From Pluralistic Ignorance to Common Knowledge with Social Assurance Contracts

Matthew Cashman
MIT Sloan School of Management
cashman@mit.edu

Abstract

Societies and organizations often fail to surface latent consensus because individuals fear social censure. A manager might suspect a silent majority would offer a criticism, support a change, report a risk, or endorse a policy—if only it were safe. Likewise, individuals with beliefs they think are rare and controversial might stay quiet for fear of consequences at work or an online mob. In both cases pluralistic ignorance produces a public discourse misaligned with privately-held beliefs. Social assurance contracts unlock latent consensus, making the public discussion more accurately reflect the underlying distribution of actual beliefs. They are akin to an open letter that publishes only when a stated threshold number of private signatures is reached. If it is not reached, nothing is revealed and no one is exposed. Whereas a single hand raised in dissent might get cut off, a thousand can be raised safely together. I build a formal model and derive rules for choosing the threshold. The mechanism (i) induces participation from those willing to speak if assured of company, resolving the core coordination problem in pluralistic ignorance; (ii) makes the threshold a transparent policy lever—sponsors can maximize success, maximize public-coalition revelation, or hit a desired success probability; and (iii) turns success into information: meeting the threshold publicly reveals hidden agreement and can widen the range of views that can be expressed in public. I consider robustness to mistrust, organized opposition, and network structure, and outline low-trust implementations like cryptographic escrow. Applications include employee voice, safety and compliance, whistleblowing, and civic expression.

Keywords: Assurance contracts, disclosure design, public goods, coordination, pluralistic ignorance, mechanism design

1 Introduction

Public silence on an issue is often misread as agreement—or, at least, the lack of a strong objection. However, in organizations, workplaces, and online communities, the missing public statement does not necessarily mean dissent is absent; it often means no one wants to be first to object. When fear of censure begets silence, observers infer there is support for the dominant narrative even when privately held beliefs are substantially different. This wedge between private belief and public expression is the core of pluralistic ignorance (Prentice and Miller 1996). Digital environments can intensify the problem: visible discourse may reflect a small set of

highly active users rather than majority sentiment (Kim et al. 2021, McClain et al. 2021, Milli et al. 2025). Experimental evidence confirms that inefficient norms are less likely to persist when credible information about peers' views becomes available (Smerdon et al. 2020). The institutional problem is therefore not measurement alone; it is coordinated public revelation.

This paper studies a simple mechanism designed to solve that problem: a *social assurance contract*. The mechanism adapts the logic of assurance (provision-point) mechanisms from the economic domain to the social, making them work for speech rather than funding. In a social assurance contract, individuals privately commit to publicly endorse a statement only when a pre-set safety threshold of a certain number of signatures is reached. Until then, commitments remain hidden from both the public and other potential signers. At a university, it might look something like this:

> We, the undersigned believe [controversial belief] and think the University should take the following steps [1, 2, 3, . . .]. Signatures on this letter will become public *only when at least [200] faculty have signed.* Until then, no one can see who has signed. If the threshold is not met, the letter and its signatures are destroyed.

By conditioning disclosure on a critical mass of support, the mechanism reduces first-mover exposure: one hand raised alone may be cut off, but a thousand may be safely raised together. The mechanism differs in important ways from a secret vote. A secret ballot can reveal that, say, 40% support a controversial position, but it does not reveal *which* 40%—and so they cannot stand together in public. A social assurance contract automatically produces that coalition when the safety threshold is met. Social assurance contracts are also importantly different from petitions or open letters in that, without secrecy, open letters will only be signed by those who can bear the costs of likely attack by adversaries alone—and that may not be enough people to surface true latent support.

These contrasts serve to clarify our scope. If a population or formal institution can map an aggregate vote directly into action, a secret ballot is usually the better tool. If an organization only needs to estimate latent support, an anonymous survey is usually simpler. Social assurance contracts belong to the intermediate class of settings where aggregate support alone is insufficient and the scarce object is a publicly revealed coalition: a set of individuals who can sign, file, or organize together without each being isolated and silenced. Section 6 formalizes this partition by comparing the three mechanisms on the same population and showing that the assurance contract is uniquely effective in an intermediate parameter region where both the survey and the open petition fail.

Corporate whistleblowing is an especially clear application. Employees are the single most common source of corporate fraud detection, yet the gap between private awareness and public reporting is large and persistent (Dyck et al. 2010). The Wells Fargo cross-selling scandal illustrates the coordination failure. Between 2002 and 2016, thousands of employees in the bank's Community Banking division knew that millions of unauthorized accounts were being opened under systemic sales pressure. Over that period, 885 employees individually contacted the internal ethics hotline—and were systematically retaliated against, often fired within days (Independent Directors of the Board of Wells Fargo & Company 2017). Each reporter acted alone, was identified, and was punished; the public appearance of compliance was sustained by the silence of

the vast majority. Regulatory action did not arrive until eleven years after the first documented complaints, at an eventual cost exceeding $5 billion in penalties (U.S. Department of Justice 2020). An escrowed, threshold-triggered collective report—a social assurance contract written on whistleblowing—would let employees submit complaints privately and reveal identities only once enough co-reporters exist to make retaliation against any individual impractical. Section 4.6 calibrates the model to this setting and quantifies the coordination failure.

Similar first-mover problems arise in labor organizing, where isolated early supporters of a union campaign are easier to identify and pressure than a large group moving together (NLRB 2026a,b); when dispersed shareholders consider backing an activist filing; and when faculty or staff consider signing a letter addressing a sensitive policy issue. In each case, what is needed is not merely a better estimate of latent support. What is needed is a way to reveal enough named supporters at once to make subsequent public action credible and safe. Recently, several implementations have emerged that now offer threshold-triggered conditional commitments and synchronized action. Examples include general-purpose platforms such as Spartacus.app, founded 2024, allegation escrows such as Callisto, collective pledges for open-science such as the Open Code Pledge (Smout et al. 2021), and Wait Until the 8th, a contract that allows parents to coordinate to delay smartphones use until the 8th grade (Wait Until the 8th 2026). Participation nevertheless depends on trust in confidentiality and trigger integrity, and I make those trust assumptions explicit and return to implementation risks in the appendices.

The core strategic problem is therefore not whether private support exists, but whether privately supportive individuals can coordinate on public expression. Some people place more value on publicly endorsing the statement than others, and some are more vulnerable to the personal or reputational costs of being identified with it. The mechanism's central force is safety in numbers: when enough others are expected to sign, public expression becomes safer, making endorsement attractive to people who would otherwise remain silent.

The paper makes three contributions. First, it develops a baseline model of privately supportive individuals who differ in both their value from public expression and their vulnerability to public exposure. Expected exposure cost falls as anticipated participation rises, and equilibrium is therefore a fixed point in anticipated participation with type-specific cutoffs. Because safety in numbers creates strategic complementarities, the participation map can admit multiple equilibria. The paper then separates two design questions: a conditional threshold-design problem evaluated at the participatory equilibrium, and a reduced-form ex ante extension that weights that objective by the probability of reaching that equilibrium. Second, a calibrated numerical exercise shows that the conditional design objective is hump-shaped and maximized at an interior threshold, with the high-vulnerability type differentially drawn into participation by the safety channel. Endogenous safety pulls the conditional optimum lower than models with fixed exposure cost, because a moderate coalition already provides sufficient protection. When coordination risk is priced in, the ex ante optimum shifts weakly further downward. A counterfactual calibration anchored to corporate whistleblowing—using the Wells Fargo cross-selling scandal and aggregate retaliation data—shows that the model's qualitative features survive in a high-retaliation, low-prevalence environment and quantifies the costs of the coordination failure the mechanism is designed to solve. A formal mechanism comparison then shows that the assurance contract is uniquely effective in an intermediate parameter region where anonymous surveys fail to produce coordination

and open petition cascades stall. Third, because outside observers hold uncertain beliefs about support prevalence, the same baseline delivers a belief-updating result: success is more likely when support is more prevalent, so observing success raises posterior beliefs on average and can, under a reduced-form mapping from beliefs to expression costs, widen the range of views that are publicly acceptable to express (the Overton window). A finite-N benchmark with exact pivotality and posterior updating over latent support is developed in an appendix for settings where individual material stakes are large, as in small groups.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 presents the mechanism and incentives. Section 4 develops the baseline model, including multiplicity, equilibrium selection, and the ex ante design layer. Section 5 studies threshold design and opposition-imposed durability. Section 6 compares the assurance contract formally with anonymous surveys and open petitions. Section 7 connects successful campaigns to public belief updating. Section 8 discusses scope and concludes. Appendices provide proofs, a finite-N benchmark with exact pivotality, implementation analysis, and extensions.

2 Related Literature

Social assurance contracts sit at the intersection of step-level public good provision, coordination under strategic uncertainty, and models of suppressed expression. This paper adds to the literature by treating *disclosure itself* as the public good in a provision-point environment with reputational inference, and by connecting equilibrium participation to belief updating about what is socially acceptable to say.

This paper builds on classic models of step-level public goods and threshold contribution games (Dybvig and Spatt 1983, Palfrey and Rosenthal 1984). Provision-point mechanisms with refunds can mitigate free riding by protecting contributors when the threshold is not met (Bagnoli and Lipman 1989), and experimental work studies how design features affect provision outcomes (Croson and Marks 2000). Dominant assurance contracts introduce refund bonuses to strengthen provision incentives (Tabarrok 1998, Zubrickas 2014), and laboratory evidence reports improved provision performance (Cason and Zubrickas 2017).

The mechanism studied here adapts this logic to *expressive* contributions: agents commit to public endorsement rather than money, and payoffs are reputational and informational. A particularly close analogue is the "information escrow" proposal in law and economics (Ayres and Unkovic 2012), in which allegations of wrongdoing are held confidentially and released under prespecified corroboration conditions. Social assurance contracts generalize this escrow logic to broader contexts of conditional collective voice. Dating applications are another analog. In common "double opt-in" implementations, only if Alice indicates interest in Bob and Bob in Alice do they find out about their mutual interest (Rad et al. 2017). This can be seen as a two-party assurance contract written on the fly (since matches are possible between any two users).

More broadly, the model relates to coordination frameworks with strategic uncertainty and complementarities, where multiple self-fulfilling outcomes can arise and institutional devices can coordinate expectations (Carlsson and van Damme 1993, Morris and Shin 2002). A particularly relevant theoretical framework is Ali and Bénabou (2020), who formalize the tradeoff between publicity (which supports incentives and compliance) and privacy (which supports information revelation about true preferences). The social assurance contract studied here navigates this tradeoff by keeping commitments private until a threshold is met, then

making them public—combining the informational advantages of privacy during the collection phase with the incentive and coordination advantages of publicity upon success. In our setting, the assurance threshold provides a credible focal point for a simultaneous deviation from silence.

Models of informational cascades and herding show how early actions can lock groups into outcomes that ignore dispersed private information (Bikhchandani et al. 1992, Raafat et al. 2009). More generally, social-learning cascades can block further learning from private information (Bikhchandani et al. 2024). By analogy, a cascade of silence can impede learning about others' views when no one speaks. By enabling coordinated, threshold-triggered disclosure, a social assurance contract is designed to interrupt such cascades by creating a public signal about latent support revealed through successful coalition formation. Recent field evidence from Hong Kong's antiauthoritarian movement further shows that protest participation responds strategically to belief updates about others' likely turnout (Cantoni et al. 2019), closely matching the same coordination logic.

The mechanism also connects to social-psychological accounts of conformity and suppressed expression, including preference falsification (Kuran 1997), the spiral of silence (Noelle-Neumann 1974), groupthink (Esser 1998), and social desirability bias (Krumpal 2013). Historical episodes suggest the durability of these dynamics, from Zhao Gao's "deer as horse" loyalty test in the Qin court—in which courtiers were forced to publicly endorse a known falsehood, revealing their willingness to comply and punishing deviation through the same reputational channel that sustains pluralistic ignorance in the baseline model—to sanctioning dynamics around public exclusion in classical Athens (Lewis 2010, Forsdyke 2009). Pluralistic ignorance has been documented across settings ranging from college campuses to workplaces and political opinion climates (Prentice and Miller 1993, Halbesleben et al. 2007, Ahler 2014). Laboratory evidence suggests that inefficient norms can persist primarily because individuals are uncertain about others' preferences, rather than because the "bad" norm is inherently stable (Smerdon et al. 2020). Experimental work by Bursztyn et al. (2020) demonstrates that misperceived social norms can suppress expression of privately held views, while Braghieri (2024) formalizes how social-image concerns generate self-censorship in equilibrium. Complementary causal evidence comes from Heese and Pérez-Cavazos (2021), who show that a 10% increase in state unemployment-insurance benefits—which reduces the cost of employer retaliation—raises OSHA employee complaints by 13.8%, with the effect concentrated in firms where employees are most vulnerable. This directly supports the model's core mechanism: reducing exposure cost increases reporting. Social assurance contracts operationalize this insight by converting dispersed private willingness to speak into a single, publicly interpretable success event.

A related literature in survey methodology develops indirect questioning techniques—such as randomized response, list experiments, and crosswise methods—to elicit truthful reports on sensitive topics while protecting respondents (Warner 1965, Glynn 2013, Yu et al. 2008). These approaches are designed to measure private beliefs without making them common knowledge among participants. If the designer's objective is simply to estimate latent support, then an anonymous survey is often the right benchmark and may dominate a social assurance contract because it minimizes exposure while still producing an aggregate statistic. What it cannot do is identify a coalition prepared to stand together in public, so it does not directly alter the incentives to speak. Relatedly, secret ballots reveal aggregate support without revealing the set of individuals prepared

to express a view publicly. They are the natural benchmark when a formal institution can map an aggregate vote directly into action. Social assurance contracts instead condition *public identity revelation* on a threshold event, creating coordinated common knowledge and associated reputational incentives. They therefore add something distinct only when the relevant output is not merely information about support, but the public emergence of a coalition large enough to speak or act together.

The Overton window describes the set of policies and ideas that public actors can endorse without being viewed as beyond the pale (Lehman 2010). Related concepts include Hallin's spheres of media discourse (Hallin 1989) and the Scandinavian "opinion corridor" (Simons 2021). These concepts are used informally in the cited literatures; the cost-threshold formalization in Section 7 is my own. Formal work on opinion dynamics and preference falsification studies how divergence between private beliefs and public statements can emerge and persist under specific assumptions (Duffy and Lafky 2018, Lütz and Wardil 2024, León-Medina et al. 2020, Apolte and Müller 2022, Fernández-Duque 2022). Here, the Overton window itself is treated as an expression-cost frontier that can respond to public belief updating, and I show how threshold-triggered disclosure raises posterior beliefs after success and can expand the window.

3 Mechanism

I model a social assurance contract as a simultaneous one-shot game with no information flow between players. (Real campaigns unfold over time, and agents may condition on elapsed time, social signals, or recruiter behavior. The static model abstracts from strategic delay and sequential arrival; Online Appendix E.2 sketches an architect-chosen intermediate-disclosure extension. Separately, beliefs are treated as exogenous, although conformity and social-influence research indicates that perceived norms can shape both public and private responses; see Capuano and Chekroun 2024.) The contract is designed by a contract *architect*, perhaps an employee representative or a compliance advocate, and administered by a contract *administrator*. The same person may fulfill both roles, and the administrator may also be a computer system. Let $\mathcal{P}$ be the overall population of individuals who could potentially be aware of or affected by the contract's statement. Within this broader population, let N denote the size of the relevant *eligible population*: the known set of individuals who could in principle sign the contract (for example, employees in a workplace, faculty in a school, shareholders meeting an eligibility rule, or users on a platform). The game is modeled among these N eligible individuals, indexed by $i = 1, \ldots, N$. A consolidated notation reference is provided in Online Appendix D.

3.1 Contract Structure

The contract is defined by a specific statement to be endorsed (which could be a proposition or normative claim, representing an underlying belief $\mathbf{b}_s$) and an assurance threshold T. This threshold T is a positive integer representing the minimum number of endorsements required for the contract's list of endorsers to become public; for success to be possible, T must satisfy $1 \leq T \leq N$, and I maintain the standing assumption $1 \leq T < N$ (excluding unanimity). Define the normalized threshold share $\tau \equiv T/N$, so $\tau \in [1/N, 1)$. When N is fixed, T and τ are one-to-one representations of the same threshold choice. The contract structure is as follows:

1. Eligible individuals $i = 1, \ldots, N$ decide simultaneously their action $a_i \in \{0, 1\}$, where $a_i = 1$ denotes choosing to endorse the contract, Sign, and $a_i = 0$ denotes Not Sign.

2. Let $M = \sum_{j=1}^{N} a_j$ be the total number of endorsers. If $M \geq T$, the statement is published along with the list of all M endorsers. The contract is then considered "successful."

3. If $M < T$, the identities of the endorsers remain confidential from each other, the public, and potentially the contract administrator. No statement is released, and the contract is considered "failed."

This structure mirrors the core logic of assurance contracts (Bagnoli and Lipman 1989, Tabarrok 1998). The model here introduces two important modifications for the expressive setting: first, there is complete secrecy about who has signed the contract; second, participation payoffs include psychological terms that are state-dependent. Reputational esteem materializes on success for signers, sanctions bite when expression is public, and non-signers may face stigma when a successful list becomes public. Treating these as utility terms rather than monetary transfers matters for both design and welfare. The equilibrium characterization continues to go through when these terms are replaced by equivalent pecuniary payments that enter additively; however, state-dependence and curvature (e.g., convex sanction costs) break simple subsidy equivalence and can shift the optimal threshold T.

In the baseline model, only supportive agents with $z_i = 1$ may rationally choose Sign; non-supportive agents with $z_i = 0$ choose Not Sign. The strategic question is therefore which supportive agents find it worthwhile to sign given their expressive benefit e_i, vulnerability α_i, and beliefs about others' participation.

We assume agents in the eligible population know π: they know how prevalent private support is among their peers, even if they do not know who is supportive. Outside observers—the broader public, institutional audiences, or media—may not know π and instead hold a prior $\pi \sim G_\pi$. When the contract succeeds and $M \geq T$ endorsers are revealed, outside observers update their beliefs about π. This asymmetry is what allows a successful social assurance contract to change public discourse: insiders coordinate on common knowledge of π, while outsiders learn about π from the public success event. The equilibrium analysis in Section 4 conditions on π known to agents; the Overton-window analysis in Section 7 adds the outside observer's learning problem.

3.2 The Agent's Decision Problem: Payoffs

The components of utility are:

Support and Common Success Value ($z_i, \bar{v}$)—A share π of the eligible population is privately supportive. Supportive agents have $z_i = 1$ and non-supportive agents have $z_i = 0$; in the baseline model non-supportive agents do not sign. If the contract succeeds, a supportive agent receives a common material value $\bar{v} > 0$ from the coordinated public revelation itself. In the baseline large-N model this common value does not drive the marginal signing decision; the finite-N benchmark in Appendix C reintroduces exact pivotality, which is where common value matters strategically.

Expressive Benefit (e_i)—If a supportive agent signs and the contract succeeds, they receive an idiosyncratic expressive or reputational benefit e_i. This captures private value from being publicly associated with the

statement once it is revealed: integrity, esteem, solidarity, or reputational credit for standing publicly with the coalition.

Vulnerability and Exposure Cost (α_i, $c_i(q)$)—If a supportive agent signs and the contract succeeds, they also incur exposure cost

$$c_i(q) = \alpha_i g(q), \qquad g'(q) < 0,$$

where q is anticipated participation and α_i is the agent's vulnerability to public exposure. The function $g(\cdot)$ captures the central safety-in-numbers channel: when more others are expected to sign, public expression is safer and the exposure cost falls. In the baseline model developed below, vulnerability takes one of two values, $\alpha_i \in \{\alpha_L, \alpha_H\}$.

