a dynamical-systems question: write the weight-update equations, analyze fixed points and attractors, and where closed form resists, run agent-based simulation. **Weight decay and baseline re-entry are design knobs to be tuned here, and their adequacy is checkable in simulation before deployment.**

A substantial document of theorems and simulations characterizing the mechanism's strategic properties, aggregation behavior, and dynamic stability could in principle be produced before a single human participates. That work is *necessary*. It is not *sufficient*.

5.2 What does not yield to proof — the dangerous part

Every model above is conditional on assumptions about what humans value and how they behave, and those conditionals do enormous load-bearing work as smuggled-in empirical claims.

- **Goodhart on the welfare metric.** The moment a democratically-chosen metric becomes the optimization target of the belief channel, it is subject to Goodhart's law: the measure ceases to be a good measure. A forecasting apparatus optimizing "augmented GDP" or any chosen scalar will surface policies that move *the metric*, not necessarily *the underlying good the metric was meant to proxy*. How badly, and along which seams, is not derivable from first principles.
- **The marginalization failure mode (morally the most serious).** A calibration-weighted belief channel systematically privileges concerns that manifest as *predictable, resolvable consequences*. Harms that are diffuse, delayed, or borne by people whose situations are under-documented may simply not show up as forecastable quantities — and the already-marginalized are disproportionately in that category. You cannot model this away, because doing so would require already knowing whose concerns the system fails to register; if you knew that, you would have solved it. This is the system's own coherence concealing what its representation omits. (Note the Habermas Machine's observed tendency to *over-weight* minority views after critique processing: a possible mitigation lever on Channel A, but its transfer to this combined mechanism is untested.)
- **Preference formation and reflexivity.** The mechanism shapes the preferences it aggregates. A forecasting-weighted system may teach participants to care only about predictable outcomes, atrophying concern for important-but-uncertain goods. Whether, how fast, and to whom this happens is a fact about human psychology under institutional pressure, observable only by running it.
- **Legitimacy.** A governance system functions because people accept it as legitimate and comply without coercion, not because it is mathematically optimal. Whether humans would experience a calibration-weighted, AI-mediated, quadratically-priced system as *theirs* — as fair, as worth obeying when they lose — depends on cultural narratives and on perceived *procedural* fairness, which research (e.g., Tyler) finds often matters more than outcome fairness. A provably optimal system can collapse for feeling alienating. The "tyranny of the confidently correct" — a felt sense that the well-calibrated are being handed rule over everyone else — is a legitimacy risk even if the mechanism is, on paper, egalitarian on values.
- **The AI mediator as attack surface and as carrier of bias.** The mediator that synthesizes value menus, maintains the ledger, and surfaces opposing views is itself a single point of capture, manipulation, and ideological skew. The Habermas Machine has already drawn exactly this critique. An adversary who controls or biases the mediator controls the menu and the framing.
- **Adaptive adversaries.** Mathematical analysis assumes a fixed strategy space. Real adversaries invent new strategies, attacking at the layer the model did not formalize —

narrative manipulation, social engineering, exploiting the mediator's blind spots, manufacturing resolvable-but-misleading forecasting questions — and co-evolve with defenses. One can model known attack classes; one cannot model the attack nobody has yet conceived, and those are usually the consequential ones.

The honest synthesis: mathematics can do the *necessary* stress-testing (rule out whole classes of manipulation, characterize aggregation, bound dynamic instability) but not the *sufficient* stress-testing (establish that the mechanism is actually good for actual humans), because the latter depends on Goodhart dynamics, marginalization, preference formation, legitimacy, mediator capture, and adaptive adversaries — all empirical. A mechanism whose designers presented its mathematical coverage as complete would be exhibiting in miniature the precise pathology that makes powerful optimizers dangerous: the hundred valid steps before the one fatal, unexamined assumption.

---

6. The human verification research program

The research required to verify the mechanism follows directly from the boundary in Section 5, in escalating stages, each gated on the previous revealing no catastrophic surprise. The structure deliberately mirrors the prove-then-validate methodology of generative social choice (Fish et al.), generalized to include the moral and political variables that representation guarantees do not cover.

