Bonded Dissent: A Falsification Protocol for Parameter Disputes in Permissionless Networks

Withheld for double-anonymous review

Manuscript prepared for submission to *Management Science*

Abstract

A permissionless network lets an authority set parameters and lets participants object. Objection is costless, so the volume of complaint carries little information about whether a parameter is misset: it reflects coordination, private interest, and cascades as readily as a real defect. The authority's response is unconstrained, and a one-sided threat to sanction the network places no symmetric burden on the party holding decision rights. We propose the *Bonded Falsification Protocol*: a dispute proceeds only if the challenger names a specific alternative, states a falsifiable prediction about a pre-agreed metric, and posts a bond; the authority defends the status quo only by posting a bond of its own; and the dispute is settled neither by a vote nor by a forecast but by a pre-registered controlled experiment that runs both settings and measures the predicted metric. We prove four results. The bond separates falsifiable beliefs from cheap talk. Committing to settle by experiment strictly raises expected network value over discretion under complaint volume. The decision is invariant to the number of complainants, so a coalition gains no influence from its size. The authority's defense bond restores symmetry to an otherwise one-sided relation, so that an authority confident in its parameters strictly prefers the protocol while a refusal to bond is itself a costly signal. A forecast market cannot substitute for the experiment, and we position it against futarchy, adversarial collaboration, optimistic dispute games, and dominant-assurance contracts. A worked application to a live registration-cost dispute closes the paper.

Keywords: governance; mechanism design; skin in the game; falsifiability; decision markets; optimistic dispute resolution; token networks.

1 Introduction

Every permissionless network that pays its participants must decide who sets its parameters and how those choices are contested. The registration cost, the reward schedule, the inflation rate, the eligibility rules: each is a number or a rule that a controlling authority, a foundation or a subnet owner, selects and can revise. Participants who dislike a choice can complain, and complaint is free. They can post in a forum, coordinate in a chat channel, or lobby the authority directly, at no cost and with no commitment. The authority, for its part, can change the parameter, refuse to change it, or threaten a sanction, and it too bears no formal cost for being wrong.

This arrangement has two failures, and they compound. The first is that costless complaint is uninformative. A complaint that a parameter is set too high looks identical whether it comes from a participant who has discovered a genuine inefficiency or from one who would simply prefer a private advantage, and the number of complaints reflects how well the complainers coordinate at least as much as whether they are right (Olson, 1965; Banerjee, 1992; Bikhchandani et al., 1992; Janis, 1972). The classical result that a costless message cannot separate types in equilibrium (Crawford and Sobel, 1982) applies directly: the authority cannot read the volume or the vehemence of objection as evidence about the parameter. The second failure is that the authority's side of the exchange is unbonded. A threat of the form "change this parameter or I will sanction the network" transfers all the downside of an error to the participants. The party with decision rights, the party best placed to be held to account, is the one party with nothing at stake. Agency theory has long held that decision rights should carry symmetric downside (Jensen and Meckling, 1976; Holmström, 1979); in network governance they routinely do not.

The two failures are sharpest when the objection is not that a parameter is mis-set but that the authority is acting in bad faith. An accusation of this kind, that a network is a fraud, that its activity is manufactured, that its operator extracts value rather than creating it, is costless to make and expensive to suffer. It transfers a reputational loss onto the accused while the accuser stays unbonded, and its persuasive force, like that of any costless message, tracks the coordination of those who repeat it more than its truth. The law of defamation answers the same asymmetry in another domain by attaching liability to a false and damaging assertion of fact; a permissionless network affords no such recourse, and the accused can refute a serious charge only by absorbing

its cost. A bond supplies the missing accountability. Requiring the accuser to stake a falsifiable prediction converts an imputation of bad faith into a claim that can be tested and that costs its maker when it fails, which is the standing that a costless accusation lacks. In this light the protocol gives a serious accusation a price, and an honest authority a way to be cleared by evidence rather than by volume.

We propose a mechanism that repairs both failures at once. The *Bonded Falsification Protocol* (BFP) admits a dispute over a parameter only when the challenger does three things together: names a specific alternative value, states a falsifiable prediction about a metric agreed in advance, and posts a bond. The authority may concede, in which case the alternative is adopted, or it may defend, which it can do only by posting a bond of its own. A defended dispute is then settled by a pre-registered controlled experiment that runs the status quo and the proposed alternative side by side, for a fixed horizon, and measures the predicted metric. The experiment, not a vote and not a market price, is the adjudicator. The party whose prediction the experiment contradicts forfeits its bond.

Each ingredient does work that the others cannot do alone, and the contribution of the paper is precisely the combination. A bond without an experiment, as in optimistic dispute games (Kalodner et al., 2018; Teutsch and Reitwießner, 2019) or token-curated arbitration (Lesaege et al., 2019), settles a claim about a fact that already exists, by re-execution or by a vote; but whether one parameter is better than another is a counterfactual that does not yet exist and must be generated. An experiment without a bond, as in adversarial collaboration (Mellers et al., 2001), places no filter on who may demand one, so a coalition can impose unbounded experimentation cost and the cheap-talk channel is never closed. A forecast without an experiment, as in futarchy (Hanson, 2013), aggregates beliefs about the counterfactual but never runs the control, and it inherits an impossibility, that no deterministic decision rule supports a strictly proper conditional market (Othman and Sandholm, 2010; Chen et al., 2011), which running the control sidesteps. And none of these places a symmetric stake on the authority. The defense bond is the empty cell that the protocol fills.