The relevant primitive at the moment of signing is anticipated participation q, not the realized coalition share M/N. Agents choose whether to sign before M is realized, so the cost term $\alpha_i g(q)$ should be read as a reduced-form ex ante representation of expected exposure under anticipated coalition size. At equilibrium, anticipated and realized participation coincide in expectation; out of equilibrium they can differ, which is precisely what creates the coordination problem.

Warm-Glow and Signing Friction (w, k)—Signing can also yield a signing-specific expressive payoff: a common warm-glow benefit from participating in a campaign one expects others to join, offset by a common signing friction $k \geq 0$ (time, hassle, or worry about exposure). I write the net signing utility as

$$\eta(q) = \bar{w}q - k, \qquad \bar{w} > 0,\ k > 0.$$

This specification makes signing an empty contract unattractive ($\eta(0) = -k < 0$) while allowing the warm-glow from joining a viable coalition to remain positive at higher participation levels. The interpretation is close to the participation bonus logic in dominant assurance contracts (Tabarrok 1998): signing is rewarding when others are expected to stand with you, but not when the contract is believed to be futile. These signing-specific terms are interpreted within the supportive subpopulation: non-supportive agents with $z_i = 0$ choose not to sign and are not assumed to receive positive expressive utility from endorsing a statement they oppose.

Stigma for Not Signing ($s(\tau)$)—If the contract succeeds and the list becomes public, a supportive non-signer may incur a stigma cost $s(\tau) \geq 0$. This captures reputational or social penalties from being publicly absent when a coalition is revealed. I treat this component as common in the main model to keep the source of heterogeneity focused on expressive benefit and vulnerability. The baseline therefore assumes public eligibility: after a successful contract, observers can identify who was eligible but did not sign. In settings where eligibility is private or only partially inferable, abstention stigma is attenuated.

The payoffs for agent i are summarized in Table 1.

Table 1 characterizes the signing decision for supportive agents ($z_i = 1$). Non-supportive agents ($z_i = 0$) choose $a_i = 0$ in the baseline model and are omitted from the table.

The baseline model does not compress private payoff heterogeneity into a fixed scalar type. Because exposure cost falls with anticipated coalition size, the relevant private state is the pair (e_i, α_i) interacting with

anticipated participation q. For agent i with action $a_i \in \{0, 1\}$ and total signatures M, write utility as:

$$U_i(a_i, M; q) \;=\; a_i\eta(q) + \mathbf{1}\{M \geq T\}\Big(\bar{v} + a_i\big(e_i - \alpha_i g(q)\big) - (1 - a_i)s(\tau)\Big).$$

Here $\mathbf{1}\{M \geq T\}$ is an indicator function equal to 1 when success occurs ($M \geq T$) and 0 otherwise, so success-contingent payoff terms switch on only in the success state.

Table 1: Payoffs for Agent i under a social assurance contract

Action a_i	**Contract Fails** ($M < T$)	**Contract Succeeds** ($M \geq T$)
$a_i = 0$ (Not Sign)	0	$\bar{v} - s(\tau)$
$a_i = 1$ (Sign)	$\eta(q)$	$\bar{v} + e_i - \alpha_i g(q) + \eta(q)$

I assume agents are rational expected utility maximizers. The structure of the game, the size N of the eligible population, and the payoff primitives and type distributions (except for individual realizations of e_i and α_i among supportive agents) are common knowledge.

If a supportive agent chooses $a_i = 1$ (Sign) and the contract succeeds, their utility is $\bar{v} + e_i - \alpha_i g(q) + \eta(q)$. If it fails, their utility is $\eta(q)$. If the same agent chooses $a_i = 0$ (Not Sign) and the contract succeeds (due to others' actions), their utility is $\bar{v} - s(\tau)$. If it fails, their utility is 0.

An important condition for choosing $a_i = 1$ emerges from comparing payoffs when success is assured. In that state, signing is preferred whenever

$$e_i + s(\tau) - \alpha_i g(q) + \eta(q) \geq 0.$$

This already reveals the baseline mechanism: higher anticipated participation lowers $g(q)$ and therefore makes the inequality easier to satisfy, especially for the more vulnerable agents.

With the contract structure and payoff primitives in hand, the paper studies two tractable decompositions of the same underlying environment. A fully general model would contain both endogenous safety in numbers (coalition size changes exposure cost) and exact pivotality under uncertain support (each agent's signing decision can materially change whether the threshold is met). The baseline model keeps endogenous safety in numbers but suppresses exact pivotality. This is the paper's conceptual center, suited to large organizations or public campaigns where coalition safety is the dominant force. Appendix C contains a Finite-N benchmark that restores exact pivotality and posterior updating over latent support θ, but holds exposure cost fixed. This analysis is suited to small committees or high-stakes settings where individual material stakes are first-order. The baseline's relevant net payoff varies with anticipated participation through $e_i - \alpha_i g(q)$; the benchmark's fixed reputational type x_i is the reduced-form analogue when that safety feedback is shut down. The two models are purposeful decompositions of the same environment, not competing alternatives.

For the baseline model developed in Sections 4–7, I assume perfect trust in the contract administrator: agents believe with certainty ($\rho = 1$) that identities remain confidential when $M < T$. This assumption captures a best-case implementation environment where confidentiality and conditional release are credibly enforced, e.g., via audited cryptographic escrow. (The appendices discuss both implementation failures and

the comparative statics of relaxing this perfect-trust assumption.)

These ingredients set up the integrated model; Sections 4 and 5 derive the core propositions and threshold design implications, and Section 7 connects successful campaigns to shifts in discourse—the Overton window.

4 Baseline Model: Endogenous Safety in Numbers

This section develops the paper's core theoretical mechanism: endogenous safety in numbers. In this setting, some supportive agents are willing to speak only if they expect enough others to speak with them. The resulting equilibrium object is a fixed point in anticipated participation.

4.1 Type-Specific Cutoffs and Fixed-Point Participation

Fix threshold T and let $q \in [0, 1]$ denote the anticipated share of the eligible population that signs. Conditional on that anticipation, the success probabilities relevant to an individual signer are

$$p_{succ|i}(q;T) = \sum_{k=T-1}^{N-1} \binom{N-1}{k} q^k (1-q)^{N-1-k},$$

$$p_{succ|\neg i}(q;T) = \sum_{k=T}^{N-1} \binom{N-1}{k} q^k (1-q)^{N-1-k}.$$

Because the baseline large-N model suppresses exact pivotality, the common material value $\bar{v}$ drops out of the private signing margin. What remains is the expressive-exposure tradeoff. For a supportive agent with expressive benefit e_i and vulnerability type $\alpha_j \in \{\alpha_L, \alpha_H\}$, the expected gain from signing is

$$\Delta U_j(e_i; q, T) = p_{succ|i}(q;T)\big[e_i - \alpha_j g(q)\big] + p_{succ|\neg i}(q;T)s(\tau) + \eta(q).$$

Proposition 1 (Type-specific cutoff characterization)**.** *Consider the baseline model with support prevalence π, two vulnerability types $\alpha_i \in \{\alpha_L, \alpha_H\}$ with $\Pr(\alpha_i = \alpha_H) = \lambda$, continuous expressive-benefit distribution F_e, and safety-in-numbers cost $c_i(q) = \alpha_i g(q)$ with $g'(q) < 0$. For any fixed (q, T) with $p_{succ|i}(q;T) > 0$, a supportive agent of type α_j signs if and only if*

$$e_i \geq e_j^*(q;T), \qquad j \in \{L, H\},$$

where

$$e_j^*(q;T) = \alpha_j g(q) - \frac{p_{succ|\neg i}(q;T)s(\tau) + \eta(q)}{p_{succ|i}(q;T)}.$$

If $\alpha_H > \alpha_L$, then $e_H^*(q;T) \geq e_L^*(q;T)$ for every (q, T).

Given the type-specific cutoffs, the implied signing share among the eligible population is

$$S(q;T) \equiv \pi\Big[(1-\lambda)\big(1 - F_e(e_L^*(q;T))\big) + \lambda\big(1 - F_e(e_H^*(q;T))\big)\Big]. \tag{1}$$

A rational-expectations equilibrium is a fixed point

$$q = S(q; T). \tag{2}$$

Proposition 2 (Existence of baseline participation equilibrium)**.** *Suppose F_e is continuous and $g(\cdot)$ is continuous on $[0, 1]$. Then for every feasible threshold T, the participation map $S(\cdot; T)$ defined in* (1) *is continuous and maps $[0, 1]$ into itself. Hence the fixed-point problem* (2) *admits at least one equilibrium participation share $q^*(T) \in [0, 1]$.*

The proposition gives the paper the tractable object it needs without reverting to a static scalar type. Vulnerability now matters: the more exposed type needs a larger expressive benefit to sign at the same anticipated participation level.

4.2 Comparative Statics and Design Intuition

The baseline model yields three immediate lessons. First, larger abstention stigma $s(\tau)$ and a more favorable signing-utility schedule (higher $\bar{w}$ or lower k) lower both cutoffs in Proposition 1, increasing participation throughout the supportive population. Second, a stronger safety-in-numbers effect works disproportionately through the high-vulnerability type: when $g(q)$ declines more steeply in q, the newly induced participation is concentrated among agents who would otherwise remain silent because exposure is too costly. Third, threshold choice no longer affects the mechanism only through informativeness. A higher threshold makes success harder, but success at a higher threshold also implies a larger publicly revealed coalition and therefore a safer disclosure event. The key design tradeoff is thus endogenous rather than purely statistical.

Correlation between expressive benefit and vulnerability can be incorporated without changing the cutoff logic. Proposition 1 already conditions on type α_j, so allowing $F_e(\cdot \mid \alpha_j)$ to differ across vulnerability types changes only the mixing weights in the participation map. If expressive benefit and vulnerability are positively correlated ($E[e_i \mid \alpha_H] > E[e_i \mid \alpha_L]$), the agents most motivated to speak are also the most exposed, which makes the tipping point harder to reach but can make the participatory equilibrium more valuable once reached. If the correlation is negative, the agents most eager to speak are also the safest, so the safety-in-numbers channel matters less.

What is distinctive relative to standard provision-point mechanisms is that the contract changes the exposure cost of acting by conditioning revelation on a critical mass. Because abstention stigma arises only when a successful coalition becomes public, the contract can sustain participation even when material pivotality is negligible.

Architect-chosen intermediate disclosure is a secondary design margin rather than the paper's main mechanism. Online Appendix E.2 sketches a disclosure extension and related technical variants.

4.3 Multiplicity, Tipping Points, and Equilibrium Selection

Because safety in numbers creates strategic complementarities—higher anticipated participation lowers exposure cost, which raises actual participation—the participation map $S(\cdot; T)$ may admit multiple fixed points.

Because $\eta(0) < 0$, signing is unattractive when no one else is expected to participate, so $S(0;T) \approx 0$ and a low fixed point near zero exists. In the calibrated range, multiplicity takes the standard coordination-failure form: a low-participation equilibrium coexists with a high (participatory) equilibrium, separated by an unstable intermediate fixed point. To make this structure explicit, we can define for each threshold T at which three fixed points exist:

- $q^L(T)$: the low stable fixed point, under which the contract fails with high probability and no public revelation occurs;
- $q^U(T)$: the unstable intermediate fixed point (tipping point);
- $q^H(T)$: the high stable fixed point (participatory equilibrium).

Proposition 3 (Tipping-point interpretation)**.** *Suppose $S(\cdot;T)$ is continuous and admits three fixed points $q^L(T) < q^U(T) < q^H(T)$ with $S'(q^L;T) < 1$, $S'(q^U;T) > 1$, and $S'(q^H;T) < 1$. Then under iterative expectations $q_{t+1} = S(q_t;T)$:*

(i) *if $q_0 < q^U(T)$, the sequence converges to $q^L(T)$;*

(ii) *if $q_0 > q^U(T)$, the sequence converges to $q^H(T)$.*

The proof follows from standard one-dimensional fixed-point stability arguments; see Appendix A.

The unstable fixed point $q^U(T)$ is the coordination threshold: the minimum level of initial expectations needed for the campaign to converge to the participatory equilibrium rather than to $q^L(T)$, where the contract fails and no public revelation occurs. This gives a clean separation between two design questions. The *conditional* design problem asks: what threshold T is best once coordination onto $q^H(T)$ is achieved? The *ex ante* design problem asks: what threshold T is best once the difficulty of reaching $q^H(T)$ is taken into account? Section 4.4 develops that reduced-form extension.

In the numerical implementation, I compute equilibria by monotone iteration from both endpoints ($q = 0$ and $q = \pi$). The conditional design results reported in Sections 4.5–5.1 use the highest stable fixed point $q^H(T)$. Appendix A verifies that the qualitative threshold-design tradeoff is robust across the calibrated range (Figure 1).

4.4 Selection, Seeding, and Ex Ante Threshold Design

The conditional design problem takes coordination onto $q^H(T)$ as given. This subsection introduces a reduced-form ex ante design extension, which accounts for the probability that the campaign reaches the participatory equilibrium in the first place. It does not resolve equilibrium selection within the simultaneous game; rather, it captures the architect's problem when pre-recruitment, trust, and campaign momentum affect whether the high-participation equilibrium is reached.

Seeding and the coordination-success probability. The cleanest foundation is an organizer pre-recruitment model. Before the contract opens, the architect pre-recruits k committed signers whose participation is common knowledge among recruits. Initial expectations therefore satisfy $q_0 \geq k/N$. By Proposition 3, the

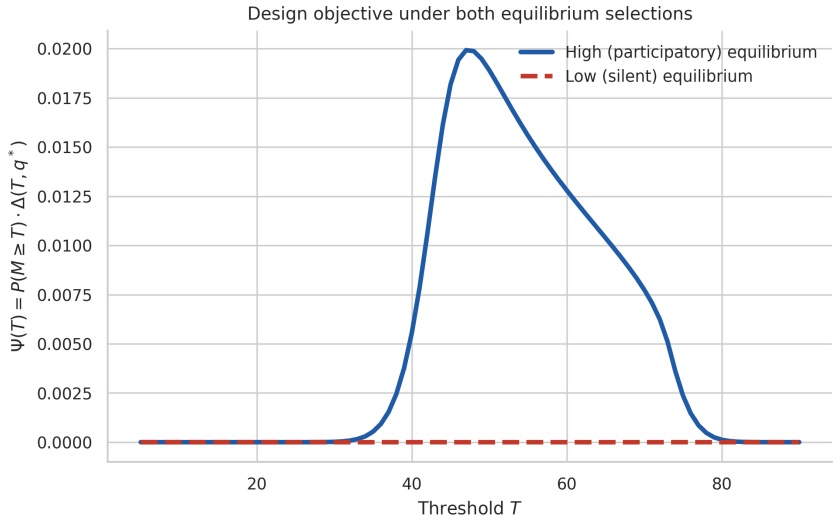


Figure 1: Conditional design objective under both equilibrium selections. Under the high (participatory) equilibrium (solid), the objective is hump-shaped with an interior optimum. Under the low-participation equilibrium (dashed), the contract fails with high probability and the objective value is near zero. Same calibration as Figure 2.

campaign reaches the participatory equilibrium if and only if $k/N \geq q^U(T)$, i.e., the seed clears the tipping point. Define the coordination-success probability

$$\sigma(T) \equiv \Pr\Big(k \geq \lceil N\, q^U(T) \rceil\Big),$$

where the probability is over whatever stochastic process governs seeding success (the organizer's recruitment effort, network reach, or trust conditions). The key economic content is that $\sigma(T)$ depends on T through the tipping-point size $q^U(T)$: higher thresholds generically raise $q^U(T)$, making coordination harder.

Trust, progress signals, and network structure also affect $\sigma(T)$ through the same channel: they shift the distribution of initial expectations q_0 relative to $q^U(T)$. Lower confidentiality trust ρ raises the effective exposure cost and therefore raises $q^U(T)$, making the tipping point harder to clear (Section 4.7). Progress signals that credibly convey campaign momentum can raise q_0 by updating agents' beliefs about others' participation. These determinants do not require separate formal treatment; they enter through the seeding distribution.

Ex ante design objective. Since the low equilibrium produces near-zero value, the ex ante design objective is

$$\Psi_{EA}(T) \equiv \sigma(T)\, \Psi_H(T), \tag{3}$$

where $\Psi_H(T) \equiv P(M \geq T \mid q^H(T), T)\, \Delta(T, q^H(T))$ is the conditional objective evaluated at the participatory equilibrium. This changes the design problem from "what threshold is best if coordination succeeds?" to "what threshold is best once one accounts for how difficult coordination is to achieve?"

Proposition 4 (Reduced-form ex ante design with coordination risk)**.** *Suppose $\Psi_H(T)$ is hump-shaped with*

interior conditional optimum T_H^*, and $\sigma(T)$ is weakly decreasing in T. Then the ex ante optimum satisfies $T_{EA}^* \leq T_H^*$.

Proof. Since $\sigma(T)$ is weakly decreasing and $\Psi_H(T)$ achieves its maximum at T_H^*, for any $T > T_H^*$ we have $\Psi_{EA}(T) = \sigma(T)\Psi_H(T) \leq \sigma(T_H^*)\Psi_H(T) < \sigma(T_H^*)\Psi_H(T_H^*) = \Psi_{EA}(T_H^*)$. Hence the ex ante optimum cannot exceed T_H^*. □

The assumption that $\sigma(T)$ is weakly decreasing is natural: higher thresholds raise $q^U(T)$, requiring a larger seed to escape the low equilibrium's basin of attraction. The proposition says that pricing in coordination risk shifts the optimal threshold weakly downward. The intuition is that high thresholds are doubly penalized: they reduce the conditional value of success (the declining part of the hump) *and* they are harder to coordinate onto.

4.5 Calibrated Baseline Numerical Exercise

To illustrate the baseline model quantitatively, I use an illustrative calibration chosen to place the model in a nondegenerate region rather than to match a specific dataset. The targets are an interior participatory equilibrium, meaningful heterogeneity in exposure risk, and signing incentives that are negative for an empty campaign but modestly positive once a coalition is expected to form. I normalize expressive benefits as $e_i \sim \mathcal{N}(0, 1)$ and set $N = 100$ as a transparent medium-sized benchmark. I let $g(q) = \exp(-\xi q)$ with $\xi = 3$, so that exposure cost falls by a factor of roughly $e^{-3} \approx 0.05$ as anticipated participation moves from zero to full. I set $\alpha_L = 0.5$ and $\alpha_H = 2.0$ with $\Pr(\alpha_i = \alpha_H) = \lambda = 0.4$, so that 40% of supportive agents are four times more vulnerable than the rest. The remaining parameters are $\pi = 0.65$, $s = 0.8$, $\bar{w} = 0.35$, and $k = 0.065$, so that $\eta(q) = 0.35q - 0.065$. This keeps the net signing utility near 0.10 around the participatory equilibrium while making signing into an empty contract unattractive ($\eta(0) = -0.065 < 0$).

The contract architect is not a welfare-maximizing planner. The architect's benchmark goal is to generate a publicly actionable coalition—a set of revealed endorsers large enough to be credible and visible. In the linear benchmark case, the value of success is proportional to the excess coalition share: the realized coalition share minus what was already anticipated. Define the *excess revealed coalition share* conditional on success as

$$\Delta(T, q^H(T)) \equiv E\left[\frac{M}{N} \;\middle|\; M \geq T\right] - q^H(T),$$

where $M \sim \text{Binomial}(N, q^H(T))$. This measures how much a successful campaign reveals about coordinated support beyond what was anticipated. Taking expectations yields the conditional design objective

$$\Psi_H(T) \equiv P(M \geq T \mid q^H(T), T)\, \Delta(T, q^H(T)).$$

This is not claimed to be a welfare criterion. It is the benchmark objective of an architect who values public coalition revelation. The paper also studies other architect objectives: durable success (Section 5.2) and Overton-window movement (Section 7). The qualitative hump shape is therefore not tied to a single payoff index; it reflects the underlying safety–coordination–revelation tradeoff.