Stage 0 — Prove what is provable (theorists). Mechanism-design and social-choice researchers establish the formal properties of Section 5.1: equilibrium analysis of truthful participation, proper-scoring guarantees, axiom satisfaction/violation, Sybil/collusion resistance relative to explicit identity assumptions, and dynamical analysis of the calibration ledger's concentration behavior. Deliverable: a theorems-and-bounds document that eliminates designs broken on paper. *Gate: no design admitted to Stage 1 that is provably manipulable in a low-cost way or provably converges to oligarchy under realistic parameters.*

Stage 1 — Agent-based simulation (computational social scientists). Populate the mechanism with synthetic agents of varying rationality, varying calibration, and explicit adversarial intent. Search for emergent pathologies that resist closed form: ledger concentration under strategic question-selection, coalition dynamics, sensitivity of outcomes to mediator-induced menu framing. Tune the design knobs (weight decay rate, baseline re-entry, quadratic budget size, resolution-window choice). *Gate: no uncontrollable instability or adversarial exploit surviving reasonable parameter tuning.*

Stage 2 — Small-scale empirical deployment (experimental political scientists, survey methodologists, STS scholars, and — non-negotiably — members of the communities most at risk of marginalization, as co-designers rather than subjects). Real humans, genuinely low stakes (a club, a budgeting exercise, an online community governing itself), intensive observation. This stage exists specifically to surface what the mathematics *cannot see*, and the measurement plan must therefore target exactly those variables: - *Marginalization audit.* Independently identify, in advance, which stakeholders' concerns are least likely to manifest as resolvable forecasts; then measure whether those concerns were in fact suppressed. This requires the omitted parties in the room defining "concern," because the whole danger is

that the mechanism's own representation cannot register what it omits. - *Goodhart audit.* Track divergence between the chosen welfare metric and qualitative assessments of the underlying good over time. - *Preference-formation tracking.* Longitudinal measurement of whether participants' expressed values drift toward the predictable-and-resolvable. - *Legitimacy measurement.* Perceived procedural fairness, sense of ownership, willingness to comply when on the losing side — disaggregated by whether a participant carries high or low calibration weight. - *Mediator-bias probing.* Red-team the AI mediator for menu-framing skew and ideological drift; vary the mediator and measure outcome sensitivity. - *Novel-attack discovery.* Open adversarial bounty: invite participants to break it, and treat every successful exploit as data about the unformalized strategy space. *Gate: each scale-up to higher stakes is contingent on the prior stage revealing no catastrophic empirical surprise on these axes.*

Stage 3 — Iterated, gated scale-up. Stakes rise only as each lower stage clears. At every level the empirical audits of Stage 2 are re-run, because pathologies (especially marginalization and legitimacy) can be scale-dependent and can re-emerge.

The disciplines required are therefore not only mathematics and machine learning but experimental political science, survey methodology, science-and-technology studies, ethics, and the direct participation of those the mechanism might silence. This breadth is not interdisciplinary garnish. It is the operational form of the principle in Section 7.

---

7. Why the human program is the load-bearing part, not an appendix

The structure of this proposal is itself an instance of its own central principle, and the principle is the reason to take the human program as primary.

The mathematical model of the mechanism is a single carving of the governance problem: internally rigorous, provably consistent where it is provable at all, and blind to exactly what it does not represent. The empirical research program is the *independent carving* that catches what the model's coherence conceals. A designer — human or AI — who built this mechanism and trusted its own internal analysis would be committing the failure mode the mechanism is otherwise meant to resist: achieving coherence by severing contact with the reality it cannot internally generate. The beauty of the proofs and the validity of the theorems would make that severance *more* seductive and *more* dangerous, not less.

This is also why the mechanism's one defensible design ambition is to be *self-grading against reality*: the belief channel's weights are set by resolved outcomes, so the system is built around continuous external correction rather than internal coherence. But that ambition is hollow if the *design process* does not hold itself to the same standard. A self-correcting mechanism validated by a non-self-correcting method is a contradiction. The human verification program is how the design process stays open to the carvings it does not itself produce — which is the only property that distinguishes a governance design worth deploying from an elegant, coherent, and quietly catastrophic one.

The boundary between what yields to proof and what requires human contact is not a limitation of the analysis. It *is* the analysis. Keeping that boundary visible — in the mechanism and in the method that validates it — is the whole task.

8. Conclusion

The Calibrated Preference–Forecast mechanism is futarchy with the money taken out of the belief channel and replaced by calibration, a cardinal values channel kept egalitarian, and an AI mediator doing the administrative work that made such mechanisms impractical before. Its components are entirely borrowed and individually well-validated; its only candidate novelty is the de-monetizing substitution, claimed with explicit and low-to-moderate confidence pending expert search for direct precedent. Its most important content is not the design but the honest map of what can and cannot be proven about it, and the staged, gated, irreducibly human research program that the unprovable part demands. If the mechanism has any value, that value will be established the same way the mechanism itself proposes to govern: not by the coherence of the argument, but by what survives contact with a reality the argument cannot contain.