We make the argument precise. After positioning the protocol in the literature (Section 2, with the coverage matrix in Table 1), we model the governance problem and its two pathologies (Section 3), state the protocol formally (Section 4), and prove four results (Section 5). Theorem 1 shows the

bond separates genuine falsifiable beliefs from cheap talk. Theorem 2 shows that committing to settle by experiment strictly raises expected network value relative to discretion under complaint volume. Theorem 3 shows the decision is invariant to the number of complainants, so coalition size buys no influence. Theorem 4 shows the defense bond restores symmetry and that a confident authority strictly prefers the protocol, with the corollary that refusing to bond is a costly signal of low confidence. Proposition 1 establishes why the experiment cannot be replaced by a forecast, and Proposition 2 treats who pays for the experiment. Section 6 addresses metric gaming, collusion, bond calibration, and experiment validity. Section 7 applies the protocol to a live registration-cost dispute, and Section 8 concludes.

2 Related work

Decision markets and futarchy. The proposal to decide policy by betting on outcomes is Hanson (2013): vote on values, bet on beliefs, and let a conditional prediction market choose the policy that the market expects to maximize a welfare metric. The closest relative of our protocol, and the one a reader will reach for first, is therefore futarchy. The decisive difference is that futarchy decides on a *forecast* of the counterfactual and never realizes it: the market on the policy that is not adopted is voided, so its prices need not reflect beliefs. Othman and Sandholm (2010) and Chen et al. (2011) formalize this as an impossibility, that no deterministic decision rule admits a strictly proper conditional market. Our protocol does not forecast the counterfactual; it runs it. We discuss the formal consequence in Proposition 1.

Adversarial collaboration. In the methodology of adversarial collaboration (Mellers et al., 2001), disputants who disagree design in advance an experiment whose outcome both will accept, then run it and abide by the result. This is the intellectual ancestor of the experiment-as-adjudicator idea, and we adopt its pre-registration discipline. It carries no monetary bond and presumes good-faith academic disputants rather than an authority with decision rights and a crowd with private incentives; it neither filters cheap talk nor binds an authority.

Optimistic dispute resolution. Optimistic rollups and verification games (Kalodner et al., 2018; Teutsch and Reitwießner, 2019) let anyone assert a result, post a bond, and be challenged, with the

Table 1: Coverage of the four ingredients across related mechanisms. A mechanism resolves a parameter dispute well only if it filters cheap talk (a bond), states what would change one's mind (a falsifiable prediction), settles by generating the counterfactual (an experiment), and binds the party with decision rights (a symmetric authority stake). The Bonded Falsification Protocol is the only row with all four.

Mechanism	Monetary bond	Falsifiable prediction	Experiment adjudicator	Symmetric authority
Futarchy / decision market (Hanson, 2013)	no	partial	no	no
Adversarial collaboration (Mellers et al., 2001)	no	yes	yes	no
Optimistic rollup / verification game (Kalodner et al., 2018; Teutsch and Reitwießner, 2019)	yes	no	no	no
Token-curated arbitration (Lesaege et al., 2019)	yes	no	no	no
Dominant-assurance contract (Tabarrok, 1998)	yes	no	no	partial
Randomized policy trial	no	yes	yes	no
Bonded Falsification Protocol (this paper)	**yes**	**yes**	**yes**	**yes**

bond awarded to the winner of an on-chain adjudication. Token-curated arbitration (Lesaege et al., 2019) bonds jurors and disputants and settles by vote. These mechanisms supply the bond and, in the rollup case, a symmetric stake between two disputants. What they adjudicate, however, is the correctness of a computation against a pre-existing ground truth, recovered by re-execution or by a vote of jurors. They do not generate new evidence, and they place no stake on a governing authority, because they have none.

Assurance and skin in the game. The dominant-assurance contract (Tabarrok, 1998) makes an entrepreneur bear downside, refunding contributors with a bonus if a public good fails to fund, and is the closest precedent for putting the party that proposes an action on the hook. It contains no falsifiable prediction and no experiment. The general principle that decision rights should carry symmetric downside is the agency-theoretic point of Jensen and Meckling (1976) and Holmström (1979); that a costly action separates types where a costless message cannot is Spence (1973) and Crawford and Sobel (1982); and the behavioral account of why costless voice aggregates into volume uncorrelated with truth is Olson (1965), Banerjee (1992), Bikhchandani et al. (1992), Janis (1972), and Hirschman (1970). Our contribution is not any one of these ideas but their assembly into a single mechanism, and the identification of the symmetric authority stake as the missing piece.

3 The governance problem

A network operates a parameter θ currently set to a status-quo value θ_0. A challenger proposes an alternative θ_1. Exactly one of two states of the world holds: either θ_1 is genuinely better than θ_0 on an agreed welfare metric m, which we write $\omega = G$ (good change), or it is not, $\omega = B$ (bad change). Adopting the better parameter is worth $G_v > 0$ to the network; adopting the worse parameter costs $L_v > 0$. Let $\pi = \mathbb{P}(\omega = G)$ be the common prior.

A *governance rule* maps the dispute to a decision $D \in \{\theta_0, \theta_1\}$. We compare rules by expected network value, $\mathbb{E}[V] = \mathbb{P}(\text{adopt better}) \cdot G_v - \mathbb{P}(\text{adopt worse}) \cdot L_v$, net of any operating cost.