Under this calibration, the baseline model delivers the following results. First, the conditional design objective $\Psi_H(T)$ is hump-shaped, confirming the central safety–coordination–revelation tradeoff (Figure 2). The conditional optimum under the participatory equilibrium is $T_H^* = 47$ ($\tau^* = 0.47$), with equilibrium participation $q^H(T_H^*) = 0.468$, success probability $P(M \geq T_H^*) \approx 0.521$, and excess revealed coalition share $\Delta(T_H^*, q^H) \approx 0.038$. Endogenous safety pulls the conditional optimum below what models with fixed exposure cost predict, because a moderate coalition already provides sufficient protection. Under the ex ante objective $\Psi_{EA}(T) = \sigma(T)\Psi_H(T)$ (Section 4.4), the optimum shifts weakly downward, consistent with Proposition 4: higher thresholds raise the tipping point $q^U(T)$ and are therefore harder to coordinate onto.

The conditional optimum is robust across calibrations. Across a parameter sweep varying $\xi \in \{1.5, 2, 3, 4, 5\}$, $\alpha_H \in \{1.5, 2, 2.5, 3\}$, and $\lambda \in \{0.2, 0.3, 0.4, 0.5\}$, the conditional optimum ranges from $T_H^* = 23$ to $T_H^* = 53$. The hump shape and interior optimum are qualitative features of every calibration in this range. Steeper safety-in-numbers effects pull T_H^* lower; greater vulnerability heterogeneity pushes T_H^* lower because the high-vulnerability type's differential response to safety is amplified. The same hump shape also survives under alternative safety functions, including matched linear and step specifications (Appendix Figure 8).

Second, safety in numbers works disproportionately through the high-vulnerability type (Figure 3, panel c). At the optimal threshold, roughly 77% of low-vulnerability supportive agents sign, compared to 64% of high-vulnerability agents. As the threshold rises and the equilibrium coalition grows, the gap narrows: the high-vulnerability type is differentially drawn into participation by the falling exposure cost.

Third, the strength of the safety-in-numbers effect is quantitatively consequential. When safety in numbers is turned off ($\xi = 0$), equilibrium participation drops to roughly 0.29 and the peak of the conditional objective shifts and shrinks. Stronger safety effects raise participation and move the conditional optimum. The mechanism's value depends on how much coalition size changes the cost of expression (Appendix Figure 7).

Computation and replication. All numerical figures are generated by Python scripts included in the replication package, available in an anonymized OSF repository: OSF replication package. The baseline fixed point is solved by damped monotone iteration from $q_0 \in \{0, \pi\}$ with damping 0.3, tolerance 10^{-8}, and a maximum of 500 iterations; fixed-point counts use a 2,000-point grid over anticipated participation. The same pipeline is used for the appendix robustness exercises.

4.6 Anchoring the Model: Corporate Whistleblowing

Social assurance contract implementations arrived on the scene very recently, so data is not available; the preceding calibration demonstrates the model's qualitative features using illustrative parameters. However, there is no shortage of incidents that might have benefited from a social assurance contract. This subsection anchors the same model to a setting where the coordination failure is empirically documented and the costs of delay are quantifiable: corporate whistleblowing. Corporate fraud often persists not because no insider knows, but because no insider can safely speak first. Employees are the single most important fraud detection channel, responsible for uncovering 19% of corporate fraud cases in a large-firm sample—more than regulators, auditors, or the media (Dyck et al. 2010). Yet the gap between private awareness and public reporting is large. In the Ethics & Compliance Initiative's Global Business Ethics Survey, 65% of employees globally reported

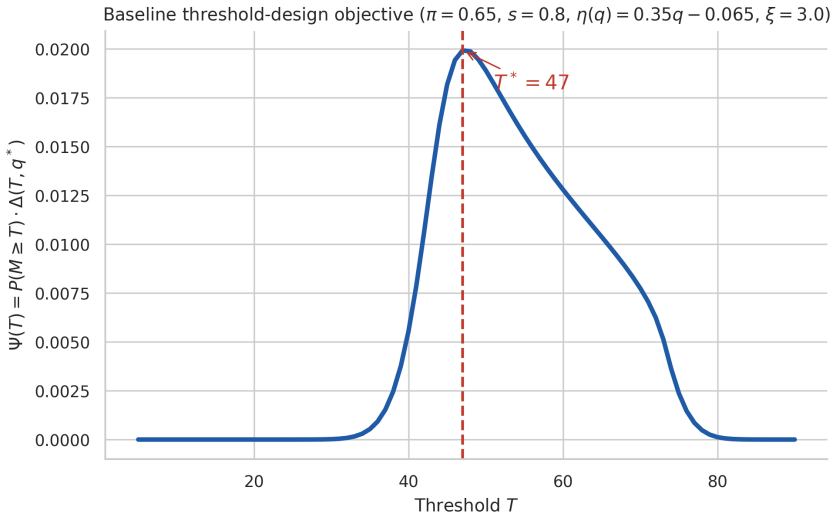


Figure 2: Conditional threshold-design objective under endogenous safety in numbers. The curve plots $\Psi_H(T) = P(M \geq T \mid q^H(T), T)\Delta(T, q^H(T))$ under the baseline calibration with $N = 100$, $\pi = 0.65$, $\alpha_L = 0.5$, $\alpha_H = 2.0$, $\lambda = 0.4$, $g(q) = e^{-3q}$, $e_i \sim \mathcal{N}(0, 1)$, $s = 0.8$, and $\eta(q) = 0.35q - 0.065$. The conditional optimum is $T^*_H = 47$.

observing workplace misconduct in 2023, but only 72% of observers reported it, leaving 28% of observers silent (Navex 2024). The silent share was 37% in 2013 (Center 2014). The reason is not apathy: 46% of those who *did* report experienced retaliation in 2023, and the figure reached 79% in the 2020–21 U.S. survey (Center 2021). Probationary federal employees who file whistleblower complaints face a termination rate of at least 10.1%, nearly ten times the 1.1% baseline for probationary employees government-wide (Office 2020). The problem this creates is large: an estimated 10% of large publicly traded firms commit securities fraud in any given year, only one-third of frauds are detected, and undetected fraud destroys roughly 1.6% of equity value annually—approximately $830 billion on a 2021 basis (Dyck et al. 2024).

This is the environment the model describes. Private awareness is widespread (π is substantial), but exposure cost for isolated reporters is high ($\alpha_H g(q)$ is large when q is small), so most observers stay silent. A social assurance contract—an escrowed, threshold-triggered collective report—could shift the equilibrium by ensuring that no reporter is exposed alone.

Wells Fargo: a case study in pluralistic ignorance

The Wells Fargo cross-selling scandal provides an especially clean illustration. Between 2002 and 2016, employees in Wells Fargo's Community Banking division opened approximately 3.5 million unauthorized customer accounts under intense sales pressure (Bureau 2016). The practice was not secret within the organization. Over five thousand employees were eventually terminated for creating fraudulent accounts, and the division employed roughly 100,000 people during the relevant period. Employees who contacted the internal ethics hotline were systematically retaliated against: an independent review identified 885 employees who called the ethics line between January 2011 and October 2016 and subsequently faced retaliation within

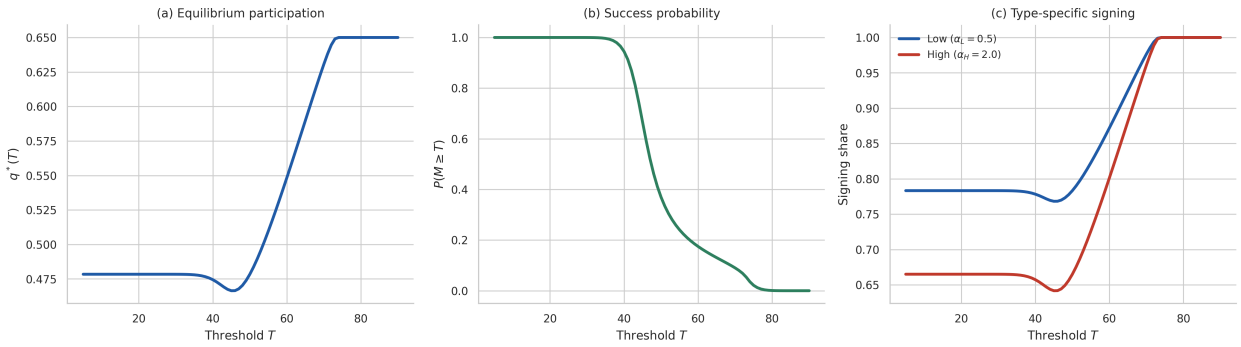


Figure 3: Equilibrium participation, success probability, and type-specific signing in the baseline model. Panel (a) plots $q^H(T)$. Panel (b) plots $P(M \geq T \mid q^H(T), T)$. Panel (c) shows type-specific signing shares: the low-vulnerability type ($\alpha_L = 0.5$) signs at higher rates throughout, but the high-vulnerability type ($\alpha_H = 2.0$) is differentially responsive to the safety-in-numbers channel. Same calibration as Figure 2.

twelve months (Independent Directors of the Board of Wells Fargo & Company 2017). Individual reporters were fired within days or weeks of calling. In 2022, the Department of Labor ordered Wells Fargo to pay over $22 million to a single whistleblower (U.S. Occupational Safety and Health Administration 2022).

The coordination failure is stark. Hundreds of employees were privately willing to report—they proved this by actually calling the hotline—but each acted alone, was identified, and was punished. The suppression extended up the hierarchy: when the Corporate HR Director expressed alarm at the approximately 1,000 annual terminations for sales integrity violations at a 2014 executive risk management committee meeting, the head of the Community Bank called to chastise her for "bullying" her people (Independent Directors of the Board of Wells Fargo & Company 2017). Senior leaders interpreted the low termination rate—roughly 1% of the workforce per year—as evidence the system was working, with one executive calling it "mind boggling. . . so low" (Independent Directors of the Board of Wells Fargo & Company 2017). The division's rolling 12-month turnover exceeded 30% throughout this period, meaning most front-line employees were relatively new and precarious at any given time. The first major external coverage appeared in 2013, but regulatory action did not arrive until September 2016—eleven years after the first documented internal ethics complaints reached the board in 2005. The cumulative cost exceeds $5 billion in regulatory penalties: $185 million in initial fines (2016), a $1 billion CFPB and OCC consent order for related abuses (2018), a $3 billion settlement with the Department of Justice and the SEC (2020), $575 million in state attorneys general settlements, and over $1 billion in shareholder litigation through 2023 (U.S. Department of Justice 2020).

Mapping the case to model parameters

The Wells Fargo case disciplines the model's parameters through observable quantities rather than illustrative targets. Table 2 summarizes the mapping. The eligible population is the Community Banking division ($N \approx 100{,}000$). To keep this exercise on the same numerical footing as the illustrative calibration, I report a stylized proxy with $N = 100$ and interpret thresholds through the normalized share $\tau = T/N$. The reported T values should therefore be read as threshold shares rather than literal headcounts, and any mapping back to the original population should be understood as a share-level heuristic rather than a count-exact structural

estimate.

Support prevalence. The 885 documented ethics-line callers imply that at least about 0.9% of the division was willing to report individually despite near-certain retaliation. The 5,300 employees terminated for participating in the fraud imply that direct knowledge was much broader. I therefore treat $\pi \in \{0.05, 0.15, 0.30\}$ as a stylized prevalence range rather than a literal count match: $\pi = 0.05$ is a conservative lower calibration above the hotline-caller share, $\pi = 0.15$ is a mid-range benchmark, and $\pi = 0.30$ reflects broader private awareness.

Vulnerability. The case exhibits extreme vulnerability heterogeneity. I set $\alpha_L = 0.5$ and $\alpha_H = 3.0$, with $\lambda = 0.6$ (60% high-vulnerability), reflecting that the majority of front-line Community Banking employees were in relatively precarious positions. The high vulnerability value is elevated relative to the illustrative calibration ($\alpha_H = 2.0$) to match the near-certain retaliation documented in the case.

Other parameters. I retain $g(q) = \exp(-3q)$ and $e_i \sim \mathcal{N}(0, 1)$. The signing-utility parameters are adjusted to $\bar{w} = 0.30$ and $k = 0.08$, reflecting higher friction for a formal whistleblowing commitment. The net signing utility $\eta(q) = 0.30q - 0.08$ remains negative for an empty campaign.

Table 2: Whistleblowing calibration: parameter mapping

Parameter	**Value**	**Empirical basis**
N	100	Community Banking div. (≈100K), normalized
π	$\{0.05, 0.15, 0.30\}$	Stylized range; 885 callers imply an ≈0.9% minimum
α_L	0.5	Employees with outside options or legal resources
α_H	3.0	At-will employees; near-certain retaliation
λ	0.6	Majority of front-line staff in precarious positions
$g(q)$	e^{-3q}	Retained from illustrative calibration
e_i	$\mathcal{N}(0, 1)$	Heterogeneous willingness to report
s	0.8	Stigma from revealed non-participation
$\bar{w}, k$	0.30, 0.08	Higher friction for formal whistleblowing

Counterfactual results

Under the mid-range whistleblowing calibration ($\pi = 0.15$), the conditional design objective $\Psi_H(T)$ retains its hump shape, with the conditional optimum at $T_H^* = 3$ ($\tau^* = 0.03$). In a 100,000-person division, that threshold share would mechanically correspond to roughly 3,000 employees, not three individual reporters. The substantive implication is therefore that the model favors a relatively permissive trigger *share*, not a literal trigger count of three. This is an order of magnitude lower than the illustrative calibration's $\tau^* = 0.47$, reflecting two forces: the lower support prevalence ($\pi = 0.15$ versus 0.65) constrains the achievable coalition, and the higher vulnerability ($\alpha_H = 3.0$, $\lambda = 0.6$) makes the safety channel more consequential. Equilibrium participation at the optimum is $q^H(T_H^*) = 0.033$, with success probability $P(M \geq T_H^*) \approx 0.64$. The high-vulnerability type is almost entirely shut out at equilibrium: only 1.4% of high-vulnerability supportive agents sign, compared to 52.8% of low-vulnerability agents (Figure 4, panel c). This is consistent with the case evidence: front-line employees could not safely report alone, and the mechanism would primarily operate through the small fraction with outside options or legal resources.

Across the support-prevalence range, the optimal threshold share rises with π: from $\tau^* = 0.02$ at $\pi = 0.05$ to $\tau^* = 0.07$ at $\pi = 0.30$ (equivalently, $T^* = 2$ to $T^* = 7$ in the normalized $N = 100$ proxy; Figure 4, panel b). In a 100,000-person division, those shares would mechanically correspond to roughly 2,000 to 7,000 employees. The hump shape and interior optimum persist throughout. These results should not be compared count-for-count to platforms such as Callisto, whose threshold of two is a literal headcount rather than a normalized share. The appropriate qualitative comparison is instead that sensitive-reporting escrows tend to use permissive triggers rather than majority-style thresholds.

These counterfactual numbers should be interpreted with appropriate caution. The exercise demonstrates that the model's qualitative features—interior optimum, differential vulnerability response, steep safety-in-numbers channel—survive calibration to an empirically documented coordination failure. However, we do not recover the exact threshold that would have surfaced the Wells Fargo fraud here. What is observed is the outcome of the coordination failure: eleven years of silence, hundreds of individually retaliated reporters, and over $5 billion in eventual penalties. Those costs are the price of the missing mechanism.

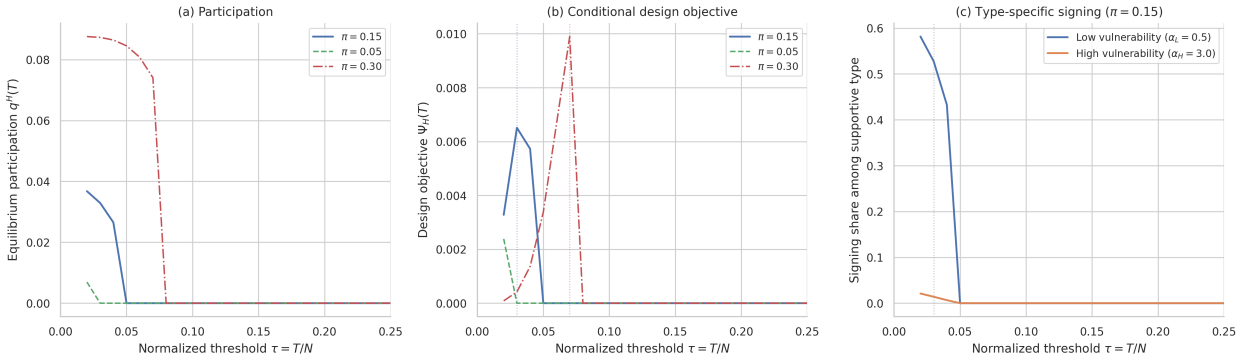


Figure 4: Whistleblowing calibration of the baseline model. Panel (a) plots equilibrium participation $q^H(T)$ across support prevalence levels $\pi \in \{0.05, 0.15, 0.30\}$. Panel (b) plots the conditional design objective $\Psi_H(T)$, which retains its hump shape at each prevalence level. Panel (c) shows type-specific signing shares at $\pi = 0.15$: the low-vulnerability type ($\alpha_L = 0.5$) drives participation, while the high-vulnerability type ($\alpha_H = 3.0$) is nearly shut out. Parameters: $N = 100$, $\lambda = 0.6$, $g(q) = e^{-3q}$, $e_i \sim \mathcal{N}(0,1)$, $s = 0.8$, $\eta(q) = 0.30q - 0.08$.

External validation

The Wells Fargo calibration is not an outlier. The 28–37% of misconduct observers who stay silent in cross-sectional surveys (Navex 2024, Center 2014) maps to the model's prediction that a substantial share of supportive agents choose $a_i = 0$ in the low-participation equilibrium. The 46–79% retaliation rate among reporters (Navex 2024, Center 2021) is consistent with $\alpha_H g(q)$ being large at low participation levels. The 10% annual fraud rate among large firms with only one-third detected (Dyck et al. 2024) implies a large stock of firms in the low-participation equilibrium. Tips account for 43% of all fraud detection, and 52% of tips come from employees (Association of Certified Fraud Examiners, Inc. 2025); the model implies that a threshold-triggered mechanism could increase this rate by converting private awareness into coordinated disclosure. Causal evidence supports the exposure-cost channel directly: Heese and Pérez-Cavazos (2021)

find that exogenous reductions in retaliation costs (through higher state unemployment-insurance benefits) increase employee whistleblowing, with the effect concentrated in firms with weak employee relations and internal control weaknesses—precisely the settings where $\alpha_H g(q)$ is largest.

4.7 Confidentiality Trust as a Baseline Comparative Static

The baseline model assumes perfect trust ($\rho = 1$): agents believe with certainty that identities remain confidential when $M < T$. Under imperfect trust ($\rho < 1$), a signer faces expected exposure cost even when the contract fails. Specifically, a signer of type α_j who expects the contract to fail receives expected payoff $\eta(q) - (1-\rho)\alpha_j g(q)$ rather than $\eta(q)$, because with probability $1-\rho$ identities leak despite failure.

This has three immediate consequences for the baseline participation model. First, lower ρ raises both type-specific cutoffs e_L^* and e_H^* in Proposition 1, reducing equilibrium participation. Second, the effect is disproportionately borne by the high-vulnerability type: the additional expected cost $(1-\rho)\alpha_H g(q)$ is larger than $(1-\rho)\alpha_L g(q)$, so trust erosion differentially deters the agents the mechanism most needs to draw in. Third, lower trust raises the tipping point $q^U(T)$, making the participatory equilibrium harder to reach and the low equilibrium more likely. Trust is therefore not merely an implementation detail—it is a first-order determinant of both participation levels and coordination success. The trust parameter can also capture false-negative failures: the administrator may fail to publish the coalition even when $M \geq T$, for instance because of external coercion, technical malfunction, or governance failure. Such failures expose signers to risk without delivering the coordination and reputational benefits of success, further raising the effective participation threshold. The fuller implementation-risk taxonomy appears in Online Appendix F, which covers Sybil attacks, insider threats, cryptographic failure modes, and blockchain deployment risks.