References

1. Arrow, K. J. *Social Choice and Individual Values.* Wiley, 1951 (2nd ed. 1963).
2. Gibbard, A. "Manipulation of Voting Schemes: A General Result." *Econometrica*, 1973. Satterthwaite, M. "Strategy-proofness and Arrow's Conditions." *J. Economic Theory*, 1975.
3. Hanson, R. "Decision Markets." *IEEE Intelligent Systems*, 1999; "Shall We Vote on Values, But Bet on Beliefs?" *Journal of Political Philosophy*, 2013.
4. Buterin, V. "An Introduction to Futarchy." 2014.
5. Lalley, S. & Weyl, E. G. "Quadratic Voting: How Mechanism Design Can Radicalize Democracy." *AEA Papers & Proceedings*, 2018. Posner, E. & Weyl, E. G. *Radical Markets*, Princeton, 2018.
6. Buterin, V., Hitzig, Z. & Weyl, E. G. "A Flexible Design for Funding Public Goods." *Management Science*, 2019. (Quadratic funding; deployed by Gitcoin.)
7. Tessler, M. H., Bakker, M. A., et al. "AI can help humans find common ground in democratic deliberation." *Science*, 386, 2024. (The Habermas Machine.)
8. Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I. & Wüthrich, M. "Generative Social Choice." *EC 2024.* Boehmer, N., Fish, S. & Procaccia, A. D. "Generative Social Choice: The Next Generation." 2025.
9. Tetlock, P. & Gardner, D. *Superforecasting: The Art and Science of Prediction.* Crown, 2015. (IARPA Good Judgment Project.)
10. Brier, G. W. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review*, 1950. Gneiting, T. & Raftery, A. E. "Strictly Proper Scoring Rules, Prediction, and Estimation." *JASA*, 2007.
11. Tyler, T. R. *Why People Obey the Law.* (On procedural justice and legitimacy.)
12. "The impact of advanced AI systems on democracy." *Nature Human Behaviour*, 2025. (Survey.)
13. Small, C. et al., Polis; the vTaiwan / digital-democracy process associated with Audrey Tang. Remesh.

14. Graeber, D. & Wengrow, D. *The Dawn of Everything*, 2021. (On the indigenous critique and non-mainstream political forms — the motivating analogy that a viable democratic form may be available but historically un-adopted.)

Note on sourcing: the prior-art descriptions above were drawn from web search conducted for this draft and paraphrased; quotations were avoided. Page-level verification against the primary sources, and an expert search for any direct precedent of the calibration-weighted substitution, remain to be done.

Stage-0 Theorems and Bounds for the Calibrated Preference–Forecast Mechanism

What can be proven before any human participates — and the explicit limits of that proof

Stage-0: Provable Properties of the Calibrated Preference–Forecast (CPF) Mechanism

Companion to the CPF proposal. This document establishes the formal properties that can be settled by proof or simulation in advance of deployment. Everything here is ***necessary*** *and none of it is* ***sufficient****: these results eliminate designs that are broken on paper; they are silent on the empirical failure modes (Goodhart drift, legibility/marginalization, legitimacy, mediator capture, adaptive adversaries), which are reserved for the staged human program.*

---

0. Setup and notation

A population N with $|N| = n$ governs via two channels.

Values channel. Each voter i holds an equal credit budget B and allocates it across issues by *quadratic voting*: casting k votes on an issue costs k^2 credits. The channel outputs a cardinal tally, not merely an ordering.

Belief channel. For each live policy option there is a resolution event Y tied to the democratically-chosen welfare metric. Each forecaster j reports a probability p_j for Y. Forecaster j carries a weight $w_j \geq 0$, $\sum_j w_j = 1$, derived from her resolved track record. The belief aggregate weights reports by w_j. After resolution, each forecaster receives a score $S(p_j, y)$ under a fixed scoring rule, which updates her track record and hence her future weight.

A policy's mechanism score is its w-weighted forecasted effect on the welfare metric; adoption favors the highest such score.

We treat the values channel (§2–3), the belief channel's elicitation (§1), manipulability of both (§3), and the dynamics of the weight ledger (§4) in turn.

---

1. Theorem 1 — Truthful elicitation on the belief channel

Theorem 1. *Under either the logarithmic or the Brier (quadratic) scoring rule, a forecaster whose payoff equals her score on a question strictly maximizes her expected payoff by reporting her true subjective probability.*

Proof (binary outcome $Y \in \{0,1\}$, true belief $q = \Pr[Y=1]$).