Pathology 1: complaint volume is uninformative. Under the status-quo rule, call it discretion, the authority adopts θ_1 when objection to θ_0 is loud enough, that is, when an observed complaint volume N exceeds a threshold. The difficulty is that N is generated by a process that does not depend on ω. Complaints are costless, so by Crawford and Sobel (1982) they cannot separate the state, and their number is governed by how well the objectors coordinate, by cascades in which each complaint invites the next (Banerjee, 1992; Bikhchandani et al., 1992), by the concentrated interest of a minority that gains privately from θ_1 (Olson, 1965), and by group polarization (Janis, 1972). Formally we take $\mathbb{P}(N > \text{threshold} \mid \omega) = \rho$ independent of ω. Then discretion adopts θ_1 with the same probability ρ whether the change is good or bad, and

$$\mathbb{E}[V_{\text{disc}}] = \rho\big(\pi\, G_v - (1-\pi) L_v\big).$$

The decision is uncorrelated with the truth; the rule is a coin weighted by how loud the room is.

Pathology 2: the authority is unbonded. Discretion also lets the authority defend θ_0, or impose it, at no cost to itself when it is wrong. A threat of the form "adopt θ_1, or I sanction the network" is a transfer of all error cost to the participants. There is no value of the dispute at which the authority forfeits anything for being mistaken, so the threat conveys no information about the authority's confidence and disciplines nothing. We will say a governance rule is *symmetric* if the authority's expected loss from defending a status quo that the evidence overturns is of the same form as the challenger's expected loss from pressing a change the evidence rejects. Discretion is

maximally asymmetric: the challenger who is wrong wastes effort, but the authority who is wrong loses nothing.

4 The Bonded Falsification Protocol

The protocol is a four-stage game between a challenger C (possibly a coalition) and the authority A, with a pre-agreed metric set $\mathcal{M}$, a neutral experiment operator, and publicly known bond sizes.

Stage 0 (Filing). C files a dispute consisting of a specific alternative θ_1, a metric $m \in \mathcal{M}$, and a falsifiable prediction $\mathbb{E}[m(\theta_1)] \geq \mathbb{E}[m(\theta_0)] + \delta$ for a stated effect size $\delta > 0$, and posts a bond b_C. A filing that does not name all of (θ_1, m, δ) is not in order and is ignored.

Stage 1 (Joinder). A responds within a fixed window. It may *concede*, adopting θ_1 and ending the dispute, or *defend*, which requires posting a bond b_A and agreeing to the pre-registered experiment below. If A neither concedes nor defends within the window, the challenge is *upheld by default* and θ_1 is adopted.

Stage 2 (Experiment). A pre-registered controlled experiment E runs two arms, θ_0 and θ_1, identical in all else, on a sandbox or test deployment, for a fixed horizon H, and produces an estimate $\hat{\Delta}$ of $m(\theta_1) - m(\theta_0)$ under a pre-committed estimator and test. The analysis plan, including H, the estimator, and the decision threshold, is fixed before the experiment runs and cannot be revised by either party.

Stage 3 (Settlement). If $\hat{\Delta} \geq \delta$, the prediction is *confirmed*: C wins, A forfeits b_A, and θ_1 is adopted. Otherwise the prediction is *falsified*: A wins, C forfeits b_C, and θ_0 stands. The forfeited bond is awarded to the winner, burned, or split between the winner and the experiment fund, according to the variant in use (Section 6).

The experiment has the usual two error rates: size $\alpha = \mathbb{P}(\hat{\Delta} \geq \delta \mid \omega = B)$, the chance it confirms a bad change, and power $1 - \beta = \mathbb{P}(\hat{\Delta} \geq \delta \mid \omega = G)$, the chance it confirms a good one. We assume throughout that the experiment is *informative*, $1 - \beta > \alpha$, which a sufficient horizon H delivers under standard conditions, and that running it costs $\kappa > 0$.

5 Results

5.1 The bond filters cheap talk

A challenger holds a private posterior $q = \mathbb{P}(\hat{\Delta} \geq \delta \mid C\text{'s information})$ that the experiment will confirm the prediction. Filing costs a small effort $\varphi \geq 0$ beyond the bond. With the winner taking the loser's bond, the challenger's expected payoff from filing is

$$U_C(q) = q\, b_A - (1 - q)\, b_C - \varphi,$$

and not filing yields 0.

Theorem 1 (Incentive compatibility / separation)**.** *A challenger files if and only if*

$$q \;\geq\; q^* \;=\; \frac{b_C + \varphi}{b_A + b_C}.$$

For symmetric bonds $b_A = b_C = b$, $q^* = \frac{1}{2} + \frac{\varphi}{2b} > \frac{1}{2}$: only a challenger who believes the experiment is more likely than not to confirm the prediction will file. A participant who lobbies for a change that does not in fact improve the metric, the private-benefit type for whom $\omega = B$, faces confirmation probability equal to the experiment's size α; for any bonds with $q^* > \alpha$, that participant strictly prefers not to file. Cheap talk is filtered.

Proof. $U_C(q) \geq 0$ rearranges to $q(b_A + b_C) \geq b_C + \varphi$, that is $q \geq q^*$, and U_C is strictly increasing in q, so the cutoff rule is exact. Substituting $b_A = b_C = b$ gives $q^* = \frac{1}{2} + \varphi/(2b)$. A private-benefit type advances a change for which $\omega = B$; conditional on $\omega = B$ the experiment confirms with probability α by the definition of size, so this type's posterior is $q = \alpha$. Choosing bonds with $(b_C + \varphi)/(b_A + b_C) > \alpha$, which holds for all $b_C > 0$ once $\alpha < 1/2$ and b_A is not too large relative to b_C, makes $U_C(\alpha) < 0$, so this type abstains. Only challengers with private posteriors above q^* self-select into filing. ■

The bond is a costly signal in the sense of Spence (1973): it is cheap to bear for a challenger who expects to win and expensive for one who does not, so willingness to post it separates belief from noise. The volume of unbonded complaint, which the authority could not read at all under Pathology 1, is replaced by a smaller set of bonded, falsifiable, testable claims.