5 Threshold Design and Durable Success

5.1 Threshold Design in the Baseline Model

Let $q^H(T)$ denote the high equilibrium participation share induced by threshold T in the baseline model. The conditional design criterion, evaluated at the participatory equilibrium, is

$$\Psi_H(T) \equiv P(M \geq T \mid q^H(T), T)\, \Delta(T, q^H(T)),$$

where $\Delta(T, q^H(T))$ denotes the excess revealed coalition share at threshold T (see Section 4.5). The threshold T enters through both factors: raising it can make success more informative and safer, but can also depress participation if the contract becomes too hard to clear. The interaction between safety, coordination, and the value of successful revelation is the core of the design problem.

5.2 Durability Under Reduced-Form Opposition

The threshold problem becomes sharper if success can trigger organized retaliation. Opposition enters here as a parameter rather than as a strategic actor: the opponent's capacity Ω is exogenous, and punishment dilutes

with coalition size. A richer model with endogenous counter-mobilization is sketched in Online Appendix H. If opposition capacity itself scales with coalition size, for example because larger campaigns provoke stronger backlash, dilution need not be monotone and a durability sweet spot can emerge. The reduced-form baseline here deliberately holds that feedback fixed.

Suppose a successful campaign faces aggregate opposition capacity $\Omega \geq 0$. If M endorsers are revealed, let each revealed endorser bear expected post-disclosure burden $r(M;\Omega)$, where $\partial r/\partial M < 0$ and $\partial r/\partial \Omega > 0$: larger coalitions dilute punishment, while stronger opposition raises it. Let $\bar{c}$ denote the largest post-disclosure burden compatible with durable public expression after the list is revealed. The durability floor acts as an effective higher threshold:

$$M_O(\Omega) \equiv \begin{cases} \min\{m \in \{1,\dots,N\} : r(m;\Omega) \leq \bar{c}\}, & \text{if this set is nonempty,} \\ N+1, & \text{otherwise.} \end{cases}$$

In the simple equal-burden case $r(M;\Omega) = \Omega/M$, this becomes $M_O(\Omega) = \lceil \Omega/\bar{c} \rceil$. A campaign produces durable success if and only if $M \geq \max\{T, M_O(\Omega)\}$. The architect who values durable outcomes should therefore solve

$$\Psi^{dur}(T;\Omega) \equiv P\big(M \geq \max\{T, M_O(\Omega)\} \mid q^H(T), T\big)\, \Delta^{dur}(T;\Omega),$$

where $\Delta^{dur}(T;\Omega) \equiv E[M/N \mid M \geq \max\{T, M_O(\Omega)\}] - q^H(T)$. Thresholds below $M_O(\Omega)$ can still generate publication, but they admit fragile victories. In the baseline calibration with a durability floor of $M_O = 60$, the durable objective shifts the conditional optimum upward, because only larger coalitions produce lasting outcomes.

6 Comparison with Alternative Mechanisms

Social assurance contracts occupy a niche between anonymous surveys (which reveal aggregate support but not identities) and open petitions (which reveal identities without safety). This section formalizes that claim with a stylized model that compares the three mechanisms on the same population, identifies precisely when each succeeds, and shows that the assurance contract is uniquely effective in an intermediate parameter region where the other two fail.

6.1 Setup

Consider N eligible agents. Fraction π are supporters, drawn independently. Among supporters, fraction μ are *invulnerable* (vulnerability $\alpha_i = 0$) and fraction $1-\mu$ are *vulnerable* ($\alpha_i = \alpha > 0$). All supporters share a common expressive benefit $e > 0$ from being publicly identified as part of a successful coalition. If M agents are publicly identified as supporters, each revealed agent with vulnerability α_i bears exposure cost $\alpha_i \cdot g(M/N)$, where $g : [0,1] \to \mathbb{R}_+$ is continuous and strictly decreasing with $g(0) = 1$. Signing an assurance contract carries a small friction $k > 0$.

The two-type structure isolates the key force. Two cost thresholds govern the analysis:

- *Full-coalition cost*: $\alpha \cdot g(\pi)$—the exposure cost when all supporters are revealed together.
- *Pioneer cost*: $\alpha \cdot g(\pi\mu)$—the exposure cost when only the invulnerable base is revealed.

Since g is decreasing and $\pi\mu < \pi$, the pioneer cost strictly exceeds the full-coalition cost. The gap $\alpha \cdot [g(\pi\mu) - g(\pi)]$ is the *safety-in-numbers wedge*: the cost reduction available to a vulnerable type who acts as part of the full coalition rather than joining only the pioneers.

6.2 Three mechanisms

(S) Post-survey expression game. An anonymous survey reveals π to all agents. Agents then simultaneously choose whether to express publicly. An agent who expresses is publicly identified regardless of others' choices.

(P) Open petition. Agents are sequentially offered the opportunity to sign. A signer's identity is immediately and publicly revealed. Each agent observes all prior signatures before deciding.

(A) Assurance contract. Agents simultaneously choose whether to sign. An assurance threshold $T \in \{1, \ldots, N\}$ is announced in advance. If $M \geq T$, all signers' identities are revealed; otherwise no information is released. Signing costs k.

6.3 When does each mechanism produce full coordination?

Proposition 5 (Mechanism comparison)**.** *The parameter space partitions into three regions.*

(i) ***Cascade-connected*** $[e \geq \alpha \cdot g(\pi\mu)]$*: All three mechanisms produce full supporter participation in equilibrium. The post-survey expression game has a unique equilibrium in which all supporters express. The open petition cascade completes. The assurance contract succeeds for any $T \leq \pi N$.*

(ii) ***Coordination gap*** $[\alpha \cdot g(\pi) \leq e < \alpha \cdot g(\pi\mu)]$*: Full supporter participation is self-sustaining but not self-starting. The post-survey expression game admits both a high-participation and a low-participation equilibrium. The open petition cascade stalls at $M = \pi\mu N$. There exists $T^* \in (\pi\mu N, \pi N]$ such that the assurance contract admits a full-participation equilibrium in which all supporters sign and $M = \pi N \geq T^*$.*

(iii) ***Fundamentally blocked*** $[e < \alpha \cdot g(\pi)]$*: No mechanism produces vulnerable-type participation. Even if all supporters acted together, exposure cost would exceed the expressive benefit. Under all three mechanisms, at most the $\pi\mu N$ invulnerable supporters participate.*

Proof. Part (i). When $e \geq \alpha \cdot g(\pi\mu)$, a vulnerable type's payoff from public expression is $e - \alpha \cdot g(M/N) \geq e - \alpha \cdot g(\pi\mu) \geq 0$ for any $M \geq \pi\mu N$. In the post-survey expression game, invulnerable types always express ($e > 0$, $\alpha_i = 0$), generating $M \geq \pi\mu N$; vulnerable types then strictly prefer to express, so the unique equilibrium is full participation. In the open petition, invulnerable types sign first; the first vulnerable agent faces cost $\alpha \cdot g(\pi\mu) \leq e$ and signs; the cascade propagates because g is decreasing in M. In the assurance contract, signing weakly dominates not signing for all supporters when $T \leq \pi N$ and $e \geq \alpha \cdot g(\pi\mu)$.

Part (ii). The high-participation equilibrium exists because $e \geq \alpha \cdot g(\pi)$ ensures that a vulnerable type who expects all supporters to act is willing to act. The low-participation equilibrium exists because $e < \alpha \cdot g(\pi\mu)$ ensures that a vulnerable type who expects only invulnerables to act is not willing to join. In the open petition, invulnerable types sign; the first vulnerable type faces pioneer cost $\alpha \cdot g(\pi\mu) > e$ and does not sign; the cascade stalls. For the assurance contract, set $T \in (\pi\mu N, \pi N]$. If all supporters sign, $M = \pi N \geq T$, and each vulnerable signer receives $e - \alpha \cdot g(\pi) - k > 0$ (for small k). If only invulnerable types sign, $M = \pi\mu N < T$, and the contract fails: no identities are revealed, and each non-signer receives 0. Both are Nash equilibria; the risk-dominance comparison in Corollary 1 shows that the assurance contract selects the high equilibrium.

Part (iii). If $e < \alpha \cdot g(\pi)$, then even at full participation $M = \pi N$, a vulnerable type's payoff $e - \alpha \cdot g(\pi) < 0$. No equilibrium of any mechanism has vulnerable types participating. □

Corollary 1 (Risk dominance)**.** *In region* (ii)*, let a vulnerable type hold belief p that the high-participation outcome obtains. In the assurance contract, signing is preferred whenever*

$$p > \frac{k}{e - \alpha \cdot g(\pi)},$$

which approaches 0 *as $k \to 0$. In the post-survey expression game, expressing is preferred whenever*

$$p > \frac{\alpha \cdot g(\pi\mu) - e}{\alpha \cdot [g(\pi\mu) - g(\pi)]},$$

which is bounded away from 0 *throughout region* (ii) *and equals* 1 *at its lower boundary $e = \alpha \cdot g(\pi)$.*

Proof. In the assurance contract with $T \in (\pi\mu N, \pi N]$: if the high outcome obtains ($M = \pi N \geq T$), signing yields $e - \alpha \cdot g(\pi) - k$; if the low outcome obtains ($M = \pi\mu N < T$), signing yields $-k$. Expected payoff from signing is $p \cdot [e - \alpha \cdot g(\pi)] - k$; from not signing, 0. The threshold follows.

In the post-survey expression game: if the high outcome obtains, expressing yields $e - \alpha \cdot g(\pi)$; if the low outcome obtains, expressing yields $e - \alpha \cdot g(\pi\mu) < 0$. Expected payoff from expressing is $p \cdot [e - \alpha \cdot g(\pi)] + (1 - p) \cdot [e - \alpha \cdot g(\pi\mu)]$; setting this to zero and solving yields the threshold. □

The payoff asymmetry driving the corollary is that the assurance contract truncates the downside: a vulnerable type who signs when the contract fails loses only the small friction k, whereas a vulnerable type who expresses publicly when few others do bears the full pioneer cost $\alpha \cdot g(\pi\mu) - e > 0$. This asymmetry makes the high-participation equilibrium the risk-dominant selection under the assurance contract throughout region (ii).

When does the open petition cascade complete? The cascade condition $e \geq \alpha \cdot g(\pi\mu)$ makes the required invulnerable base explicit. With the exponential safety function $g(q) = \exp(-\xi q)$ used in the baseline calibration, the cascade completes if and only if

$$\pi\mu \geq \frac{1}{\xi} \ln \frac{\alpha}{e}.$$

With parameters from the whistleblowing calibration ($\xi = 3$, $\alpha/e \approx 3$), this requires $\pi\mu \geq 0.37$—more than a third of the eligible population must be effectively invulnerable for sequential revelation to propagate. In the

Wells Fargo setting, where the vast majority of employees faced severe retaliation and the invulnerable base was negligible, the cascade condition fails by a wide margin. This is consistent with the observed pattern: individual whistleblowers were isolated and punished one by one, while the silent majority sustained the appearance of compliance. The assurance contract is designed precisely for this gap.

Mapping to the baseline model. The two-type comparison model maps directly onto the baseline. The invulnerable share μ corresponds to the low-vulnerability fraction $1 - \lambda$ of the baseline's two-type structure. The coordination-gap region (ii) corresponds to the parameter range in which the baseline participation map $S(q;T)$ admits multiple fixed points: full participation is self-sustaining (the high fixed point exists) but not self-starting (the low fixed point also exists and is the cascade outcome). The threshold T^* in region (ii) is the analogue of the optimal interior threshold T_H^* identified in Section 5.

7 Public Belief Updating After Disclosure: The Overton Window

The baseline model identifies how threshold choice changes the coalition that appears and how safety in numbers shapes who is willing to sign. This section shows that the same model also delivers a belief-updating result: because success is more likely when support prevalence is higher, observing success raises outside observers' posterior beliefs about π in expectation. Under a reduced-form mapping from beliefs to expression costs, this posterior shift can widen the Overton window.

Agents in the eligible population know π and play the equilibrium characterized in Section 4. Outside observers—the broader public, institutional audiences, or media—do not know π and hold a prior $\pi \sim \text{Beta}(a_\pi, b_\pi)$. When the contract succeeds and $M \geq T$ endorsers are publicly revealed, outside observers update their beliefs about π.

Proposition 6 (Public belief updating from the baseline model)**.** *Let $q^H(T;\pi)$ denote the highest stable equilibrium participation share as a function of support prevalence π, and let outside observers hold prior $\pi \sim \text{Beta}(a_\pi, b_\pi)$. Define the success likelihood*

$$L_T(\pi) = P(M \geq T \mid q^H(T;\pi), T) = \sum_{m=T}^{N} \binom{N}{m} q^H(T;\pi)^m \big(1 - q^H(T;\pi)\big)^{N-m}.$$

(i) *Under the highest-stable-fixed-point selection rule, $q^H(T;\pi)$ is nondecreasing in π.*

(ii) *Hence $L_T(\pi)$ is nondecreasing in π: success is weakly more likely when support is more prevalent.*

(iii) *If success occurs with probability strictly between* 0 *and* 1*,*

$$\Delta_\pi(T) \equiv E[\pi \mid M \geq T, T] - E[\pi] > 0.$$

Corollary 2 (Reduced-form Overton implication)**.** *If expected individual expression cost for a focal belief $\mathbf{b}_0$ depends on the public belief state through the posterior mean of π and is locally affine,*

$$c_{indiv}(\mathbf{b}_0, \mathcal{B}) = \bar{c}_R(\mathbf{b}_0) - \kappa_R(\mathbf{b}_0) E[\pi \mid \mathcal{B}], \qquad \kappa_R(\mathbf{b}_0) > 0,$$

then success reduces expression cost by $\kappa_R(\mathbf{b}_0)\Delta_\pi(T)$, and the focal belief enters the Overton window after a successful contract if and only if

$$\Delta_\pi(T) \geq \frac{c_{indiv}(\mathbf{b}_0, \mathcal{B}_t) - \bar{\tau}}{\kappa_R(\mathbf{b}_0)}.$$

The proof of Proposition 6 is in Appendix A. The MLRP step in part (iii) is standard Bayesian updating: success is more likely when π is higher, so observing success raises the posterior mean in expectation. The contribution of the participation model is to generate the likelihood structure in parts (i)–(ii): because the participation map $S(q;T)$ is proportional to π, the selected high fixed point $q^H(T;\pi)$ is nondecreasing in π, which in turn ensures that $L_T(\pi)$ satisfies the monotone likelihood property. The safety channel creates multiplicity and shapes the level of q^H; the monotone comparative static in π is driven by the extensive margin of supportive agents.

Proposition 6 conditions on the coarse event $\{M \geq T\}$. The actual mechanism reveals the full count M, not only whether the threshold was met. A realized coalition that barely clears T could attenuate the belief update relative to the coarse-event posterior and, in some cases, even move the exact-count posterior below the prior. The unambiguous positive result in Proposition 6 is therefore an expected-value statement for the success event.

If the contract's existence is common knowledge, failure is informative as well. The same monotone-likelihood argument implies

$$E[\pi \mid M < T, T] < E[\pi].$$

A failed contract can therefore shrink the perceived room for public expression rather than expand it. That downside risk is not priced into the benchmark coalition-revelation objective $\Psi_H(T)$, which values only the success side. It strengthens the case against excessively high thresholds, because a threshold that is too ambitious risks not only non-publication but also a publicly discouraging signal when failure is observed. If the contract is covert unless it succeeds, by contrast, failure reveals nothing to outsiders.

The Overton-gap reduction in Corollary 2 depends on the local affine assumption. Convex expression costs would dampen the gap reduction from a given $\Delta_\pi(T)$, because the marginal effect of a belief shift is weaker in the region where the posterior mean lands. Dependence on higher moments of the posterior (e.g., if observers are uncertainty-averse) would introduce additional channels beyond the mean shift. These departures do not invalidate the baseline result but clarify that the clean proportionality between $\Delta_\pi(T)$ and gap reduction is a consequence of the maintained functional form. More generally, the belief-updating result in Proposition 6 holds without the affine assumption; the Overton interpretation is the reduced-form extension.

Define the Overton window as $O_R(\mathcal{B}_t, \bar{\tau}) = \{\mathbf{b} \in \mathcal{I} \mid c_{indiv}(\mathbf{b}, \mathcal{B}_t) \leq \bar{\tau}\}$ and the focal belief's Overton gap as $g_R(\mathbf{b}_0) \equiv c_{indiv}(\mathbf{b}_0, \mathcal{B}_t) - \bar{\tau}$. The ex ante expected gap reduction from choosing threshold T is $\kappa_R(\mathbf{b}_0)\Psi_O(T)$, where $\Psi_O(T) \equiv P(M \geq T \mid q^H(T), T)\,\Delta_\pi(T)$ is the baseline Overton design objective. Threshold design therefore directly chooses the expected discourse shift. If opposition matters (Section 5.2), the same logic applies with the durable-success threshold replacing T.

Under the baseline calibration with $\pi \sim \text{Beta}(13, 7)$ (so that $E[\pi] = 0.65$, matching Section 4.5), $\Psi_O(T)$ is hump-shaped and maximized at $T_O^* = 46$, slightly below the threshold $T_H^* = 47$ that maximizes the conditional participation objective (Figure 5). The posterior shift $\Delta_\pi(T)$ is monotonically increasing in T:

higher thresholds are more informative conditional on success, but they succeed less often. The magnitude of that posterior shift is sensitive to prior concentration: more diffuse priors produce larger discourse shifts because outside observers have more to learn from a successful campaign (Appendix Figure 9).

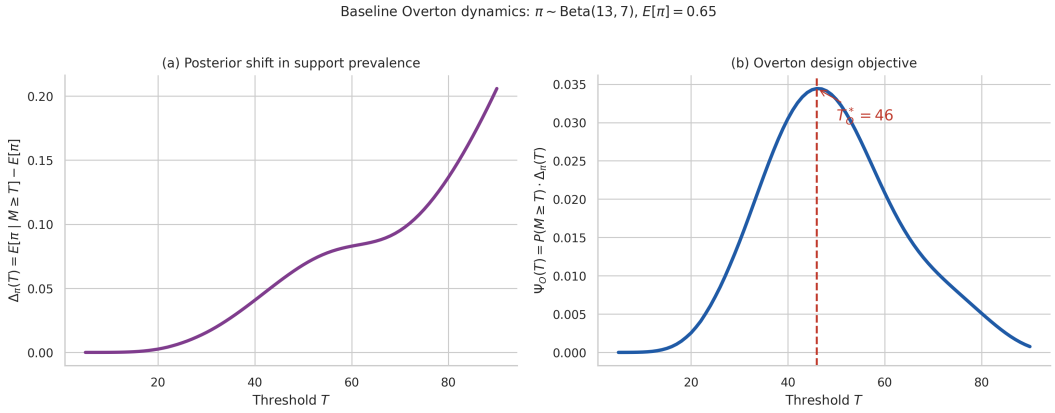


Figure 5: Baseline Overton dynamics under $\pi \sim \text{Beta}(13, 7)$. Panel (a) plots the posterior shift in support prevalence $\Delta_\pi(T) = E[\pi \mid M \geq T] - E[\pi]$ for the coarse success event. Panel (b) plots the Overton design objective $\Psi_O(T) = P(M \geq T)\Delta_\pi(T)$, maximized at $T_O^* = 46$. Same baseline calibration as Figure 2.