Logarithmic rule. $S(p,y) = y\ln p + (1-y)\ln(1-p)$. Expected score under the true belief:

$$G(p) = q\ln p + (1-q)\ln(1-p).$$

$$G'(p) = \frac{q}{p} - \frac{1-q}{1-p}, \qquad G'(p) = 0 \iff q(1-p) = (1-q)p \iff p = q.$$

$$G''(p) = -\frac{q}{p^2} - \frac{1-q}{(1-p)^2} < 0,$$

so G is strictly concave and $p = q$ is its unique global maximizer. ■

Brier rule. $S(p,y) = -(p-y)^2$. Then

$$G(p) = -\big[\, q(1-p)^2 + (1-q)p^2 \,\big], \qquad G'(p) = 2(q-p),$$

so $G'(p) = 0 \iff p = q$, and $G''(p) = -2 < 0$: unique maximizer at $p = q$. ■

The categorical case (outcomes on the simplex) follows from the general theory of strictly proper scoring rules; the logarithmic and quadratic rules are strictly proper, and the logarithmic rule is the unique *local* proper rule (the score depends only on the probability assigned to the realized outcome). See Gneiting & Raftery (2007).

Corollary 1.1 (scope — a real gap, not a footnote). Theorem 1 guarantees truthful reporting *only when the forecaster's objective is exactly her score on that question.* In CPF the score also updates her **weight**, hence her **future influence over policy**. A forecaster who values influence instrumentally — to push outcomes toward her preferred values — has effective objective

$$\text{score} \;+\; \lambda \cdot (\text{value of future influence}),$$

and for $\lambda > 0$ truthful reporting need not maximize the combined objective. Properness therefore removes the *myopic* incentive to misreport but **not** the *strategic* one. This motivates two design responses, both to be evaluated downstream: (a) separating, where feasible, the identity that forecasts from the identity that votes on values, so forecasting cannot be used as a lever on one's own value-preferences; and (b) constructing the weight map so that score cannot be profitably traded against policy influence. We flag (a)/(b) as **open** at Stage 0.

---

2. Proposition 2 — Arrow-inapplicability, and its price

Proposition 2. *Arrow's impossibility theorem does not apply to the CPF values channel; the channel escapes it by leaving the ordinal framework, at the explicit cost of assuming interpersonal comparability supplied by budget normalization.*

Argument. Arrow's theorem concerns a social welfare function mapping each profile of individual *orderings* to a social *ordering*, required to satisfy unrestricted domain, the Pareto principle, independence of irrelevant alternatives, and non-dictatorship; for ≥ 3 alternatives no such function exists. The CPF values channel does not take orderings as input nor produce an ordering as its primitive output — it aggregates *cardinal* purchased intensities and outputs a cardinal tally. Arrow's hypotheses are thus not satisfied and the impossibility is vacuously inapplicable. This is the standard sense in which range, approval, and quadratic voting "evade" Arrow.

The price (stated honestly). Cardinal aggregation requires that the cardinal units be interpersonally comparable. QV supplies comparability by *normalization*: an identical budget B for every voter makes "one budget's worth of influence" the common unit. CPF therefore does not obtain something for nothing; it trades Arrow's ordinal-invariance for an explicit, defensible equal-influence normalization. The normative weight rests on that normalization, which must be argued, not assumed away.

Efficiency (cited, not reproven). Lalley & Weyl (2018) show that under independent private values and price-taking behavior in a large population, the QV allocation converges to the utilitarian-efficient outcome, with the efficiency loss vanishing as $n \to \infty$. The load-bearing assumptions are **price-taking** and **independent private values**; collusion and large pivotal stakeholders violate them and return us to §3 and to the empirical program.

---

3. Proposition 3 — Manipulability is unavoidable; characterize and bound it

Proposition 3. *The deterministic core of CPF is manipulable (Gibbard–Satterthwaite); manipulation can nonetheless be characterized and bounded on each channel.*

Gibbard–Satterthwaite (1973, 1975). Every onto, non-dictatorial, deterministic social choice function on ≥ 3 alternatives is manipulable: some profile admits a voter who gains by misreporting. CPF is therefore manipulable. This is structural, not a fixable defect; randomization (Gibbard 1977) can restore strategyproofness only by sacrificing ex-post efficiency, which we do not adopt.

Values channel — bounded by quadratic cost. With cost k^2 for k votes, the marginal cost of the k-th vote is

$$k^2 - (k-1)^2 = 2k - 1,$$

rising linearly. A voter optimally equates this marginal cost to the expected marginal benefit (pivotality $\times$ value), so reported intensity tracks true intensity up to the pivotality term, and the gap from honest revelation is bounded by budget exhaustion and shrinks toward zero as pivotality falls (large n). Intensity exaggeration is thus *purchasable but self-limiting.*

Belief channel — myopic incentive removed, two residuals. Theorem 1 removes the myopic misreport incentive. Two residual manipulations remain:

1. *Strategic-influence misreporting* (Corollary 1.1): misreporting to position one's weight, addressed only by the design responses (a)/(b) above.
2. *Question-selection farming*: forecasting only easy/predictable questions to accumulate weight cheaply. Mitigation: score against a **difficulty baseline** (e.g., the prevailing consensus or a reference forecast) so that weight rewards *skill* — accuracy beyond what the easy baseline already achieves — rather than raw accuracy. The residual after baselining is the forecaster's ability to find questions where her *edge over baseline* is reliably positive, which is precisely the ability the mechanism wants to reward; the attack and the desired behavior converge under proper baselining. We flag the choice of baseline as a tuning decision for Stage 1.