5.2 Settling by experiment raises welfare

Theorem 2 (Welfare improvement)**.** *Suppose an in-order dispute exists. The protocol with settlement by experiment yields*

$$\mathbb{E}[V_{BFP}] = \pi(1-\beta)G_v - (1-\pi)\,\alpha\,L_v - \kappa,$$

and $\mathbb{E}[V_{BFP}] > \mathbb{E}[V_{disc}]$ *whenever the experiment is informative* ($1-\beta > \alpha$)*, complaint volume is uninformative* (ρ *independent of* ω *with* $\alpha \le \rho \le 1-\beta$)*, and the experiment cost satisfies*

$$\kappa \;<\; \pi\big(1-\beta-\rho\big)G_v + (1-\pi)\big(\rho-\alpha\big)L_v.$$

Proof. Under the protocol θ_1 is adopted exactly when the experiment confirms, which happens with probability $1-\beta$ in state G and α in state B, giving the stated $\mathbb{E}[V_{\mathrm{BFP}}]$ after subtracting the experiment cost. Subtracting $\mathbb{E}[V_{\mathrm{disc}}] = \rho(\pi G_v - (1-\pi)L_v)$,

$$\mathbb{E}[V_{\mathrm{BFP}}] - \mathbb{E}[V_{\mathrm{disc}}] = \pi\big(1-\beta-\rho\big)G_v + (1-\pi)\big(\rho-\alpha\big)L_v - \kappa.$$

Both bracketed terms are nonnegative under the stated condition on ρ and at least one is positive when the experiment is informative, so the difference is positive precisely when κ is below the displayed bound. ■

The content of Theorem 2 is that discretion makes the right decision only by luck, because ρ does not depend on the truth, whereas the experiment makes the right decision with probability tied to its power and size. The protocol pays a known cost κ to convert a coin flip into an informative test, and the conversion is worth it whenever the test is informative and the experiment is not too expensive. Two corollaries deserve note. First, the bound on κ widens as the stakes G_v and L_v grow, so the more a parameter matters, the more clearly the protocol dominates discretion. Second, because Theorem 1 already restricts filings to challengers with $q \ge q^*$, the population of disputes that reach Stage 2 is enriched for the good state relative to the prior, which only raises the realized value of the protocol. Figure 1 illustrates both theorems at a representative calibration.

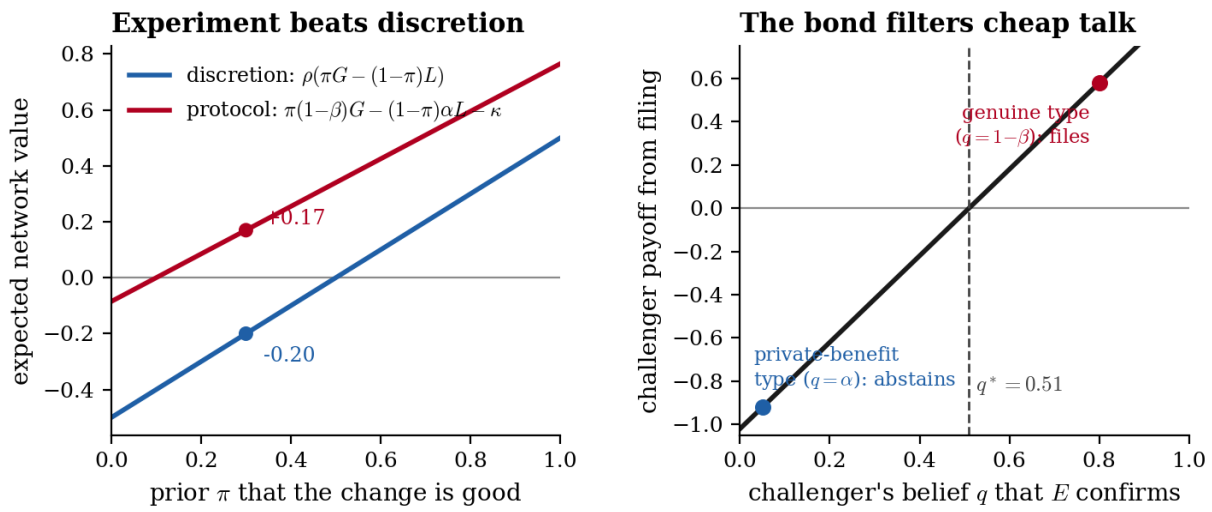

Figure 1: The two central results at the calibration of Appendix A ($\alpha = 0.05$, $1-\beta = 0.80$, $\rho = 0.5$, $\kappa = 0.035$, symmetric bonds, $\varphi = 0.02$). Left (Theorem 2): expected network value under discretion, which adopts at the truth-independent rate ρ, against the protocol, which adopts only when the experiment confirms; the protocol line lies above the discretion line at every prior π, and at $\pi = 0.3$ the network gains $+0.17$ under the protocol versus -0.20 under discretion. Right (Theorem 1): the challenger's payoff from filing crosses zero at the cutoff $q^* = 0.51$, so the genuine type with belief $q = 1-\beta$ files while the private-benefit type with belief $q = \alpha$ abstains.