Remark. The finite-N benchmark in Appendix C delivers a structurally parallel result (Proposition 10) with a latent support state θ replacing π and signal-indexed cutoffs replacing the baseline's type-specific cutoffs. That benchmark adds exact pivotality and posterior weighting over θ conditional on private signals. The baseline version here is simpler but sufficient for the discourse-shift channel: what outside observers learn from success is, on average, that support prevalence π is higher than they expected, which is enough to reduce expression costs under the reduced-form mapping.

8 Discussion and Conclusion

Social assurance contracts are useful when the problem is not just to measure support, but to reveal a coalition large enough to speak together. The model given here makes that coordination problem explicit: endogenous safety in numbers creates strategic complementarities, so threshold choice simultaneously affects who is willing to sign, how likely success is, and how much a successful revelation says about latent coalition strength. Because outside observers do not know support prevalence, a successful campaign also updates public beliefs about π on average. When opposition can retaliate, the durability of success imposes a higher effective threshold.

Scope and normative considerations. The mechanism is content-neutral. That is part of its attraction, but also part of its danger. A social assurance contract can surface suppressed but constructive views; it can also coordinate support for positions that other observers regard as harmful, exclusionary, or destabilizing. The positive results in this paper therefore do not imply that more revelation is always better. They show when conditional disclosure can move public expression closer to privately held sentiment within a specified eligible population. Whether that movement is socially desirable depends on the content of the claim, the institutional

setting, and the downstream consequences of making a coalition publicly legible. The mechanism is most defensible when public coalition formation has independent value, the eligible population and threshold rule are transparent, and fallback mechanisms (anonymous surveys, secret ballots) remain available when private measurement or formal aggregation would achieve the institution's goal with less exposure.

Applications. The Wells Fargo case illustrates the core pattern: hundreds of employees were privately willing to report fraud—they proved it by calling the ethics hotline—but each was isolated and punished, while the silent majority sustained the appearance of compliance. An escrowed, threshold-triggered collective report would let employees submit complaints privately and reveal identities only once enough co-reporters exist to make retaliation impractical. The calibrated model shows that in high-retaliation environments the optimal threshold is low, and the mechanism operates primarily through the subset of employees with some insulation from retaliation. Similar logic applies to labor organizing, where isolated early supporters of a union campaign are easier to pressure than a large group moving together (NLRB 2026a,b); to shareholder activism, where dispersed shareholders consider backing a filing (U.S. Securities and Exchange Comission 1998, 2023); and to open letters on sensitive policy issues. Across these settings, the institution needs not merely an estimate of support, but a coalition that can be revealed together.

Limitations and Future directions. The framework is intentionally stylized in several respects. First, the model is static and simultaneous, so it abstracts from the recruitment path by which campaigns build momentum, from sequential revelation, and from belief updating after earlier successes or failures. Second, the ex ante coordination layer is reduced-form: the paper does not provide a structural model of seeding, organizer effort, or equilibrium selection, but instead summarizes those forces through the probability of reaching the participatory equilibrium. Third, trust, disclosure design, and retaliation are only partially endogenous in the main text. The model shows why those forces matter, but it does not endogenize administrator behavior, legal enforcement, or strategic counter-mobilization. The numerical exercises—both the illustrative calibration and the whistleblowing counterfactual—are calibrated rather than structurally estimated, so the quantitative results should be read as mechanism-demonstrating rather than directly predictive. These limitations point to a broader empirical and theoretical agenda: platform data, laboratory experiments, or field experiments could test whether higher stigma raises participation, whether threshold choice produces the predicted hump-shaped objective, and whether stronger retaliation pushes successful campaigns toward larger coalitions (Cantoni et al. 2019, Tabarrok 1997). An additional open question is how multiple assurance contracts interact when they represent rival positions.

References

Douglas J. Ahler. Self-Fulfilling Misperceptions of Public Polarization. *The Journal of Politics*, 76(3):607–620, July 2014. ISSN 0022-3816. doi: 10.1017/S0022381614000085.

S. Nageeb Ali and Roland Bénabou. Image versus Information: Changing Societal Norms and Optimal Privacy. *American Economic Journal: Microeconomics*, 12(3):116–164, August 2020. ISSN 1945-7669. doi: 10.1257/mic.20180052.

Thomas Apolte and Julia Müller. The persistence of political myths and ideologies. *European Journal of Political Economy*, 71:102076, January 2022. ISSN 0176-2680. doi: 10.1016/j.ejpoleco.2021.102076.

Association of Certified Fraud Examiners, Inc. ACFE Report to the Nations | 2024 Global Fraud Study. Technical report, Association of Certified Fraud Examiners, Inc., 2025.

Ian Ayres and Cait Unkovic. Information Escrows. *Michigan Law Review*, 111:145, 2012.

Mark Bagnoli and Barton L. Lipman. Provision of Public Goods: Fully Implementing the Core through Private Contributions. *The Review of Economic Studies*, 56(4):583–601, October 1989. ISSN 0034-6527. doi: 10.2307/2297502.

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, October 1992. ISSN 0022-3808, 1537-534X. doi: 10.1086/261849.

Sushil Bikhchandani, David Hirshleifer, Omer Tamuz, and Ivo Welch. Information Cascades and Social Learning. *Journal of Economic Literature*, 62(3):1040–1093, September 2024. ISSN 0022-0515. doi: 10.1257/jel.20241472.

Luca Braghieri. Political Correctness, Social Image, and Information Transmission. *American Economic Review*, 114 (12):3877–3904, December 2024. ISSN 0002-8282. doi: 10.1257/aer.20210039.

U.S. Consumer Financial Protection Bureau. Consumer Financial Protection Bureau Fines Wells Fargo $100 Million for Widespread Illegal Practice of Secretly Opening Unauthorized Accounts. https://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-fines-wells-fargo-100-million-widespread-illegal-practice-secretly-opening-unauthorized-accounts/, September 2016.

Leonardo Bursztyn, Alessandra L. González, and David Yanagizawa-Drott. Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *American Economic Review*, 110(10):2997–3029, October 2020. ISSN 0002-8282. doi: 10.1257/aer.20180975.

Callisto. Callisto. https://www.projectcallisto.org, 2025.

Jan Camenisch and Anna Lysyanskaya. An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Birgit Pfitzmann, editors, *Advances in Cryptology — EUROCRYPT 2001*, volume 2045, pages 93–118. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-42070-5. doi: 10.1007/3-540-44987-6_7.

Davide Cantoni, David Y Yang, Noam Yuchtman, and Y Jane Zhang. Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement*. *The Quarterly Journal of Economics*, 134(2): 1021–1077, May 2019. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjz002.

Carla Capuano and Peggy Chekroun. A Systematic Review of Research on Conformity. *International Review of Social Psychology*, 37(1), July 2024. doi: 10.5334/irsp.874.

Hans Carlsson and Eric van Damme. Global Games and Equilibrium Selection. *Econometrica*, 61(5):989–1018, 1993. ISSN 0012-9682. doi: 10.2307/2951491.

Timothy N. Cason and Robertas Zubrickas. Enhancing fundraising with refund bonuses. *Games and Economic Behavior*, 101:218–233, January 2017. ISSN 08998256. doi: 10.1016/j.geb.2015.11.001.

Ethics Resource Center. 2013 National Business Ethics Survey. Technical report, Ethics Resource Center, 2014.

Ethics Resource Center. Global Business Ethics Survey 2020. Technical report, Ethics Resource Center, 2021.

Rachel T. A. Croson and Melanie Beth Marks. Step returns in threshold public goods: A meta- and experimental analysis. *Experimental Economics*, 2(3):239–259, March 2000. ISSN 1386-4157, 1573-6938. doi: 10.1007/BF01669198.

John R. Douceur. The Sybil Attack. In Peter Druschel, Frans Kaashoek, and Antony Rowstron, editors, *Peer-to-Peer Systems*, pages 251–260, Berlin, Heidelberg, 2002. Springer. ISBN 978-3-540-45748-0. doi: 10.1007/3-540-45748-8_24.

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

John Duffy and Jonathan Lafky. Living a Lie: Theory and Evidence on Public Preference Falsification. *Department of Economics Working Paper Series*, September 2018.

Philip H. Dybvig and Chester S. Spatt. Adoption externalities as public goods. *Journal of Public Economics*, 20(2):231–247, March 1983. ISSN 0047-2727. doi: 10.1016/0047-2727(83)90012-9.

Alexander Dyck, Adair Morse, and Luigi Zingales. Who Blows the Whistle on Corporate Fraud? *The Journal of Finance*, 65(6):2213–2253, December 2010. ISSN 0022-1082, 1540-6261. doi: 10.1111/j.1540-6261.2010.01614.x.

Alexander Dyck, Adair Morse, and Luigi Zingales. How pervasive is corporate fraud? *Review of Accounting Studies*, 29(1):736–769, March 2024. ISSN 1573-7136. doi: 10.1007/s11142-022-09738-5.

James K Esser. Alive and Well after 25 Years: A Review of Groupthink Research. *Organizational Behavior and Human Decision Processes*, 73(2-3):116–141, February 1998. ISSN 07495978. doi: 10.1006/obhd.1998.2758.

David Evans, Vladimir Kolesnikov, and Mike Rosulek. A Pragmatic Introduction to Secure Multi-Party Computation. *Foundations and Trends in Privacy and Security*, 2(2-3):70–246, December 2018. ISSN 2474-1558. doi: 10.1561/3300000019.

Mauricio Fernández-Duque. The probability of pluralistic ignorance. *Journal of Economic Theory*, 202:105449, June 2022. ISSN 0022-0531. doi: 10.1016/j.jet.2022.105449.

Sara Forsdyke. *Exile, Ostracism, and Democracy: The Politics of Expulsion in Ancient Greece*. Princeton University Press, Princeton, 2009. ISBN 978-1-4008-2686-5 978-0-691-11975-5.

Adam N. Glynn. What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment. *Public Opinion Quarterly*, 77(S1):159–172, January 2013. ISSN 0033-362X. doi: 10.1093/poq/nfs070.

Oded Goldreich and Yair Oren. Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, 7(1):1–32, December 1994. ISSN 1432-1378. doi: 10.1007/BF00195207.

Jonathon R.B. Halbesleben, Anthony R. Wheeler, and M. Ronald Buckley. Understanding pluralistic ignorance in organizations: Application and theory. *Journal of Managerial Psychology*, 22(1):65–83, January 2007. ISSN 0268-3946. doi: 10.1108/02683940710721947.

Daniel C. Hallin. *The Uncensored War: The Media and Vietnam*. University of California Press, Berkeley, 1989. ISBN 978-0-520-06543-7.

Jonas Heese and Gerardo Pérez-Cavazos. The effect of retaliation costs on employee whistleblowing. *Journal of Accounting and Economics*, 71(2-3):101385, April 2021. ISSN 01654101. doi: 10.1016/j.jacceco.2020.101385.

Independent Directors of the Board of Wells Fargo & Company. Sales Practices Investigation Report. Technical report, Independent Directors of the Board of Wells Fargo & Company, April 2017.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6):922–946, December 2021. ISSN 0021-9916. doi: 10.1093/joc/jqab034.

Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, 47 (4):2025–2047, June 2013. ISSN 1573-7845. doi: 10.1007/s11135-011-9640-9.

Timur Kuran. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, Cambridge, Massachusetts London, England, 1997. ISBN 978-0-674-70758-0.

Joseph G Lehman. 'The Overton Window': Made in Michigan. *Viewpoint on public issues*, 2010(19):2, July 2010. ISSN 1093-2240.

Francisco J. León-Medina, Jordi Tena-Sánchez, and Francisco J. Miguel. Fakers becoming believers: How opinion dynamics are shaped by preference falsification, impression management and coherence heuristics. *Quality & Quantity*, 54(2):385–412, April 2020. ISSN 1573-7845. doi: 10.1007/s11135-019-00909-2.

Mark Edward Lewis. *The Early Chinese Empires: Qin and Han*. History of Imperial China. Belknap Press of Harvard University Press, Cambridge, Mass., 1. harvard univ. press paperback ed edition, 2010. ISBN 978-0-674-05734-0 978-0-674-02477-9.

Alessandra F. Lütz and Lucas Wardil. The evolution of pluralistic ignorance. *Physica A: Statistical Mechanics and its Applications*, 647:129920, August 2024. ISSN 0378-4371. doi: 10.1016/j.physa.2024.129920.

Colleen McClain, Regina Widjaya, Gonzalo Rivero, and Aaron Smith. The Behaviors and Attitudes of U.S. Adults on Twitter: A minority of Twitter users produce a majority of tweets from U.S. adults, and the most active tweeters are less likely to view the tone or civility of discussions as a major problem on the site. Technical report, Pew Research Center, 2021.

Patrick McCorry, Siamak F. Shahandashti, and Feng Hao. A Smart Contract for Boardroom Voting with Maximum Voter Privacy. In Aggelos Kiayias, editor, *Financial Cryptography and Data Security*, pages 357–375, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70972-7. doi: 10.1007/978-3-319-70972-7_20.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3):pgaf062, March 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf062.

Stephen Morris and Hyun Song Shin. Social Value of Public Information. *American Economic Review*, 92(5):1521–1534, November 2002. ISSN 0002-8282. doi: 10.1257/000282802762024610.

Navex. ECI's 2023 Global Business Ethics Survey Reveals Harsh Realities About E&C Programs | NAVEX. https://www.navex.com/en-us/blog/article/ecis-2023-global-business-ethics-survey-reveals-harsh-realities-about-ec-programs/, January 2024.

U.S. NLRB. Conduct Elections | National Labor Relations Board. https://www.nlrb.gov/about-nlrb/what-we-do/conduct-elections, 2026a.

U.S. NLRB. Your Right to Form a Union | National Labor Relations Board. https://www.nlrb.gov/about-nlrb/rights-we-protect/the-law/employees/your-right-to-form-a-union, 2026b.

Elisabeth Noelle-Neumann. The Spiral of Silence a Theory of Public Opinion. *Journal of Communication*, 24(2):43–51, June 1974. ISSN 0021-9916, 1460-2466. doi: 10.1111/j.1460-2466.1974.tb00367.x.

U.S. Government Accountability Office. Whistleblowers: Office of Special Counsel Should Require Information on the Probationary Status of Whistleblowers | U.S. GAO. https://www.gao.gov/products/gao-20-436, May 2020.

Thomas R. Palfrey and Howard Rosenthal. Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics*, 24(2):171–193, July 1984. ISSN 0047-2727. doi: 10.1016/0047-2727(84)90023-9.

Deborah A Prentice and Dale T Miller. Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm. *Journal of Personality and Social Psychology*, 64(2):243–256, 1993.

Deborah A. Prentice and Dale T. Miller. Pluralistic Ignorance and the Perpetuation of Social Norms by Unwitting Actors. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 28, pages 161–209. Academic Press, January 1996. doi: 10.1016/S0065-2601(08)60238-5.

Ramsey M. Raafat, Nick Chater, and Chris Frith. Herding in humans. *Trends in Cognitive Sciences*, 13(10):420–428, October 2009. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2009.08.002.

Sean Rad, Todd M Carrico, and Kenneth B Hoskins. Matching process system and method, August 2017.

Victor Shoup. Practical Threshold Signatures. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, and Bart Preneel, editors, *Advances in Cryptology — EUROCRYPT 2000*, volume 1807, pages 207–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000. ISBN 978-3-540-67517-4. doi: 10.1007/3-540-45539-6_15.

Greg Simons. Swedish Media, Fundamental Values and the Opinion Corridor in the 2018 Election. *Etkileşim*, 1(8): 12–34, October 2021. ISSN 2636-7955, 2636-8242. doi: 10.32739/etkilesim.2021.4.8.136.

David Smerdon, Theo Offerman, and Uri Gneezy. 'Everybody's doing it': On the persistence of bad social norms. *Experimental Economics*, 23(2):392–420, June 2020. ISSN 1573-6938. doi: 10.1007/s10683-019-09616-z.

Cooper Smout, Dawn Liu Holford, Kelly Garner, Ruddy Manuel Illanes Beyuma, Paula Andrea Martinez, Megan Ethel Janine Campbell, Ibrahim Khormi, Dylan G. E. Gomes, Johanna Margarete Marianne Bayer, Claire Bradley, Antonio Schettino, Nami Sunami, Sina L. Mansour, and Luis Pedro Coelho. An open code pledge for the neuroscience community, December 2021.

Spartacus.app. Spartacus.app. https://spartacus.app, 2025.

Jennifer G. Steiner, B. Clifford Neuman, and Jeffrey I. Schiller. Kerberos: An Authentication Service for Open Network Systems. *Usenix Winter*, pages 191–202, 1988.

Nick Szabo. Formalizing and Securing Relationships on Public Networks. *First Monday*, 2(9), September 1997. ISSN 13960466. doi: 10.5210/fm.v2i9.548.

Alexander Tabarrok. A simple model of crime waves, riots, and revolutions. *Atlantic Economic Journal*, 25(3):274–288, September 1997. ISSN 1573-9678. doi: 10.1007/BF02298409.

Alexander Tabarrok. The private provision of public goods via dominant assurance contracts. *Public Choice*, 96(3): 345–362, September 1998. ISSN 1573-7101. doi: 10.1023/A:1004957109535.

U.S. Department of Justice. Wells Fargo Agrees to Pay $3 Billion to Resolve Criminal and Civil Investigations into Sales Practices Involving the Opening of Millions of Accounts without Customer Authorization. https://www.justice.gov/archives/opa/pr/wells-fargo-agrees-pay-3-billion-resolve-criminal-and-civil-investigations-sales-practices, February 2020.

U.S. Occupational Safety and Health Administration. US Department of Labor orders Wells Fargo to pay more than $22M for retaliating against executive that alleged financial misconduct. https://www.osha.gov/news/newsreleases/osha-national-news-release/20220901, September 2022.

U.S. Securities and Exchange Comission. Amendments To Rules On Shareholder Proposals. https://www.sec.gov/rules-regulations/1998/05/amendments-rules-shareholder-proposals, 1998.

U.S. Securities and Exchange Comission. Modernization of Beneficial Ownership Reporting. https://www.sec.gov/rules-regulations/2023/10/33-11180, 2023.

Anna Vacca, Andrea Di Sorbo, Corrado A. Visaggio, and Gerardo Canfora. A systematic literature review of blockchain and smart contract development: Techniques, tools, and open challenges. *Journal of Systems and Software*, 174: 110891, April 2021. ISSN 01641212. doi: 10.1016/j.jss.2020.110891.

Wait Until the 8th. Wait Until the 8th: How to Start. https://www.waituntil8th.org/resources, 2026.

Stanley L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965. ISSN 0162-1459. doi: 10.1080/01621459.1965.10480775.

Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, April 2002. doi: 10.1073/pnas.082090499.

Jun-Wu Yu, Guo-Liang Tian, and Man-Lai Tang. Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, 67(3):251–263, April 2008. ISSN 1435-926X. doi: 10.1007/s00184-007-0131-x.

Robertas Zubrickas. The provision point mechanism with refund bonuses. *Journal of Public Economics*, 120:231–234, December 2014. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2014.10.006.

A Proofs and Derivations

The main text centers a two-type endogenous-safety baseline in which equilibrium is a fixed point in anticipated participation q. This appendix records supporting technical ingredients: equilibrium multiplicity and selection, comparative statics derivations for the reduced-form benchmark (Appendix C), and the fixed-point arguments behind the safety-in-numbers extension.