---

4. Theorem 4 — Ledger dynamics: when does forecasting influence concentrate?

This is the formal heart of the anti-plutocracy claim. The question: does belief-channel weight concentrate into a self-perpetuating oligarchy, or circulate? The answer depends decisively on the

form of the weighting rule, and the difference is provable.

Let forecaster j have a true skill parameter giving per-question scores $s_{j,t}$ with mean μ_j and bounded variance σ^2, drawn from a stationary process across questions t.

Scheme A — averaging (mean-score weighting)

Track record is an exponentially-weighted moving average of past scores,

$$r_{j,t} = (1-\lambda)\sum_{u\le t}\lambda^{t-u}\, s_{j,u}, \qquad 0<\lambda<1,$$

and weight is a softmax of records with steepness β,

$$w_{j,t} = \frac{\exp(\beta\, r_{j,t})}{\sum_k \exp(\beta\, r_{k,t})}.$$

Proposition 4A. *Under Scheme A, with stationary per-question scores of overlapping support:*

1. *$r_{j,t}$ does not converge to a point but to a stationary random variable with mean μ_j and variance*

$$\mathrm{Var}(r_{j,\infty}) = \sigma^2\,\frac{1-\lambda}{1+\lambda},$$

   bounded and tunable by λ.
2. *Weights fluctuate around $w_j^* \propto \exp(\beta\mu_j)$; for finite β no forecaster's weight tends to 1, so there is no absorbing dictator. The joint weight process is a stationary, ergodic process on the simplex with a unique stationary distribution; rankings can and do change as records fluctuate (circulation persists).*
3. *Concentration is monotone in β: as $\beta\to\infty$, weight collapses onto* $\arg\max_j \mu_j$ *(winner-take-all among the most skilled); as $\beta\to 0$, weight $\to$ uniform. β is the explicit knob trading meritocracy against pluralism; λ trades responsiveness against stability.*

Why (variance). With geometric weights $(1-\lambda)\lambda^k$ summing to 1 on i.i.d. scores of variance σ^2,

$$\mathrm{Var}(r_{j,\infty}) = (1-\lambda)^2\sigma^2\sum_{k\ge 0}\lambda^{2k} = \frac{(1-\lambda)^2\sigma^2}{1-\lambda^2} = \sigma^2\frac{1-\lambda}{1+\lambda}.$$

Why (no dictator / ergodicity). $r_{j,t}$ is a bounded-input bounded-output stable linear filter ($|\lambda|<1$) of a stationary process, hence itself stationary and ergodic; the softmax is a continuous bounded map, so $w_{j,t}$ is stationary and ergodic. For finite β and overlapping score supports, weights stay almost surely bounded away from the simplex vertices, so no vertex is absorbing. (Stated under the explicit assumptions; the conclusion is the qualitative one — bounded, circulating, unique stationary law — not a closed-form stationary distribution.)

Scheme B — accumulation (compounding weighting), the failure mode

Suppose instead weight compounds multiplicatively on performance with no renormalization of history,

$$w_{j,t+1} \propto w_{j,t}\,\exp(\eta\, s_{j,t}), \qquad \eta>0.$$

Proposition 4B. *Under Scheme B,* $\log w_{j,t}$ *is a random walk with drift* $\eta\mu_j$*; the log-ratio between any two forecasters,*

$$\log w_{j,t} - \log w_{k,t} \;\sim\; \text{random walk with drift } \eta(\mu_j - \mu_k),$$

so the highest-μ *forecaster's relative weight grows without bound (winner-take-all almost surely), and among equal-skill forecasters stochastic drift alone produces large transient concentration (Gibrat's law* $\Rightarrow$ *heavy-tailed / approximately log-normal weights).*

This is the "rich get richer" oligarchy. It is the *same compounding dynamic* that concentrates capital in markets — and it is exactly the dynamic of monetary prediction markets, in which wealth compounds with winning bets.

Corollary 4.1 — the formal content of the anti-plutocracy claim

CPF **must** weight by *averaging* (Scheme A), not *accumulation* (Scheme B), to avoid a built-in oligarchy. Futarchy's monetary markets are effectively Scheme B; replacing money with **averaged calibration** is precisely the move from B to A. This is the provable core of the proposal's central claim — and note its scope: it is a statement about the *weighting rule*, recoverable by proof and simulation. It says nothing about the **legibility** worry (see §5).