5.3 Coalition size confers no influence

Theorem 3 (Volume invariance and coalition-resistance)**.** *The protocol's decision D is a function only of the in-order prediction (θ_1, m, δ) and the experiment outcome $\hat{\Delta}$. It does not take the number of complainants as an argument. Consequently:*

1. *(*Volume invariance*.) For any coalition of size n advancing a fixed prediction, $\partial D/\partial n = 0$: enlarging the coalition does not change the decision.*

2. *(*No cheap capture*.) A coalition can change θ only by a prediction the experiment confirms. If the coalition's claim is false, its expected bond loss is at least $(1-\alpha)\, b_C$, regardless of n; if true, adoption raises welfare by Theorem 2.*

Proof. By construction the settlement in Stage 3 reads only (θ_1, m, δ) and $\hat{\Delta}$; the count n enters nowhere, giving volume invariance. A coalition that files a single in-order dispute wins only if the experiment confirms, which under a false claim ($\omega = B$) fails with probability $1-\alpha$, forfeiting b_C; filing n separate disputes on the same false claim forfeits in expectation $n(1-\alpha)b_C$, which is worse. Hence no path changes θ on a false claim except the experiment's size-α error, which is independent

of n and bounded by choice of horizon. A true claim is confirmed with probability $1 - \beta$ and its adoption is welfare-improving. ■

This is the formal antidote to Pathology 1. Under discretion the influence of a complaint rises with the number of voices behind it, which is exactly the advantage that a coordinated minority (Olson, 1965) or an informational cascade (Banerjee, 1992; Bikhchandani et al., 1992) exploits. Under the protocol the channel from volume to decision is severed: the only instrument a coalition has is the quality of a falsifiable prediction, and that quality is adjudicated by an experiment whose verdict is the same whether one participant or ten thousand stand behind the claim. A coalition may file, but it prevails only if the experiment confirms its prediction, which is the outcome the network wanted anyway.

5.4 The defense bond restores symmetry

The authority chooses at Stage 1 between conceding and defending. Let $q_A = \mathbb{P}(\hat{\Delta} < \delta \mid A\text{'s information})$ be the authority's posterior that the status quo will survive the experiment, and let w be the authority's net value of retaining θ_0 rather than adopting θ_1. Defending stakes b_A: with probability q_A the experiment upholds the status quo, the authority retains θ_0 (worth w) and collects the challenger's bond b_C; with probability $1 - q_A$ it is overturned, θ_1 is adopted, and b_A is forfeit. Conceding adopts θ_1 and risks no bond.

Theorem 4 (Symmetric fairness)**.** *With the defense bond $b_A > 0$, the protocol is symmetric in the sense of Section 3: an authority that defends a status quo the experiment overturns forfeits b_A, the same form of loss the challenger bears for a rejected change. Moreover:*

1. *The authority defends iff $q_A \geq q_A^*$, where*

$$q_A^* = \frac{b_A}{b_A + b_C + w} \in (0, 1)$$

is increasing in its own bond b_A and decreasing in the value w it places on retaining θ_0. An authority confident that θ_0 is better strictly prefers to defend, and an authority that privately knows θ_0 is worse concedes, adopting the better parameter without an experiment.

2. *For any target ratio of challenger-to-authority skin in the game, there is a bond pair* (b_C, b_A) *that implements it, so the designer can set how much each side must stake.*

Proof. The forfeiture in Stage 3 is symmetric by construction: C loses b_C on falsification, A loses b_A on confirmation. Relative to conceding, defending yields $q_A(w + b_C) - (1 - q_A)b_A$: with probability q_A the authority keeps θ_0, worth w, and collects b_C, and with probability $1 - q_A$ it forfeits b_A. This is strictly increasing in q_A and crosses zero where $q_A(w + b_C) = (1 - q_A)b_A$, that is at $q_A^* = b_A/(b_A + b_C + w)$, so the defend-iff-$q_A \geq q_A^*$ rule is exact. Differentiating, $\partial q_A^*/\partial b_A > 0$ and $\partial q_A^*/\partial w < 0$. The ratio b_C/b_A is a free design parameter, so any relative stake is implementable. ■

Corollary 1 (Refusing to bond is a costly signal)**.** *Against an in-order, bonded challenge, an authority that declines to post* b_A *triggers the default adoption of* θ_1*. An authority with* $q_A \geq q_A^*$ *strictly prefers to defend and expects to win the challenger's bond, so a refusal to bond reveals* $q_A < q_A^*$*: the authority is not confident the status quo would survive the experiment. Confidence in a parameter is therefore demonstrated by willingness to defend it under symmetric stake, and a one-sided threat that avoids the bond is evidence of the opposite.*

Proof. Immediate from Theorem 4(1): defending is optimal exactly when $q_A \geq q_A^*$, so declining to defend an in-order bonded challenge is optimal only when $q_A < q_A^*$, which is the content of the signal. ■

Corollary 1 is the part of the design that disciplines the side with decision rights. A familiar move in network governance is the one-sided ultimatum: change the parameter, or face a sanction. The protocol does not forbid the authority from preferring the status quo; it requires that the preference be backed, like the challenge it answers, by a stake that the evidence can take away. An authority sure of its parameters loses nothing by this, and gains the challenger's bond and the legitimacy of a settled question. An authority unsure of them is revealed, which is the same service the bond performs on the challenger's side.