A.1 Constructive characterization and finite-grid existence of the selected-equilibrium construction

Fix threshold T and a candidate monotone cutoff function $x(m;T)$ for supportive agents. Conditional on the latent support state θ, the probability that another eligible agent signs is

$$\alpha(\theta;T) = \theta \int \left[1 - F_{x|z=1}(x(m;T))\right] f(m \mid \theta)\, dm,$$

so $M_{-i} \mid \theta \sim \text{Binomial}(N-1, \alpha(\theta;T))$. For a supportive agent with signal m_i, define the posterior-weighted exact probabilities

$$piv(m_i;T) = E[\Pr(M_{-i} = T-1 \mid \theta) \mid m_i],$$

$$p_+(m_i;T) = E[\Pr(M_{-i} \geq T-1 \mid \theta) \mid m_i], \qquad p_0(m_i;T) = E[\Pr(M_{-i} \geq T \mid \theta) \mid m_i].$$

Then the exact payoff gain from signing is

$$\Delta U_i(x_i, m_i;T) = piv(m_i;T)\bar{v} + p_+(m_i;T)x_i + p_0(m_i;T)s(\tau) + \eta,$$

so the selected-equilibrium cutoff identity is, whenever $p_+(m;T) > 0$,

$$x^*(m;T) = -\frac{piv(m;T)\bar{v} + p_0(m;T)s(\tau) + \eta}{p_+(m;T)}.$$

Lemma 1 (Best response is a cutoff in reputational type)**.** *Fix threshold T and any signal-dependent conjecture about others' play that induces posterior-weighted objects $piv(m;T)$, $p_+(m;T)$, and $p_0(m;T)$. If $p_+(m;T) > 0$ on the signal support, then for each signal realization m the payoff gain $\Delta U_i(x_i, m;T)$ is affine and strictly increasing in x_i. Hence a supportive agent's best response is a cutoff rule in reputational type, given by* (8)*.*

Proof. For fixed m, the payoff gain is

$$\Delta U_i(x_i, m;T) = p_+(m;T)x_i + \left[piv(m;T)\bar{v} + p_0(m;T)s(\tau) + \eta\right].$$

The slope with respect to x_i is $p_+(m;T) > 0$, so the sign of $\Delta U_i(x_i, m;T)$ changes at most once as x_i rises. The unique indifference point is therefore the cutoff in (8). □

To formalize the selected-equilibrium construction used in the numerics, discretize the signal space at

grid points $m_1 < \cdots < m_K$ and impose a cutoff bound $X > 0$. Let

$$C_K(X) = \{x \in [-X, X]^K : x_1 \geq \cdots \geq x_K\}$$

denote the bounded monotone cone of cutoff vectors. Given $x \in C_K(X)$, let $\widetilde{B}_{T,\varepsilon,X}(x)$ be the raw best-response vector obtained by evaluating (8) on the grid using the same quadrature rule, denominator trimming $p_+ \mapsto \max\{p_+, \varepsilon\}$, and clipping to $[-X, X]$ as in the numerical solver, where $\varepsilon > 0$ is fixed. This trim stabilizes grid points where the raw denominator would otherwise vanish; at such points $piv(m_k; T) = p_0(m_k; T) = 0$ and the exact signing decision is governed directly by the sign of η. Let $\Pi_{C_K(X)}$ denote the weighted least-squares isotonic projection onto $C_K(X)$ and define the projected operator

$$B_{T,\varepsilon,X}(x) \equiv \Pi_{C_K(X)}\big(\widetilde{B}_{T,\varepsilon,X}(x)\big).$$

Proposition 7 (Finite-grid existence of the selected-equilibrium construction)**.** *Suppose $F_{x|z=1}$ is continuous and the signal density $f(m \mid \theta)$ is continuous in (m, θ). Then, for fixed T, $\varepsilon > 0$, $X < \infty$, and grid $\{m_k\}_{k=1}^K$, the projected operator $B_{T,\varepsilon,X}$ is continuous and maps $C_K(X)$ into itself. Since $C_K(X)$ is nonempty, compact, and convex, Brouwer's fixed-point theorem implies that $B_{T,\varepsilon,X}$ admits at least one fixed point $x^{*,K}(T) \in C_K(X)$.*

Proof. Continuity of $F_{x|z=1}$ implies continuity of the supportive signing probabilities $1 - F_{x|z=1}(x_k)$ in the cutoff vector. The quadrature map from these probabilities to $\alpha(\theta; T)$ is linear and therefore continuous. Binomial point masses and tails are continuous in α, so the posterior-weighted objects $piv(m_k; T)$, $p_+(m_k; T)$, and $p_0(m_k; T)$ vary continuously with the cutoff vector. The denominator trim $\max\{p_+(m_k; T), \varepsilon\}$ and the coordinatewise clip to $[-X, X]$ preserve continuity, so the raw update $\widetilde{B}_{T,\varepsilon,X}$ is continuous. Weighted isotonic projection onto a closed convex set is continuous, hence $B_{T,\varepsilon,X}$ is continuous. By construction $B_{T,\varepsilon,X}(x) \in C_K(X)$ for every $x \in C_K(X)$. The set $C_K(X)$ is a closed bounded convex subset of $\mathbb{R}^K$, so Brouwer yields a fixed point. □

This does not prove uniqueness of the continuum selected equilibrium, and the paper does not claim such a theorem. What it does prove is the part needed for the manuscript's numerical core: the exact best response remains a cutoff in reputational type, and the projected monotone fixed-point problem solved in the numerical implementation is well posed on every finite grid. Under standard mesh-refinement and dominated-convergence arguments, convergent subsequences of these grid fixed points inherit the continuum cutoff identity (8) at points where $p_+(m; T) > 0$; this is the numerical sense in which the paper's selected-equilibrium construction should be read.

A.2 Reduced-form benchmark derivation

This subsection records the reduced-form benchmark derivation in the paper's notation. The mapping back to the baseline primitives is: $x_i^{img} = e_i - \alpha_i g(q)$, $x^{info} = 0$ (absorbed into the expressive-benefit term), and $s^{img} + s^{info} = s(\tau)$. The reduced-form type is therefore $x_i = x_i^{img} + x^{info}$ and the reduced-form stigma is $s = s^{img} + s^{info}$.

A full Bayesian Nash equilibrium would require agents to form strategies based on their two-dimensional type (v_i, x_i). Agent i chooses $a_i = 1$ if $E[U_i(a_i = 1)|v_i, x_i] \geq E[U_i(a_i = 0)|v_i, x_i]$. Comparing the payoffs from Table 1, a supportive agent with $v_i > 0$ signs if:

$$p_{succ|i}\left(v_i + x_i^{img} + x^{info}\right) \geq p_{succ|\neg i}\left(v_i - (s^{img} + s^{info})\right)$$

where $p_{succ|i}$ is the probability of success if i signs, and $p_{succ|\neg i}$ is the probability of success if i does not sign (i.e., success is due to others), and $p_{succ|i} > p_{succ|\neg i}$. This can be rewritten considering agent i's pivotality. Let $P(\text{pivotal}_i)$ be the probability that $M_{-i} = T - 1$. Then the condition to sign is:

$$P(\text{pivotal}_i)v_i + P(M_{-i} \geq T - 1)(x_i^{img} + x^{info}) + P(M_{-i} \geq T)(s^{img} + s^{info}) \geq 0$$

This simplifies to $P(\text{pivotal}_i)v_i + p_{succ|i}(x_j \geq \sigma^*(v_j, x_j))(x_i^{img} + x^{info}) + p_{succ|\neg i}(x_j \geq \sigma^*(v_j, x_j))(s^{img} + s^{info}) \geq 0$, where $\sigma^*(\cdot)$ denotes the symmetric equilibrium strategy mapping from a player's private information (v_i, x_i) to an action $a_i \in \{0, 1\}$, and $p_{succ|i}$ and $p_{succ|\neg i}$ are the probabilities that at least $T - 1$ or T others sign based on their own (potentially complex) strategies $\sigma^*(v_j, x_j)$. Non-supportive agents with $v_j \leq 0$ choose not to sign in the baseline model, so the relevant heterogeneity for participation is the distribution of x_j among supportive agents. Solving for such a multi-dimensional equilibrium strategy $\sigma^*(v_j, x_j)$ is complex.

For benchmark comparison, the paper also studies a reduced-form large-population signing model. Instead of solving the exact finite-N game over the two-dimensional type (v_i, x_i), Proposition 9 approximates the signing rule by a cutoff in the reduced-form type x_i and treats information-based reputation (x^{info}) and stigma components (s^{img}, s^{info}) as common across agents. This is an approximation rather than an exact reduction: when N is small or material stakes v_i are large, the exact best response can depend on both v_i and x_i.

Lemma 2 (Order of the omitted pivotal term)**.** *Suppose $|v_i| \leq \bar{v}$ for all supportive agents and the equilibrium signing probability of another eligible agent satisfies $q(x^*) \in [\epsilon, 1 - \epsilon]$ for some fixed $\epsilon \in (0, 1/2)$. Then in the finite-N game,*

$$P(pivotal_i) = \Pr(M_{-i} = T - 1) = O\Big((N - 1)^{-1/2}\Big), \qquad |P(pivotal_i)v_i| = O\Big((N - 1)^{-1/2}\Big).$$

Moreover, a standard local-limit approximation for the binomial distribution gives

$$P(pivotal_i) = \frac{1}{\sqrt{2\pi(N-1)q(x^*)(1-q(x^*))}} \exp\left(-\frac{\big(T - 1 - (N-1)q(x^*)\big)^2}{2(N-1)q(x^*)(1-q(x^*))}\right) + o\Big((N-1)^{-1/2}\Big),$$

so the omitted material-benefit term is uniformly second-order whenever $q(x^*)$ stays away from 0 and 1.

Sketch of justification. In the finite-N bridge, $M_{-i} \sim \text{Binomial}(N - 1, q(x^*))$, so $P(\text{pivotal}_i)$ is a single binomial point mass. Standard local-limit approximations for binomial probabilities yield the displayed expansion uniformly when $q(x^*) \in [\epsilon, 1 - \epsilon]$. Multiplying by the bounded valuation $|v_i| \leq \bar{v}$ gives $|P(\text{pivotal}_i)v_i| = O((N - 1)^{-1/2})$. □

This lemma clarifies how Proposition 9 should be read. The single-cutoff representation is most appropriate

in large eligible populations, or when material stakes v_i are modest relative to stigma and esteem incentives, because then the omitted pivotal term is second-order. In smaller groups or in high-stakes settings, the exact finite-N strategy can depend materially on both v_i and x_i and the reduced-form cutoff should be viewed as a rougher guide. For illustration, when $q \approx 0.5$ and T is near the midpoint, the exact pivotal probability is about 0.08 at $N = 100$ and about 0.25 at $N = 10$. The familiar paradox-of-voting logic points in the same direction: collective participation often persists even when pivotality is small, which makes reputational and expressive motives first-order in large populations (Downs 1957).

Under this reduced-form approximation, agent i chooses $a_i = 1$ (Sign) if the expected payoff from signing, focusing on the image and information components relevant to the act of signing itself, is non-negative. If agent i signs, they receive $x_i^{img} + x^{info}$ only when the contract succeeds (which happens with probability $p_{succ|i}(x^*)$ if i signs and others use cutoff x^*). If the agent does not sign, they incur stigma $s^{img} + s^{info}$ if the contract succeeds (with probability $p_{succ|\neg i}(x^*)$). In addition, signing confers a warm-glow or commitment utility η regardless of the contract's outcome. Thus, agent i signs if:

$$p_{succ|i}(x^*)(x_i^{img} + x^{info}) + p_{succ|\neg i}(x^*)(s^{img} + s^{info}) + \eta \geq 0$$

To keep the image- vs. information-based channels explicit while matching the reduced-form notation used elsewhere, use the decomposition $x_i = x_i^{img} + x^{info}$ and $s = s^{img} + s^{info}$. The participation condition becomes $p_{succ|i}(x^*)x_i + p_{succ|\neg i}(x^*)s(\tau) + \eta \geq 0$, implying a cutoff rule: sign if $x_i \geq -\frac{p_{succ|\neg i}(x^*)}{p_{succ|i}(x^*)}s(\tau) - \frac{\eta}{p_{succ|i}(x^*)}$. The equilibrium cutoff x^* for the reduced-form type x_i must therefore satisfy:

$$x^* = -\frac{p_{succ|\neg i}(x^*)}{p_{succ|i}(x^*)}s(\tau) - \frac{\eta}{p_{succ|i}(x^*)} \tag{4}$$

where $p_{succ|i}(x^*)$ is the probability that at least $T - 1$ of the other $N - 1$ eligible agents sign. Since each of the $N - 1$ other eligible agents is supportive with probability π, and conditional on support has type x_j drawn from $F_{x|z=1}$, the unconditional probability that any given other agent signs under cutoff x^* is $q(x^*) \equiv \pi[1 - F_{x|z=1}(x^*)]$. Therefore $p_{succ|i}(x^*)$ is given by:

$$p_{succ|i}(x^*) = \sum_{k=T-1}^{N-1} \binom{N-1}{k} q(x^*)^k \big(1 - q(x^*)\big)^{N-1-k}. \tag{5}$$

Similarly, the probability of success if i does not sign is

$$p_{succ|\neg i}(x^*) = \sum_{k=T}^{N-1} \binom{N-1}{k} q(x^*)^k \big(1 - q(x^*)\big)^{N-1-k}. \tag{6}$$

Here $q(x) = \pi[1 - F_{x|z=1}(x)]$ is the implied signing probability of an arbitrary other eligible agent. Therefore, the equilibrium cutoff x^* solves (4), with $p_{succ|i}(\cdot)$ and $p_{succ|\neg i}(\cdot)$ defined in (5) and (6).

A.3 Proof of Proposition 9

Proof. Fix the distribution $F_{x|z=1}$ and define

$$H(x) \;\equiv\; x\, p_{succ|i}(x) + s(\tau)\, p_{succ|\neg i}(x) + \eta,$$

where $p_{succ|i}(x)$ and $p_{succ|\neg i}(x)$ are given by Eqs. (5)–(6) with $q(x) \equiv \pi[1 - F_{x|z=1}(x)]$:

$$p_{succ|i}(x) \;=\; \sum_{k=T-1}^{N-1} \binom{N-1}{k} q(x)^k \,[1-q(x)]^{\,N-1-k}.$$

and

$$p_{succ|\neg i}(x) \;=\; \sum_{k=T}^{N-1} \binom{N-1}{k} q(x)^k \,[1-q(x)]^{\,N-1-k}.$$

Continuity and monotonicity. Since $F_{x|z=1}$ is continuous, so is $q(\cdot)$, hence $p_{succ|i}(\cdot)$ and $p_{succ|\neg i}(\cdot)$ are finite sums of continuous functions and therefore continuous in x. Moreover, $q(x) = \pi[1 - F_{x|z=1}(x)]$ is nonincreasing in x because $F_{x|z=1}$ is nondecreasing, and each binomial tail probability is nondecreasing in q. Thus $p_{succ|i}(\cdot)$ and $p_{succ|\neg i}(\cdot)$ are nonincreasing in x, and H is continuous on $\mathbb{R}$.

Boundary signs. By assumption, $H(0) \geq 0$ and $H(x) < 0$ for some sufficiently negative x. A sufficient primitive condition is that the support of $F_{x|z=1}$ extends far enough to the left, relative to $s(\tau)$ and η, that the linear term $x\, p_{succ|i}(x)$ dominates the positive terms for large negative x.

Existence. By the intermediate value theorem there exists $x^* \in \mathbb{R}$ with $H(x^*) = 0$. By the positivity assumption, $p_{succ|i}(x^*) > 0$, so the ratio form below is well defined. Thus

$$x^* \;=\; -\frac{p_{succ|\neg i}(x^*)}{p_{succ|i}(x^*)}\, s(\tau) - \frac{\eta}{p_{succ|i}(x^*)},$$

which is Eq. (4). Finally, $x^* \leq 0$ because $H(0) = s(\tau)\, p_{succ|\neg i}(0) + \eta \geq 0$ (when $s(\tau) > 0$ and $\eta \geq 0$) and $H(x) < 0$ for sufficiently negative x; if $s(\tau) > 0$ and $p_{succ|\neg i}(0) > 0$, then $x^* < 0$. This establishes existence of a symmetric cutoff equilibrium and the sign property stated in Proposition 9. □

A.4 Proof of Proposition 6

Proof. Part (i): The participation map $S(q;T) = \pi[(1-\lambda)(1 - F_e(e^*_L(q;T))) + \lambda(1 - F_e(e^*_H(q;T)))]$ is proportional to π, so $S(q;T;\pi') \geq S(q;T;\pi)$ whenever $\pi' \geq \pi$. Under the highest-stable-fixed-point selection, increasing the map pointwise raises the selected fixed point.

Part (ii): $L_T(\pi) = P(M \geq T \mid q^H(T;\pi))$ is the upper tail of Binomial$(N, q^H(T;\pi))$, which is nondecreasing in $q^H(T;\pi)$ and therefore nondecreasing in π.

Part (iii): Since $L_T(\pi)$ is nondecreasing in π and not constant almost everywhere (for example, $L_T(0) = 0$ whenever $T \geq 1$), and the success event has probability strictly between 0 and 1, standard Bayesian monotone-likelihood arguments give $E[\pi \mid M \geq T] > E[\pi]$. □

A.5 Baseline Multiplicity and Equilibrium Selection

As noted in Section 4.3, the baseline participation map $S(\cdot; T)$ can admit multiple fixed points because safety in numbers creates strategic complementarities. In the calibrated range, the typical pattern is three fixed points: a low-participation equilibrium near $q = 0$, an unstable intermediate crossing, and a high (participatory) equilibrium at an interior $q > 0$. Figure 6 confirms this pattern. Panel (a) shows that iteration from the optimistic initial condition ($q = \pi$) converges to the high equilibrium at every threshold, while iteration from the pessimistic initial condition ($q \approx 0$) converges to the low equilibrium, under which the contract fails with high probability and no coalition is publicly revealed. Panel (b) shows that the grid-based fixed-point count is two or three throughout the threshold range.

The conditional design objective $\Psi_H(T)$ reported in the main text uses the highest stable fixed point $q^H(T)$. Focusing on the high equilibrium is without loss of generality for the design problem: under the low equilibrium the contract fails with high probability and the design value is near zero everywhere, so the nontrivial threshold-choice problem lives on $q^H(T)$. The interior optimum and the differential response of the high-vulnerability type are properties of $q^H(T)$ across the full threshold range.

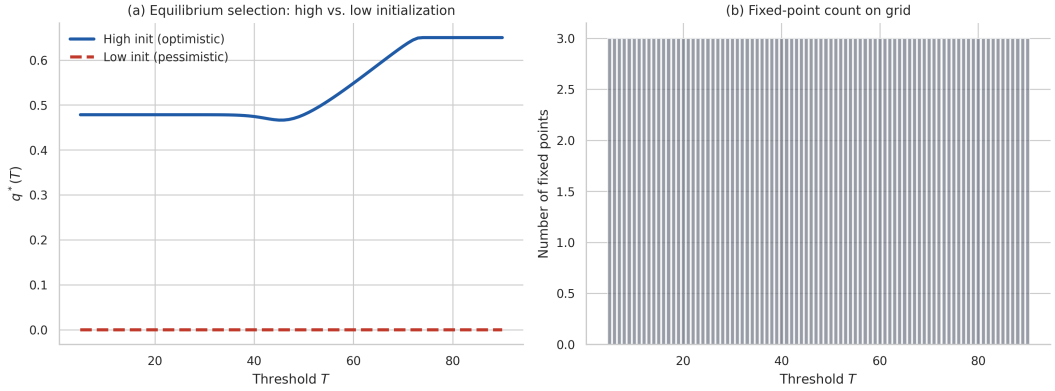


Figure 6: Equilibrium multiplicity in the baseline model. Panel (a) plots $q^H(T)$ and $q^L(T)$ from optimistic and pessimistic initial conditions. Panel (b) shows the number of fixed points detected on a fine grid. Same calibration as Figure 2.

A.6 Formal Derivations of Comparative Statics

Throughout this subsection, the derivative expressions treat s and η as exogenous constants; if one instead allows s or η to depend on $\tau = T/N$, then comparative statics in T or N require chain-rule terms.