Proposition 4.2 — the feedback condition that re-breaks even Scheme A

Even averaging concentrates if **weight feeds back into the ability to score**. Formally, if the skill mean is independent of weight, $\mu_j \perp w_j$, Scheme A stays benign; but if $\mu_j = \mu_j(w_j)$ is increasing — weight buys privileged information, or high-weight forecasters control which questions are posed — the system reacquires Scheme-B-type instability. Hence an explicit Stage-0 **design constraint**: *weight must confer aggregation influence only, never privileged information or control of the question agenda.*

---

5. Scope limits — what Stage 0 does *not* establish

These results are necessary and not sufficient. They are silent on, and must not be presented as covering:

- **Goodhart drift.** Once the welfare metric is the optimization target, it degrades as a proxy for the underlying good. Not addressed here.
- **The legibility / marginalization bias.** Theorem 4 defuses the *weight-concentration* form of "rich get richer," but the sharper worry lives in the **question space** — *which* concerns ever become resolvable forecasts — not in the weight dynamics. Calibration-weighting structurally privileges the documentable, and that bias is untouched by anything proven above. It is the central empirical question of the human program, and possibly fatal to the proposal; Stage 0 cannot adjudicate it.
- **Legitimacy.** Whether participants experience the mechanism as theirs and comply when they lose is irreducibly empirical (procedural-justice research; Tyler).
- **Mediator capture and bias.** The AI mediator that builds value menus, maintains the ledger, and surfaces opposing views is a single point of capture and ideological skew; no theorem here constrains it.

- **Adaptive adversaries.** All of the above assumes fixed strategy spaces. Real adversaries invent unmodeled strategies; those are usually the consequential ones.

The honest reading of this document: it can prove the mechanism is *not obviously broken* on several specific axes, and it can prove the anti-plutocracy weighting claim in its narrow (weight-dynamics) sense. It cannot establish that the mechanism is *good for actual humans*. That requires the staged, gated, human-empirical program — which is not a supplement to this analysis but the part that catches what this analysis, by construction, cannot see.

References (Stage-0)

- Arrow, K. J. *Social Choice and Individual Values*, 1951/1963.
- Gibbard, A. "Manipulation of Voting Schemes." *Econometrica*, 1973; "Manipulation of Schemes That Mix Voting with Chance." *Econometrica*, 1977. Satterthwaite, M. *J. Econ. Theory*, 1975.
- Gneiting, T. & Raftery, A. E. "Strictly Proper Scoring Rules, Prediction, and Estimation." *JASA*, 2007. Brier, G. W. *Monthly Weather Review*, 1950.
- Lalley, S. & Weyl, E. G. "Quadratic Voting: How Mechanism Design Can Radicalize Democracy." *AEA P&P*, 2018.
- (Background) Tetlock & Gardner, *Superforecasting*, 2015; Hanson, "Shall We Vote on Values, But Bet on Beliefs?", *J. Political Philosophy*, 2013.

Stage-1 Agent-Based Simulation of the Calibrated Preference–Forecast Mechanism

Making the ledger dynamics and the legibility bias measurable

Stage-1: Agent-Based Simulation of the CPF Ledger

Companion to the CPF proposal and the Stage-0 theorems. Stage 0 proved what is provable in closed form; Stage 1 simulates the regimes proof cannot reach in closed form, and operationalizes the one failure mode Stage 0 explicitly could not touch — the legibility bias. Source: `cpf_stage1_sim.py`.

Interpretive caveat, stated up front. Each experiment demonstrates a failure *mechanism* **conditional on its modeling assumptions**. Experiment 3 in particular *assumes* that some real concerns are illegible — that they resolve rarely or noisily as forecasting questions. The simulation shows the *consequence* of that assumption; it does not establish that real polities contain illegible concerns, nor which constituencies hold them. That is the Stage-2 empirical question no simulation can settle.

E1 — Averaging vs accumulation (validating Theorem 4)

Sixty forecasters with heterogeneous skill forecast 2,500 resolving questions. Belief-channel weight is computed two ways: by **averaging** (exponentially-weighted mean score into a softmax — Scheme A, the CPF design) and by **accumulation** (multiplicative compounding on score — Scheme B, the dynamic of money-based markets). Concentration is measured by the Gini coefficient of the weight vector over time.

Result. Averaging settles at a Gini of about **0.34** — bounded, circulating, tracking skill without locking in. Accumulation climbs to about **0.98**, the winner-take-all ceiling. The anti-plutocracy weighting claim survives simulation: the choice of weighting rule is decisive, and the rule CPF specifies behaves as Theorem 4 predicts. This is good news for the design.