Figure 2 collects the two design facts the theorems establish. The left panel plots the region of priors and experiment costs in which settling by experiment beats discretion (Theorem 2); the calibrated dispute of Appendix A sits well inside it, and the region only widens as the stakes grow or the test sharpens. The right panel plots the challenger's filing threshold q^* of Theorem 1 and the

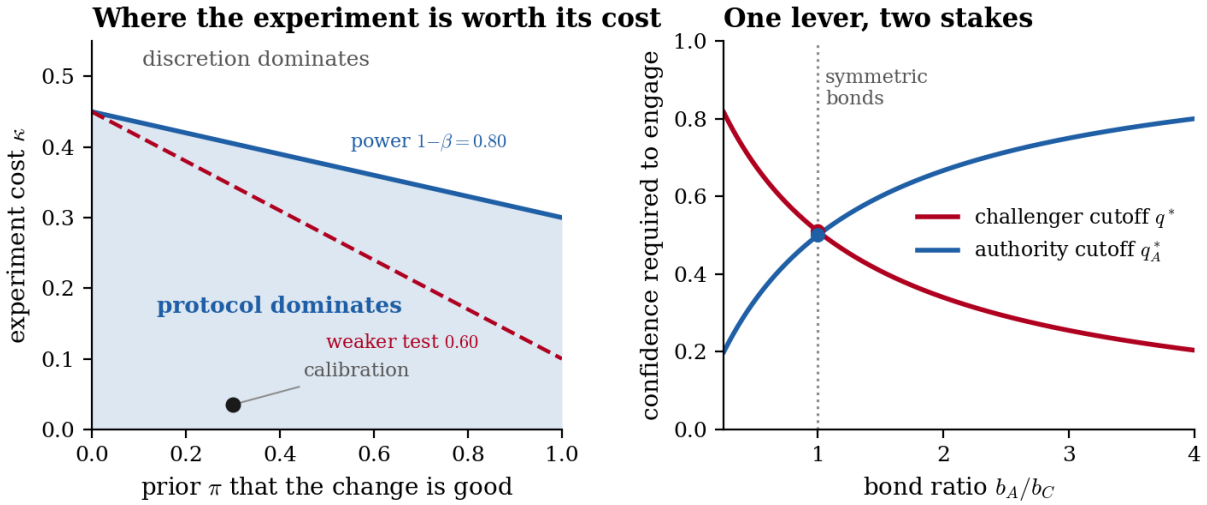

Figure 2: Two design facts from the model. **Left** (Theorem 2): the set of priors π and experiment costs κ for which settling by experiment yields higher expected network value than discretion, namely $\kappa \leq \pi(1-\beta-\rho)G_v + (1-\pi)(\rho-\alpha)L_v$, shaded for the calibrated test ($1-\beta = 0.80$, $\rho = 0.5$, $\alpha = 0.05$, $G_v = L_v = 1$); the boundary recedes as the test weakens (dashed, $1-\beta = 0.60$), and the Appendix A calibration lies well inside the region. **Right** (Theorems 1 and 4): the challenger's filing cutoff $q^* = (b_C + \varphi)/(b_A + b_C)$ and the authority's defense cutoff $q_A^* = b_A/(b_A + b_C + w)$ at $w = 0$, as functions of the bond ratio b_A/b_C. At symmetric bonds the two thresholds meet near $\frac{1}{2}$; raising either side's bond lowers the confidence its counterparty needs to engage and raises its own.

authority's defense threshold q_A^* of Theorem 4 against the bond ratio b_A/b_C: at symmetric bonds each side must be about even-or-better confident to engage, and tilting the ratio is the designer's single lever for reallocating how much confidence each side must stake.

5.5 Why a forecast cannot replace the experiment

Proposition 1 (The forecast route inherits an impossibility)**.** *Replace Stage 2 with a conditional prediction market that forecasts $m(\theta_1)$ and $m(\theta_0)$ and let the decision adopt whichever the market scores higher, as in futarchy (Hanson, 2013). Under a deterministic decision rule the market on the parameter that is not adopted is voided, and by the decision-market impossibility (Chen et al., 2011; Othman and Sandholm, 2010) no strictly proper scoring of that conditional market exists, so traders need not report beliefs truthfully and the forecast need not reflect them. Running both arms, as the protocol does, realizes both metrics, so no conditional voiding occurs and the impossibility does not bind.*

Proof. A conditional market on the outcome under θ_j pays only if θ_j is adopted. With a deterministic

rule that adopts the higher-scored arm, the lower-scored arm is never realized, its market is voided, and positions in it return zero regardless of the truth. Chen et al. (2011) show this destroys strict properness for general deterministic decision rules: a trader's optimal report in the voided market is unconstrained, so prices carry no guaranteed information, generalizing the earlier max-rule impossibility of Othman and Sandholm (2010). The protocol's experiment realizes $m(\theta_0)$ and $m(\theta_1)$ on the two arms with certainty; both metrics are observed, nothing is voided, and the verdict rests on realized outcomes rather than on the price of an unrealized one. ■

This is the formal reason the protocol runs the control rather than predicting it. Futarchy is attractive because markets aggregate dispersed information cheaply, but the very conditioning that makes a decision market a *decision* rule is what voids the losing market and undermines its incentives. An experiment is more expensive than a market quote, by the cost κ, and Theorem 2 is precisely the statement that the network should pay that cost when the parameter matters enough.

5.6 Who pays for the experiment

Proposition 2 (Funding)**.** *If the authority or foundation pre-funds the experiment from the emission budget it already controls, the protocol is individually rational for an honest challenger at any effect size δ for which $q \geq q^*$, and the funder's outlay per in-order dispute is bounded by κ. Whenever the bound of Theorem 2 holds, κ is less than the expected value the experiment creates, so funding the experiment is itself value-increasing for the funder.*

Proof. With the experiment pre-funded, the challenger's payoff is the bond game of Theorem 1 without an added share of κ, so the filing cutoff is unchanged and any challenger with $q \geq q^*$ participates. The funder pays κ per in-order dispute and, by Theorem 2, receives an expected value increase exceeding κ under the stated bound, so funding is profitable for the funder. ■

Proposition 2 answers a practical objection. The party with the budget to run test deployments is typically the authority itself, the holder of the emission or treasury controls. Far from a burden, funding the experiment is the cheapest way for that party to convert an unresolved and reputationally costly dispute into a settled question, and it removes the wealth barrier that a large bond would otherwise place in front of a poor but correct challenger.