Equilibrium Definition and Assumptions: For $T > 1$, the equilibrium cutoff x^* is implicitly defined by:

$$H(x^*, \vartheta) = x^* p_{succ|i}(x^*; \vartheta) + s\, p_{succ|\neg i}(x^*; \vartheta) + \eta = 0$$

where ϑ represents parameters of interest (e.g., s, η, distribution parameters of $F_{x|z=1}$, threshold T, or population size N).

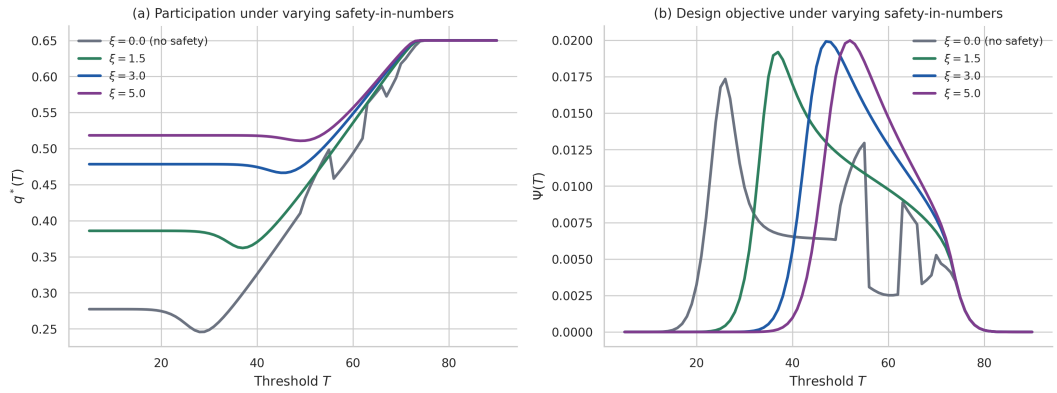


Figure 7: The effect of safety-in-numbers strength. Panel (a) shows equilibrium participation $q^H(T)$ and panel (b) shows the conditional design objective $\Psi_H(T)$ under $\xi \in \{0, 1.5, 3, 5\}$. When $\xi = 0$ (no safety effect), participation is lower and the conditional optimum shifts. Same calibration as Figure 2 except for ξ.

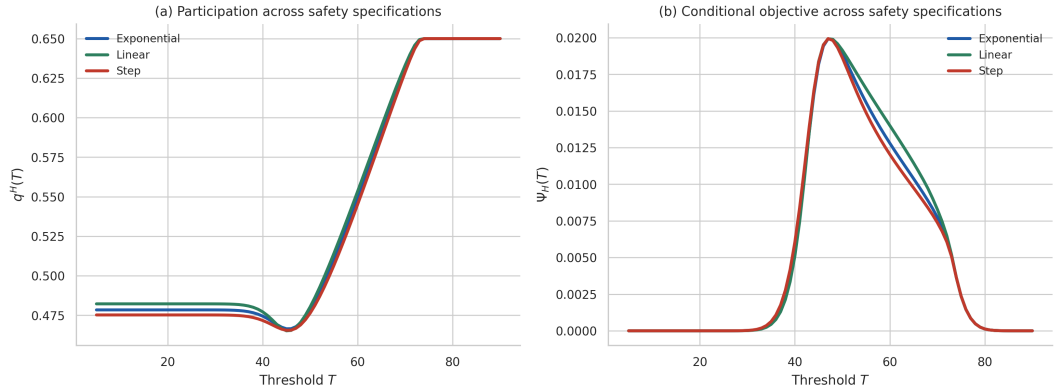


Figure 8: Appendix robustness to alternative safety-in-numbers functions. Panel (a) plots $q^H(T)$ and panel (b) plots $\Psi_H(T)$ under matched exponential, linear, and step specifications for $g(q)$. The hump shape and interior optimum survive across these alternative functional forms.

We apply the Implicit Function Theorem (IFT):

$$\frac{dx^*}{d\vartheta} = -\frac{\partial H/\partial \vartheta}{\partial H/\partial x^*}.$$

Assumption (Stability/Uniqueness): We assume (see Appendix A.3):

$$\frac{\partial H}{\partial x^*} = p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*) > 0.$$

Derivation of Comparative Statics:

1. **Stigma (s):** We have $\frac{\partial H}{\partial s} = p_{succ|\neg i}(x^*)$, hence

$$\frac{dx^*}{ds} = -\frac{p_{succ|\neg i}(x^*)}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)} < 0.$$

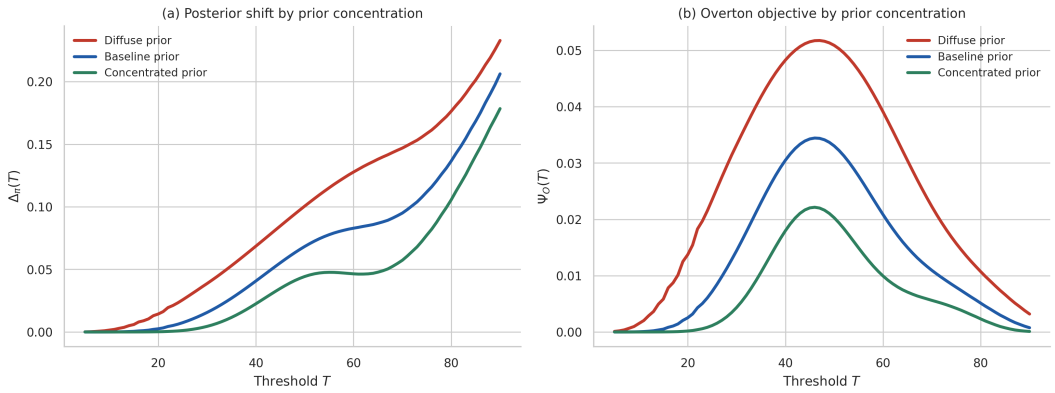


Figure 9: Appendix robustness of Overton dynamics to prior concentration. The mean of the outside-observer prior is fixed at $E[\pi] = 0.65$, while concentration varies across diffuse, baseline, and concentrated Beta priors. More diffuse priors generate larger posterior shifts and larger values of the Overton objective.

Thus, increasing s decreases x^*, increasing participation.

2. **Warm-glow / commitment utility (η):** We have $\frac{\partial H}{\partial \eta} = 1$, hence

$$\frac{dx^*}{d\eta} = -\frac{1}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)} < 0$$

under the stability condition $\partial H/\partial x^* > 0$. Thus, increasing η decreases x^*, increasing participation. This is the second signed comparative-static result alongside $dx^*/ds < 0$.

3. **FOSD Improvement in Distribution $F_{x|z=1}$:** Let an increase in λ_{shift} represent an FOSD improvement (stochastically higher types x_i). We have $\frac{\partial p_{succ|i}}{\partial \lambda_{shift}} > 0$ and $\frac{\partial p_{succ|\neg i}}{\partial \lambda_{shift}} > 0$. Thus,

$$\frac{dx^*}{d\lambda_{shift}} = -\frac{x^*(\partial p_{succ|i}/\partial \lambda_{shift}) + s(\partial p_{succ|\neg i}/\partial \lambda_{shift})}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)}.$$

The sign depends on how the shift changes $p_{succ|i}$ relative to $p_{succ|\neg i}$ through the numerator of the IFT expression.

4. **Threshold (T):** Increasing T makes success harder ($\frac{\partial p_{succ|i}}{\partial T} < 0$ and $\frac{\partial p_{succ|\neg i}}{\partial T} < 0$). Thus,

$$\frac{dx^*}{dT} = -\frac{x^*(\partial p_{succ|i}/\partial T) + s(\partial p_{succ|\neg i}/\partial T)}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)}.$$

The sign depends on how T changes $p_{succ|i}$ relative to $p_{succ|\neg i}$ through the numerator of the IFT expression.

5. **Population Size (N):** Increasing N makes success easier ($\frac{\partial p_{succ|i}}{\partial N} > 0$ and $\frac{\partial p_{succ|\neg i}}{\partial N} > 0$). Thus,

$$\frac{dx^*}{dN} = -\frac{x^*(\partial p_{succ|i}/\partial N) + s(\partial p_{succ|\neg i}/\partial N)}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)}.$$

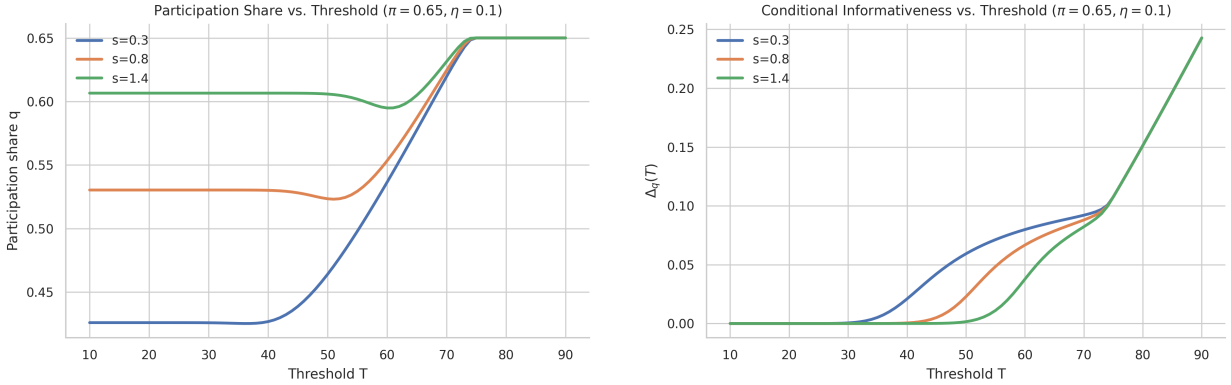
The sign depends on how N changes $p_{succ|i}$ relative to $p_{succ|\neg i}$ through the numerator of the IFT expression.

6. **Cost Convexity (Increase in Costs):** Let λ_{cvx} represent increased costs, causing a stochastic decrease in types ($\frac{\partial p_{succ|i}}{\partial \lambda_{cvx}} < 0$ and $\frac{\partial p_{succ|\neg i}}{\partial \lambda_{cvx}} < 0$). Thus,

$$\frac{dx^*}{d\lambda_{cvx}} = -\frac{x^*(\partial p_{succ|i}/\partial \lambda_{cvx}) + s(\partial p_{succ|\neg i}/\partial \lambda_{cvx})}{p_{succ|i}(x^*) + x^* p'_{succ|i}(x^*) + s\, p'_{succ|\neg i}(x^*)}.$$

The sign depends on how the cost shift changes $p_{succ|i}$ relative to $p_{succ|\neg i}$ through the numerator of the IFT expression.

Summary: Under the maintained assumption of a stable and unique equilibrium ($\partial H/\partial x^* > 0$), $dx^*/ds < 0$ and $dx^*/d\eta < 0$: both stigma and warm-glow unambiguously lower the cutoff and raise participation. The signs for distribution shifts, T, N, and cost convexity depend on how those parameters move $p_{succ|i}$ relative to $p_{succ|\neg i}$ in the numerator of the IFT expression.



(a) Participation share q vs threshold T

(b) Informativeness $\Delta_q(T)$ vs threshold T

Figure 10: Additional numerical comparative-statics illustrations under the baseline calibration $N = 100$, $\pi = 0.65$, $x_i \mid z_i = 1 \sim \mathcal{N}(0, 1)$, and $\eta = 0.1$. Both panels vary T over the same grid as Figure 11 and compare $s \in \{0.3, 0.8, 1.4\}$; panel (b) uses the prior $q \sim \text{Beta}(2, 2)$.

B Trust and Implementation Note

This brief note records how imperfect administrator trust affects participation incentives and complements the implementation risk discussion in the Online Appendix.

B.1 Implementation and trust: crypto escrow

This subsection describes a crypto-escrow implementation that operationalizes the model's conditional-revelation rule without adding a new strategic layer. A campaign specifies threshold T and a deadline;

participants submit encrypted commitments (endorsement or identity claims), and commitments are revealed only if $M \geq T$.

Mechanism sketch— A minimal implementation uses three primitives: (i) binding commitments to endorsements/identities, (ii) an aggregate counter for valid commitments, and (iii) a release rule that unlocks endorsements only on success. The commitments are escrowed *on-chain* (i.e., recorded and verifiable on a blockchain ledger) and remain sealed if the threshold is missed. In this design, the contract acts as a cryptographic referee for timing and threshold execution rather than interpreting participants' preferences.

Trust and verification framing— The central trust question is custody plus conditional execution. In model terms, this is summarized by ρ, the probability the escrow behaves as specified. Under audited smart contracts, open-source verification, and multi-signature controls, ρ can be close to one, so the baseline payoff structure is unchanged in form: the same success condition $M \geq T$ governs realization of success-contingent terms.

Residual risks— Residual risk remains in implementation quality (software bugs, key custody, governance) and in participation dynamics (coordination frictions even when escrow is correct). These are separated in this paper: implementation failures are treated in Online Appendix F, while coordination/network constraints are discussed in Online Appendix G.

Interpretation— Crypto escrow shifts trust from discretionary custodians to verifiable computation while preserving the model's timing and information structure. The mechanism's incentive logic therefore carries over directly, with deviations from $\rho \approx 1$ interpreted as concrete implementation or governance failures.

B.2 Imperfect Trust in the Administrator and Mechanism Integrity

A crucial assumption underpinning an agent's willingness to participate in a social assurance contract, especially when exposure cost $c_i(q) = \alpha_i g(q)$ is large, is their trust in the contract administrator and the overall mechanism to uphold confidentiality if the assurance threshold T is not met. Let ρ be the subjective probability an agent assigns to the administrator being trustworthy and the system functioning as intended (i.e., secrecy is maintained upon failure). Consequently, $1 - \rho$ is the probability of a "betrayal", where a signature is revealed despite $M < T$.

This imperfect trust ($\rho < 1$) alters the expected payoff for an agent i who chooses $a_i = 1$ (Sign) if the contract fails. Instead of receiving $\eta(q)$ with certainty, their expected utility in this scenario becomes:

$$E[U_i(a_i = 1, M < T)] = \rho \cdot \eta(q) + (1 - \rho)\big(\eta(q) - \alpha_i g(q)\big) = \eta(q) - (1 - \rho)\alpha_i g(q)$$

The agent faces an expected additional cost of $(1 - \rho)\alpha_i g(q)$ due to the risk of premature, unsanctioned exposure. The revised payoffs are shown in Table 3:

Table 3: Payoffs for Agent i with Imperfect Trust ($\rho < 1$)

Action a_i	**Contract Fails** ($M < T$)	**Contract Succeeds** ($M \geq T$)
$a_i = 0$ (Not Sign)	0	$\bar{v} - s(\tau)$
$a_i = 1$ (Sign)	$\eta(q) - (1 - \rho)\alpha_i g(q)$	$\bar{v} + e_i - \alpha_i g(q) + \eta(q)$

Impact when success is unlikely— The condition for agent i to prefer $a_i = 1$ when they expect the contract

to fail (i.e., $M_{-i} < T-1$) changes from indifference to a strict disincentive: $a_i = 1$ yields $\eta(q) - (1-\rho)\alpha_i g(q)$ while $a_i = 0$ yields 0. When $\eta(q) < (1-\rho)\alpha_i g(q)$, the all-abstaining profile can re-emerge as a Nash equilibrium, especially if trust ρ is low or exposure costs are high.

Impact on Equilibrium Behavior— Imperfect trust makes signing less attractive. The simplified equilibrium condition in (4) was based on the premise that the net payoff from signing if the contract fails was η. With imperfect trust, this payoff is now $\eta(q) - (1-\rho)\alpha_i g(q)$. The same trust parameter can also capture false-negative failures in the success state: the administrator may fail to publish the coalition even when $M \geq T$, in which case signers bear risk without receiving the coordination benefits of publication.

To connect with the reduced-form benchmark notation, write $c_i \equiv \alpha_i g(q)$ and $x_i = e_i - c_i$. Adapting the simplified style of (4), a supportive agent i with type x_i and specific cost c_i signs if

$$p_{succ|i}(x^*)x_i + p_{succ|\neg i}(x^*)s + \eta - (1 - p_{succ|i}(x^*))(1-\rho)c_i \geq 0.$$

This condition now depends on e_i and c_i separately, not just their difference x_i. This means that the agent's type for the signing decision effectively becomes multi-dimensional (e.g., depending on (e_i, c_i) or even (v_i, e_i, c_i) when the finite-N pivotal term is retained), and a single cutoff x^* for $x_i = e_i - c_i$ no longer fully characterizes the equilibrium strategy in a simple way. A formal analysis of the BNE with imperfect trust would require solving for an equilibrium strategy over a multi-dimensional type space, which is an interesting avenue for future work. However, the clear qualitative implication is that lower trust ρ (i.e., a higher perceived probability $1-\rho$ of premature exposure) acts as an additional expected cost for signing, leading to lower participation rates and a reduced likelihood of the social assurance contract achieving its assurance threshold T. This underscores the importance of the insider attack mitigation strategies discussed in Subsection F.2 (particularly cryptographic guarantees) and the overall need for contract organizers and platforms to establish and maintain a high degree of trustworthiness (i.e., ensure $\rho \approx 1$). Without sufficient trust, the assurance mechanism itself is undermined.

C Finite-N Benchmark with Exact Pivotality

This appendix develops a finite-N benchmark that complements the baseline model by adding exact pivotality and posterior belief updating over a latent support state θ. It abstracts from endogenous safety in numbers—reputational type x_i is a fixed draw rather than a function of anticipated coalition size—and should therefore be interpreted as a companion analysis for settings where individual material stakes or exact pivotality are first-order concerns.

C.1 Model Setup and Selected-Equilibrium Construction

In the benchmark model, a latent support state θ determines the share of supportive agents, supportive agents share a common material value $\bar{v}$, each supportive agent draws a reputational type x_i, and each observes a

private signal $m_i = \theta + \sigma\varepsilon_i$. Conditional on conjectured play, define the exact pivotal and success probabilities

$$piv(m_i;T) = E[\Pr(M_{-i} = T-1 \mid \theta) \mid m_i],$$

$$p_+(m_i;T) = E[\Pr(M_{-i} \geq T-1 \mid \theta) \mid m_i], \qquad p_0(m_i;T) = E[\Pr(M_{-i} \geq T \mid \theta) \mid m_i].$$

The exact payoff gain from signing is

$$\Delta U_i(x_i, m_i; T) = piv(m_i;T)\bar{v} + p_+(m_i;T)x_i + p_0(m_i;T)s(\tau) + \eta, \tag{7}$$

so whenever $p_+(m;T) > 0$ the implied signal-indexed cutoff identity is

$$x^*(m;T) = -\frac{piv(m;T)\bar{v} + p_0(m;T)s(\tau) + \eta}{p_+(m;T)}. \tag{8}$$

Given a cutoff profile $x^*(\cdot;T)$, the state-contingent probability that another eligible agent signs is

$$\alpha(\theta;T) = \theta \int \left[1 - F_{x|z=1}(x^*(m;T))\right] f(m \mid \theta)\, dm. \tag{9}$$

Proposition 8 (Finite-N selected-equilibrium benchmark)**.** *For each fixed threshold T, supportive best responses in the finite-N benchmark are cutoffs in reputational type x_i conditional on the private signal m_i. On any finite signal grid with a bounded cutoff set, the projected monotone best-response operator admits at least one fixed point, yielding the selected cutoff profile used in the paper's numerical benchmark. Section A gives the constructive existence argument.*

C.2 Reduced-Form Scalar-Cutoff Approximation

For transparency, this appendix also records the scalar-cutoff approximation that suppresses exact pivotality and treats the signing decision as one-dimensional in the reduced-form type x_i. If another eligible agent signs with probability $q(x^*) \equiv \pi[1 - F_{x|z=1}(x^*)]$, then the cutoff condition becomes

$$x^* = -\frac{p_{succ|\neg i}(x^*)}{p_{succ|i}(x^*)}s(\tau) - \frac{\eta}{p_{succ|i}(x^*)}, \tag{10}$$

where

$$p_{succ|i}(x^*) = \sum_{k=T-1}^{N-1} \binom{N-1}{k} q(x^*)^k \big(1 - q(x^*)\big)^{N-1-k} \tag{11}$$

and

$$p_{succ|\neg i}(x^*) = \sum_{k=T}^{N-1} \binom{N-1}{k} q(x^*)^k \big(1 - q(x^*)\big)^{N-1-k}. \tag{12}$$

Proposition 9 (Reduced-form scalar-cutoff benchmark)**.** *Suppose $F_{x|z=1}$ is continuous, $H(0) \geq 0$ (which holds whenever $s(\tau) > 0$ or $\eta \geq 0$), and $H(x) < 0$ for some sufficiently negative x, where*

$$H(x) \equiv x\, p_{succ|i}(x) + s(\tau)\, p_{succ|\neg i}(x) + \eta.$$

Then there exists a symmetric cutoff equilibrium x^* solving (4)*, provided $p_{succ|t}(x^*) > 0$ at the equilibrium cutoff; moreover $x^* \leq 0$.*

Under the stability condition $\partial H/\partial x^* > 0$, this benchmark yields the two signed comparative statics: $dx^*/ds < 0$ and $dx^*/d\eta < 0$. Larger abstention stigma and larger commitment utility both lower the relevant cutoff, making signing more attractive throughout the supportive population.