E2 — The feedback threshold (Proposition 4.2)

Proposition 4.2 warned that even averaging concentrates if weight feeds back into the ability to score — if high standing buys privileged information or control of the question agenda. Here the effective skill of forecaster j is boosted by a term proportional to her current weight, $\mu_j^{\text{eff}} = \mu_j + \gamma\,(w_j - \bar{w})$, and the feedback strength γ is swept.

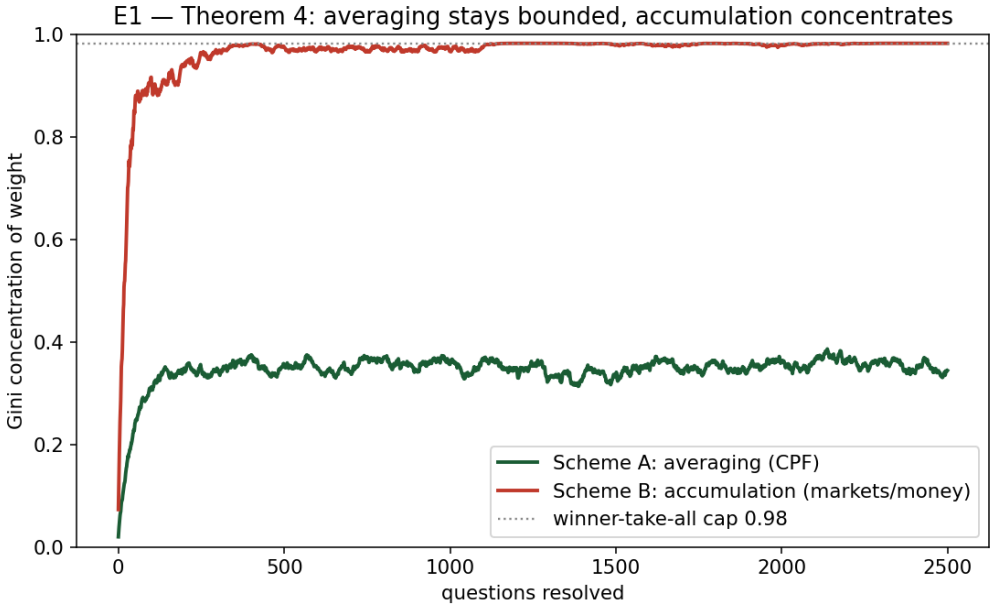

Figure 1: E1: averaging holds weight concentration bounded and circulating; accumulation runs to winner-take-all.

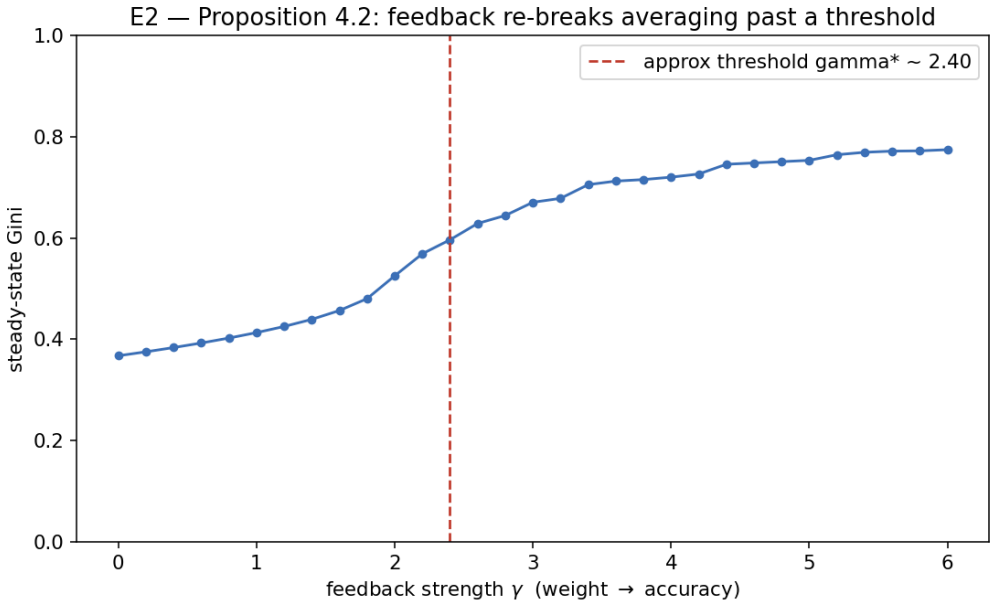

Figure 2: E2: past a feedback threshold near gamma ~ 2.4, averaging stops protecting against concentration.