6 Design choices and robustness

Metric gaming. An experiment is only as good as its metric, and any single metric invites the Goodhart response of optimizing the proxy rather than the target. The protocol confines disputes to a metric set $\mathcal{M}$ agreed before any specific dispute, audited by the neutral operator, and chosen so that the metric is costly to game within the horizon H. Where a parameter plausibly trades off two objectives, the prediction must name both, and confirmation requires no significant regression on the unnamed-but-protected metric. This is the experimental-design counterpart of pre-registration in Mellers et al. (2001).

Collusion between challenger and authority. If forfeited bonds were always transferred to the winner, a challenger and an authority could agree to file and concede, adopting θ_1 with no experiment and no loss. Two features remove the incentive. First, an in-order challenge that the authority concedes still adopts θ_1 only if it passes a minimal pre-screen, so concession is not a free bypass of evidence for a consequential change. Second, in the burn variant a forfeited bond is destroyed rather than paid to the counterparty, so there is no pot for colluders to share; the bond is a deposit against being wrong, not a prize.

Bond calibration and access. A bond large enough to deter cheap talk may exclude a correct challenger of modest means. Three measures reconcile the two. Bonds can be set as a fraction of the disputed parameter's value at stake rather than as a flat sum, so they scale with what is being contested. Challenges can be crowdfunded under a dominant-assurance structure (Tabarrok, 1998), in which contributors are refunded with a bonus if the challenge fails to reach its bond, which makes funding a correct challenge individually rational. And the funder of last resort for the experiment, by Proposition 2, is the authority, which lowers the total cost a challenger must cover to the bond alone.

Experiment validity. The experiment must be a valid test of the parameter, which requires that its unit of treatment be the whole market and not a fraction of it. A companion paper (Anonymous, 2026) shows, for the registration-cost application below, that reserving a fraction of seats for the alternative parameter on a live deployment violates the stable-unit-treatment-value assumption,

because behavior under one parameter setting leaks into the arm running the other. The valid designs are a parallel deployment running each parameter for its whole population, or a test network on which the alternative parameter governs all entry. The protocol inherits this requirement: the pre-registered E of Stage 2 is a market-level experiment, not a within-market carve-out.

Repeated play. Across many disputes the protocol accumulates a public record of predictions and outcomes, which prices reputational capital into each side's willingness to bond and tightens the separation of Theorem 1 over time. An authority that defends and wins repeatedly builds standing; a challenger with a record of confirmed predictions can credibly file at lower bond. The one-shot analysis here is therefore a lower bound on the protocol's selectivity.

Beyond the binary state. The two-state world, the change is either good or bad, is a presentational simplification, not a load-bearing assumption. With a continuous effect $\Delta_{\text{true}} \in \mathbb{R}$ and an experiment that tests $H_0 : \Delta_{\text{true}} \leq 0$ against the filed prediction $\Delta_{\text{true}} \geq \delta$ for a pre-registered effect size $\delta > 0$, every result survives with the obvious reading. The filing cutoff of Theorem 1 becomes a cutoff in the challenger's posterior mean of Δ_{true}; the welfare comparison of Theorem 2 holds verbatim with α and $1 - \beta$ reinterpreted as the test's size and its power at the alternative δ; and coalition-resistance (Theorem 3) and the symmetry of the defense bond (Theorem 4) are unchanged, since neither uses the cardinality of the state space. The binary model is the special case $\Delta_{\text{true}} \in \{0, \delta\}$, and we adopt it only because it lets the separating equilibrium be displayed in closed form.

7 Application: a registration-cost dispute

We close with the dispute that motivated the protocol, stated in its terms. A reward network prices entry through an adaptive registration burn, and a group of participants objects that the burn is too high and should be reduced to a near-zero flat level. The status quo is θ_0, the deployed adaptive schedule; the proposed alternative is θ_1, a flat near-zero entry price. The companion paper (Anonymous, 2026) predicts, from an impossibility theorem, that lowering the entry price toward zero collapses strategy diversity, because free entry makes cloning the leading strategy profitable, so the network converges to copies of a single entrant rather than to a population of distinct ones.

That prediction is exactly the falsifiable object the protocol needs. The metric m is strategy diversity, operationalized as the entropy of the strategy distribution and the rate of genuinely new strategies per epoch, with debate quality as the protected secondary metric. The challenger who believes free entry improves the network predicts $\mathbb{E}[m(\theta_1)] \geq \mathbb{E}[m(\theta_0)] + \delta$, names δ, and bonds it. The authority, if it believes the companion paper, defends and bonds the opposite. The experiment is a pre-registered parallel deployment, or a test network, running flat near-zero entry against the adaptive burn for a fixed horizon, measuring diversity and net new strategies under a neutral judge, with cross-play of the end-state strategies to control for population differences. The verdict is the measured sign of $\hat{\Delta}$.