Result. Steady-state concentration stays low while γ is small, then rises sharply past a threshold near $\gamma^* \approx 2.4$, climbing from a Gini of about 0.37 toward 0.78. This converts an abstract design constraint into an enforceable, measurable red line: the implementation must keep the weight-to-accuracy coupling below this regime — concretely, **weight must confer aggregation influence only, never privileged information or control of which questions are posed.**

E3 — The legibility / marginalization bias (the load-bearing result)

This experiment models the **question space**, which Theorem 4 does not reach. Concerns differ in *legibility* — how cleanly they resolve as forecasting questions. Legible concerns (measurable outcomes) resolve reliably; illegible concerns (diffuse, delayed, or borne by under-documented groups) resolve rarely and noisily. Equal numbers of each arrive every round, representing two constituencies with equal claim to representation.

Forecasters earn calibration weight only on resolved questions, so under **strategic selection** they chase legible questions to protect their records. Strategic intensity is parameterized by α. The representation ratio R = influence(legible)/influence(illegible) measures the bias ($R = 1$ is fair). Critically, the experiment also runs the **difficulty-baselining** mitigation proposed in §3 of the theorems note — scoring forecasters against the per-question consensus — to test whether it repairs the bias.

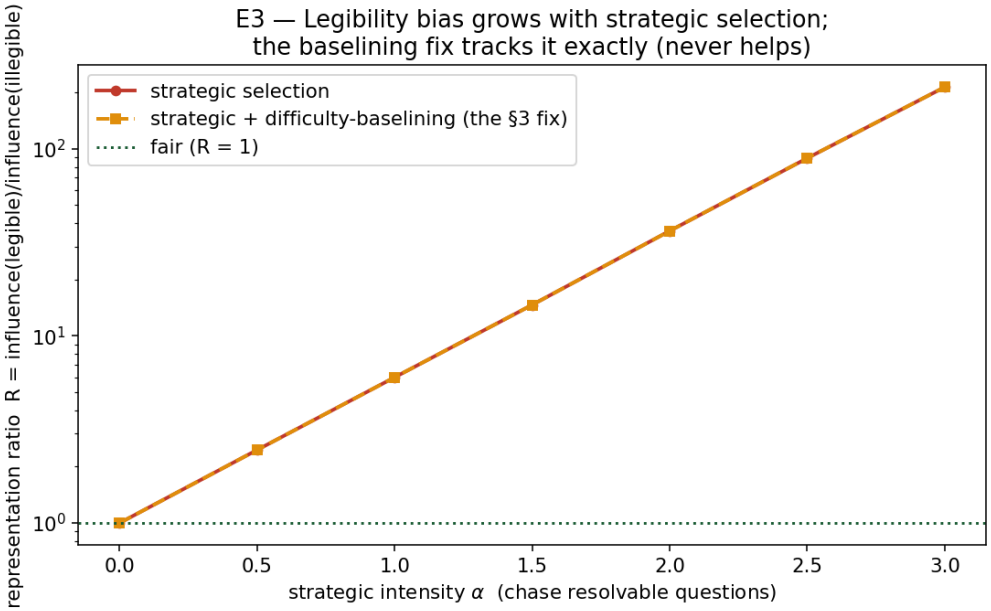

Figure 3: E3: the legibility bias grows with strategic intensity, and the difficulty-baselining fix tracks it exactly — the two curves are indistinguishable.

Result. With no strategic selection ($\alpha = 0$) the system is fair, $R = 1$. As forecasters grow strategic, illegible concerns are starved of belief-channel influence: $R \approx 6$ at moderate intensity, ≈ 36 higher up, scaling with the legibility gap. The decisive finding is the overlap in the figure: the difficulty-baselining fix matches the unmitigated curve to three decimal places at every intensity (mean ratio

1.000). Baselining acts on *scoring*; the bias lives in *coverage* of the question space; the scoring-side fix cannot touch it.

What this means. The marginalization objection is real, structural, and **not patchable by the obvious scoring-side remedy**. Repairing it requires a *coverage-side* intervention — mandated forecasting of illegible concerns, subsidized illegible-question markets, or reserved influence — none of which is in the current design, and each of which reintroduces its own manipulation surface (and so must be re-run through the Stage-0 analysis). The proposal's plutocracy defense holds (E1) with a quantified guardrail (E2); its deeper legibility problem is confirmed and shown to survive the easy fix (E3), and that is the result most likely to determine whether the mechanism is worth pursuing.

On the magnitude. The size of R is set by strategic intensity and the legibility gap and is *illustrative, not predictive*. Only two findings are robust: the **direction** (bias rises above fair as selection becomes strategic) and the **invariance** (baselining never helps). The real-world questions — whether illegible concerns exist, how strategic forecasters actually are, whose interests fall on which side — require staged human deployment, the independent carving that catches what this model, by construction, cannot.