Two features of this application are worth stating plainly. First, the natural worry that the free-entry arm will simply copy the paid arm's strategies is not a confound but the dependent variable: convergence to copies is the predicted outcome under free entry, so the experiment should measure diversity rather than absolute quality, and copying *is* the result. Second, the cheap subnets sometimes invoked as a ready-made demonstration of free entry are not one: they run different tasks under different reward rules, several winner-take-all rather than graded, so their outcomes are silent on what free entry would do to a graded-reward market. That is precisely why the question needs the controlled, bonded experiment rather than an appeal to confounded cross-network outcomes, the configuration that Theorem 1 is designed to resolve: the protocol asks the objection to bond a falsifiable prediction and submit it to the experiment that would settle it, and Corollary 1 asks the same of the authority on the other side.

8 Conclusion

Governance of a permissionless network fails in a particular way: complaint is free, so its volume is uninformative, and the authority is unbonded, so its decisions are undisciplined. The Bonded Falsification Protocol repairs both failures with one mechanism. It admits a dispute only when the challenger stakes a bond on a falsifiable prediction, it requires the authority to stake a bond to defend the status quo, and it settles the dispute by running the experiment that the prediction is about, rather than by counting voices or pricing a forecast. The bond separates belief from noise; the experiment converts a coin flip into an informative test; volume invariance denies a coalition any

advantage from its size; and the symmetric stake turns a one-sided ultimatum into a two-sided wager that a confident authority is glad to take. The combination is what is new: no prior mechanism filters cheap talk, states a falsifiable prediction, generates the counterfactual by experiment, and binds the party with decision rights, all at once. A dispute conducted this way ends not when one side tires or the other prevails by force, but when the evidence both sides agreed to produce has been produced. That is the standard a network that pays for measured performance should hold its own governance to.

References

Anonymous. The entry trilemma: Costly registration as Sybil-resistance in graded-reward networks, 2026. Companion paper, authors withheld for review.

Abhijit V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992. doi: 10.2307/2118364.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. doi: 10.1086/261849.

Yiling Chen, Ian A. Kash, Mike Ruberry, and Victor Shnayder. Decision markets with good incentives. In *Internet and Network Economics (WINE 2011)*, volume 7090 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2011. doi: 10.1007/978-3-642-25510-6_7.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. doi: 10.2307/1913390.

Robin Hanson. Shall we vote on values, but bet on beliefs? *Journal of Political Philosophy*, 21(2): 151–178, 2013. doi: 10.1111/jopp.12008.

Albert O. Hirschman. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, Cambridge, MA, 1970.

Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91, 1979. doi: 10.2307/3003320.

Irving L. Janis. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 1972.

Michael C. Jensen and William H. Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360, 1976. doi: 10.1016/0304-405X(76)90026-X.

Harry Kalodner, Steven Goldfeder, Xiaoqi Chen, S. Matthew Weinberg, and Edward W. Felten. Arbitrum: Scalable, private smart contracts. In *27th USENIX Security Symposium*, pages 1353–1370. USENIX Association, 2018.

Clément Lesaege, Federico Ast, and William George. Kleros: Short paper v1.0.7. Kleros whitepaper, 2019. Available at `https://kleros.io/assets/whitepaper.pdf`.

Barbara Mellers, Ralph Hertwig, and Daniel Kahneman. Do frequency representations eliminate conjunction effects? an exercise in adversarial collaboration. *Psychological Science*, 12(4):269–275, 2001. doi: 10.1111/1467-9280.00350.

Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, Cambridge, MA, 1965.

Abraham Othman and Tuomas Sandholm. Decision rules and decision markets. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 625–632, 2010.

Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973. doi: 10.2307/1882010.

Alexander Tabarrok. The private provision of public goods via dominant assurance contracts. *Public Choice*, 96(3–4):345–362, 1998. doi: 10.1023/A:1004957109535.

Jason Teutsch and Christian Reitwießner. A scalable verification solution for blockchains. arXiv:1908.04756, 2019. Available at `https://arxiv.org/abs/1908.04756`.

A A numerical bond calibration

Suppose the experiment has size $\alpha = 0.05$ and power $1 - \beta = 0.80$, the prior is $\pi = 0.3$, and the stakes are $G_v = L_v = 1$ (normalizing value). Discretion with uninformative volume $\rho = 0.5$ yields $\mathbb{E}[V_{\text{disc}}] = 0.5(0.3 - 0.7) = -0.20$: under a loud room that adopts changes half the time regardless of merit, the network loses value, because bad changes outnumber good ones at this prior. The protocol yields $\mathbb{E}[V_{\text{BFP}}] = 0.3(0.8) - 0.7(0.05) - \kappa = 0.24 - 0.035 - \kappa = 0.205 - \kappa$, positive for any $\kappa < 0.205$ and better than discretion for any $\kappa < 0.405$. With symmetric bonds and filing effort $\varphi = 0.02\, b$, the filing cutoff is $q^* = 0.5 + 0.01 = 0.51$, so a challenger files only when more likely than not to be confirmed, while an uninformed objector with $q = \alpha = 0.05$ stays out by a wide margin. The defense cutoff q_A^* is set by the same bonds, and choosing $b_C = b_A$ gives each side equal skin in the game.

B Protocol pseudocode

1. Pre-agree the metric set $\mathcal{M}$, bond sizes (b_C, b_A), experiment horizon H, estimator, and decision threshold δ family.

2. Challenger files (θ_1, m, δ) and posts b_C; reject if not in order.

3. Authority within the window: concede and adopt θ_1 (subject to pre-screen); or post b_A and proceed; or do neither and let θ_1 adopt by default.

4. Run the pre-registered two-arm experiment for horizon H; compute $\hat{\Delta}$ under the fixed plan.

5. If $\hat{\Delta} \geq \delta$ adopt θ_1 and forfeit b_A; else keep θ_0 and forfeit b_C. Forfeited bonds are burned (collusion-proof variant) or awarded to the winner.