C.3 Overton Dynamics from the Finite-N Benchmark

Proposition 10 (Endogenous Overton-window dynamics from a selected cutoff profile)**.** *For each threshold T, let $x^*(\cdot;T)$ denote a selected finite-N monotone cutoff profile satisfying Proposition 8, and let $\alpha(\theta;T)$ be defined by* (9)*. Then conditional on θ, success has likelihood*

$$L_T(\theta) = \Pr(M \geq T \mid \theta, T) = \sum_{m=T}^{N} \binom{N}{m} \alpha(\theta;T)^m \big(1 - \alpha(\theta;T)\big)^{N-m}.$$

If observers hold prior $\theta \sim \text{Beta}(a, b)$, then observing success yields posterior density proportional to $L_T(\theta)\theta^{a-1}(1-\theta)^{b-1}$. Hence

$$\Delta_\theta(T) \equiv E[\theta \mid M \geq T, T] - E[\theta] > 0$$

whenever success is more likely at higher values of θ and occurs with probability strictly between 0 *and* 1*.*

The finite-N Overton design objective is $\Psi_{FN}(T) \equiv P(M \geq T \mid T, x^*(\cdot;T))\, \Delta_\theta(T)$. Under a benchmark calibration with $N = 100$, $\theta \sim \text{Beta}(13, 7)$, $x_i \mid z_i = 1 \sim \mathcal{N}(0, 1)$, $\bar{v} = 0.5$, $s = 0.8$, $\eta = 0.1$, and private signals $m_i = \theta + 0.1\varepsilon_i$, the objective is hump-shaped and maximized at $T^\star_{FN} = 56$.

C.4 Strategic Opposition in the Finite-N Benchmark

Under the selected finite-N cutoff construction, strategic opposition imposes a durability floor $M_O(\Omega)$ as in Section 5.2. A campaign produces durable success if and only if $M \geq \max\{T, M_O(\Omega)\}$. The architect who values durable belief change should solve

$$\Psi^{dur}_{FN}(T;\Omega) \equiv P(M \geq \max\{T, M_O(\Omega)\} \mid T, x^*(\cdot;T))\, \Delta^{dur}_\theta(T;\Omega),$$

where $\Delta^{dur}_\theta(T;\Omega) \equiv E[\theta \mid M \geq \max\{T, M_O(\Omega)\}, T] - E[\theta]$. In the benchmark calibration, imposing a durability floor $M_O = 60$ shifts the maximizing threshold from $T^\star_{FN} = 56$ to $T^{dur}_{FN} = 60$.

Table 4: Contract structure and population

Notation	Description
N	Size of the eligible population
$\mathcal{P}$	Overall population who could be aware of or affected by the contract
$i = 1, \ldots, N$	Index over eligible agents
$a_i \in \{0, 1\}$	Agent i's action: 1 = Sign, 0 = Not Sign
T	Assurance threshold; integer with $1 \leq T < N$
$\tau \equiv T/N$	Normalized threshold share, $\tau \in [1/N, 1)$
$M = \sum_{i=1}^{N} a_i$	Realized number of signers; success when $M \geq T$

Table 5: Support and types

Notation	Description
π	Support prevalence: share of eligible population that is privately supportive
$z_i \in \{0, 1\}$	Support indicator; non-supportive agents ($z_i = 0$) do not sign
e_i	Idiosyncratic expressive/reputational benefit from signing
F_e	CDF of e_i among supportive agents
$\alpha_i \in \{\alpha_L, \alpha_H\}$	Vulnerability to public exposure; $\Pr(\alpha_i = \alpha_H) = \lambda$
λ	Share of supportive agents who are high-vulnerability

Table 6: Payoff components

Notation	Description
$\bar{v}$	Common material value of success within the supportive population
$g(q)$	Safety-in-numbers function; $g'(q) < 0$
$c_i(q) = \alpha_i g(q)$	Agent i's exposure cost, decreasing in anticipated participation
ξ	Steepness of safety-in-numbers effect: $g(q) = \exp(-\xi q)$ in calibration
$s(\tau)$	Success-contingent abstention stigma
$\eta(q)$	Baseline net signing-specific utility; $\eta(q) = \bar{w}q - k$
$\bar{w}$	Marginal warm-glow from anticipated participation
w	Warm-glow benefit from endorsing
k	Signing friction (time, hassle, worry)

Table 7: Equilibrium objects (baseline)

Notation	Description
q	Anticipated participation share among eligible agents
$q^L(T)$	Low stable fixed point (contract fails with high probability; no public revelation)
$q^U(T)$	Unstable fixed point (tipping point / coordination threshold)
$q^H(T)$	High stable fixed point (participatory equilibrium)
$S(q; T)$	Participation map
$e_L^*(q; T),\ e_H^*(q; T)$	Type-specific cutoffs in expressive benefit
$p_{succ\|i}(q; T)$	Probability of success given agent i signs
$p_{succ\|\neg i}(q; T)$	Probability of success given agent i does not sign

Table 8: Design objectives and Overton window

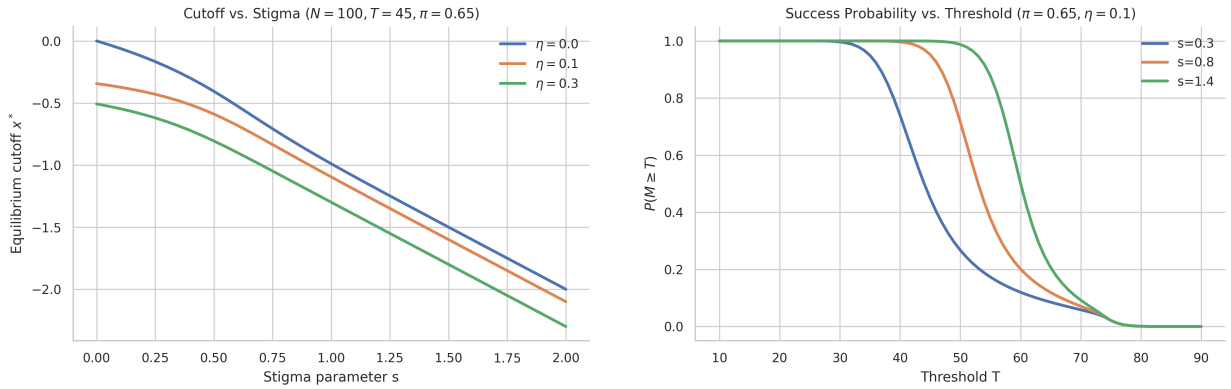
Notation	Description
$\Delta(T, q^H(T))$	Excess revealed coalition share conditional on success
$\Psi_H(T)$	Conditional threshold-design objective: $P(M \geq T) \cdot \Delta(T, q^H(T))$
$\sigma(T)$	Coordination-success probability: $\Pr(k \geq \lceil Nq^U(T) \rceil)$
$\Psi_{EA}(T)$	Ex ante threshold-design objective: $\sigma(T) \cdot \Psi_H(T)$
$\Psi_O(T)$	Baseline Overton design objective: $P(M \geq T) \cdot \Delta_\pi(T)$
G_π	Outside observers' prior over π; when specified, $\pi \sim \text{Beta}(a_\pi, b_\pi)$
$\Delta_\pi(T)$	Posterior shift: $E[\pi \mid M \geq T] - E[\pi]$
$c_{indiv}(\mathbf{b}_0, \mathcal{B})$	Expected unilateral expression cost for focal belief $\mathbf{b}_0$
$\bar{\tau}$	Expression-cost tolerance threshold defining the Overton window
$O_R(\mathcal{B}_t, \bar{\tau})$	The Overton window: set of beliefs with expression cost $\leq \bar{\tau}$

Table 9: Strategic opposition

Notation	Description
Ω	Aggregate opposition capacity
$r(M; \Omega)$	Per-endorser post-disclosure burden; $\partial r/\partial M < 0$, $\partial r/\partial \Omega > 0$
$\bar{c}$	Largest post-disclosure burden compatible with durable expression
$M_O(\Omega)$	Opposition-imposed durability floor
$\Psi^{dur}(T; \Omega)$	Durable-success design objective

Table 10: Auxiliary and extensions

Notation	Description
ρ	Trust in administrator; probability confidentiality is maintained if $M < T$
$y = \delta(M)$	Disclosure statistic under architect-chosen disclosure rule
$\phi(y)$	Continuation anticipated participation after disclosure event y
θ	Latent support state in the finite-N benchmark (Appendix C)
m_i	Private signal; $m_i = \theta + \sigma\varepsilon_i$
x_i, $x^*(m; T)$	Reputational type and signal-indexed cutoff in the finite-N benchmark
$\Psi_{FN}(T)$	Finite-N informational-impact benchmark objective



(a) Cutoff x^* vs stigma s

(b) Success probability $P(M \geq T)$ vs threshold T

Figure 11: Reduced-form benchmark comparative statics under $N = 100$, $\pi = 0.65$, and $x_i \mid z_i = 1 \sim \mathcal{N}(0, 1)$. Panel (a) fixes $T = 45$ and plots x^* against $s \in [0, 2]$ for $\eta \in \{0, 0.1, 0.3\}$. Panel (b) fixes $\eta = 0.1$ and plots $P(M \geq T)$ against T for $s \in \{0.3, 0.8, 1.4\}$.

C.5 Benchmark Figures

D Notation Reference

E Extensions

E.1 Safety in Numbers: Common-Cost Specialization and Uniqueness

The baseline model in the main text already features safety in numbers through exposure cost that falls with anticipated participation. This appendix specializes that mechanism to a common-cost form and conditions on the supportive subpopulation in order to derive a tractable contraction-based uniqueness result and related technical variants (Tabarrok 1997). Let the anticipated participation share be $q \in [0, 1]$, and let

$$c_i(q; \xi) \;=\; \alpha\, g(q; \xi), \qquad g_q(q; \xi) \;=\; -\xi\, h(q), \;\; h(q) > 0, \tag{13}$$

with a maintained common-cost assumption $\alpha > 0$ (so costs shift types uniformly), g continuously differentiable, and $\xi \geq 0$ indexing the steepness of the safety-in-numbers effect (the case $\xi = 0$ turns this channel off). The net type is $x_i = e_i - c_i(q; \xi)$. If $e_i \sim F_e$ with density f_e, the induced distribution of x at anticipated q is

$$F_x(z; q, \xi) \;=\; F_e\big(z + c(q; \xi)\big), \qquad c(q; \xi) \equiv c_i(q; \xi) = \alpha\, g(q; \xi). \tag{14}$$

This appendix specialization suppresses the extensive-margin support parameter π and conditions on the supportive subpopulation to isolate the intensive-margin safety-in-numbers feedback. If one reinstates latent support explicitly, unconditional participation in the eligible population is the product of support prevalence and the supportive-agent signing share studied here.

The cutoff identity remains the right organizing equation but now depends on q (and ξ) through both

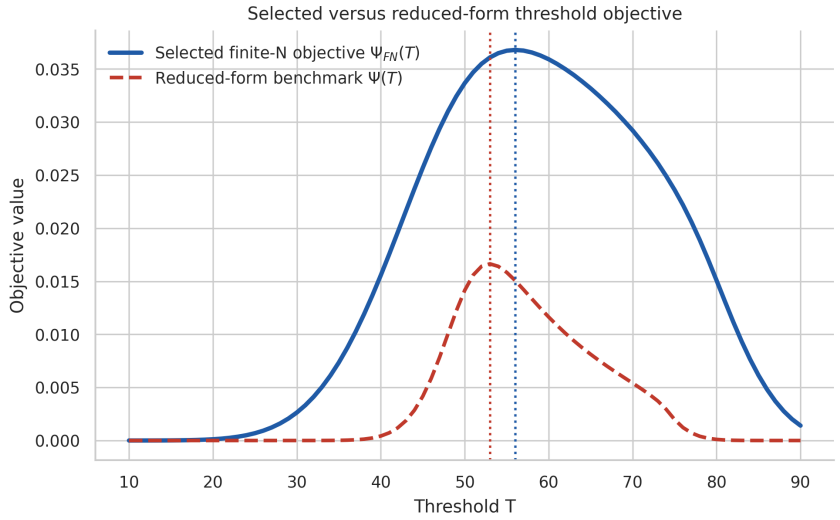


Figure 12: Calibrated threshold-choice comparison in the finite-N benchmark. The solid curve plots $\Psi_{FN}(T)$ under $N = 100$, $\theta \sim \text{Beta}(13, 7)$, $\bar{v} = 0.5$, $s = 0.8$, $\eta = 0.1$, and signal noise $\sigma = 0.1$. The dashed curve plots the matched-prevalence reduced-form comparison. The finite-N benchmark is maximized at $T^{\star}_{FN} = 56$, while the reduced-form comparison is maximized at $T^{\star} = 53$.

success probabilities and the shifted type distribution:

$$H\big(x^*, q, \xi\big) \;\equiv\; x^*\, p_{succ|i}\big(x^*; q, \xi\big) + s(\tau)\, p_{succ|\neg i}\big(x^*; q, \xi\big) + \eta \;=\; 0, \tag{15}$$

where $p_{succ|i}(x; q, \xi)$ and $p_{succ|\neg i}(x; q, \xi)$ are computed as in Eqs. (5)–(6) using $F_x(\cdot; q, \xi)$.

Equilibrium as a fixed point in participation. Given anticipated q, the share who sign is

$$S(q; \xi) \;=\; 1 - F_x\big(x^*(q; \xi); q, \xi\big) \;=\; 1 - F_e\big(x^*(q; \xi) + \alpha\, g(q; \xi)\big), \tag{16}$$

and a rational-expectations equilibrium is a fixed point $q \in [0, 1]$ of

$$q \;=\; S(q; \xi). \tag{17}$$

Differentiating (15) by the Implicit Function Theorem,

$$\frac{dx^*}{dq} \;=\; -\,\frac{x^*\, p_{i,q}(x^*; q, \xi) + s(\tau)\, p_{\neg i,q}(x^*; q, \xi)}{p_{succ|i}(x^*; q, \xi) + x^*\, p_{i,x}(x^*; q, \xi) + s(\tau)\, p_{\neg i,x}(x^*; q, \xi)},$$

where subscripts i, x and i, q denote partial derivatives of $p_{succ|i}$, and likewise for $\neg i$. Differentiating (16) gives

$$S'(q; \xi) \;=\; -\, f_e\big(x^*(q; \xi) + c(q; \xi)\big)\left(\frac{dx^*}{dq} + c_q(q; \xi)\right), \tag{18}$$

where $c_q(q; \xi) = \partial c(q; \xi)/\partial q < 0$. Because $c_q(q; \xi) < 0$, the best–response map $S(\cdot; \xi)$ is increasing whenever the safety–in–numbers effect $-c_q(q; \xi)$ dominates the indirect cutoff response dx^*/dq (standard

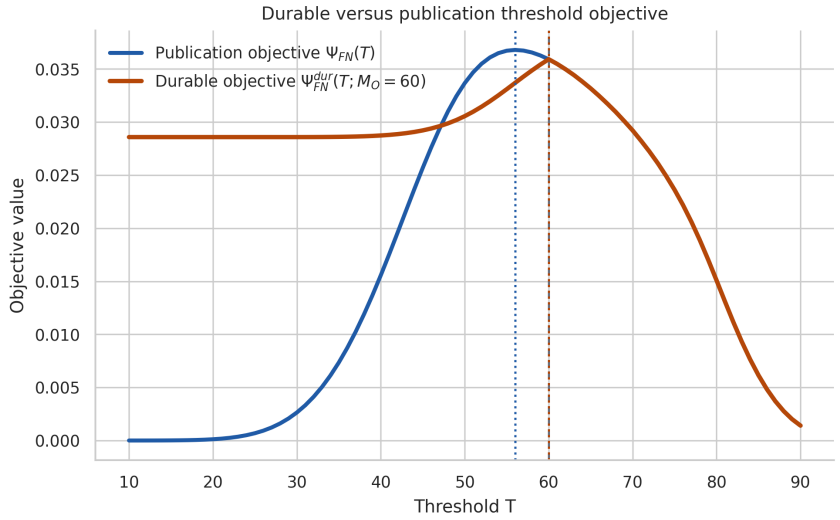


Figure 13: Publication versus durable threshold design in the finite-N benchmark. The blue curve plots $\Psi_{FN}(T)$. The orange curve plots $\Psi_{FN}^{dur}(T;\Omega)$ under a durability floor of $M_O = 60$.

critical–mass logic).

Proposition 11 (Uniqueness via contraction under modest safety in numbers)**.** *Suppose:* (i) f_e *is bounded by* $\bar{f} < \infty$*;* (ii) $p_{succ|i}$ *and* $p_{succ|\neg i}$ *are continuously differentiable with* $p_{succ|i}(x;q,\xi) > 0$ *and*

$$H_x(x,q,\xi) \equiv p_{succ|i}(x;q,\xi) + x\, p_{i,x}(x;q,\xi) + s(\tau)\, p_{\neg i,x}(x;q,\xi) \;\geq\; \underline{h} > 0$$

for all (x,q) *in the relevant domain;* (iii) $|x^*\, p_{i,q} + s(\tau)\, p_{\neg i,q}| \leq \bar{p}_q$ *for a uniform bound* $\bar{p}_q$*; and* (iv) $c_q(q;\xi) \in [-\bar{c}(\xi), 0)$ *with* $\bar{c}(\xi) \equiv \alpha\, \xi\, \bar{h} < \infty$ *for* $\bar{h} \equiv \sup_{q\in[0,1]} h(q)$*. Then every fixed point of* (17) *is unique provided*

$$\bar{f}\left(\bar{c}(\xi) + \frac{\bar{p}_q}{\underline{h}}\right) \;<\; 1 \quad \textit{for all } q \in [0,1]. \tag{19}$$

The condition has a transparent interpretation: $\bar{c}(\xi)$ captures the steepness of the safety–in–numbers effect $-c_q(q;\xi)$; larger values push $S'(q;\xi)$ upward, making multiple equilibria more likely. The term $\bar{p}_q/\underline{h}$ measures how strongly the cutoff responds to anticipated participation via the success probabilities.

Proof.— Fix $q \in [0,1]$ and consider the cutoff identity

$$H(x,q,\xi) \;\equiv\; x\, p_{succ|i}(x;q,\xi) + s(\tau)\, p_{succ|\neg i}(x;q,\xi) + \eta \;=\; 0, \qquad \text{with } p_{succ|i} \text{ and } p_{succ|\neg i} \text{ as defined above.}$$

By assumption, $H_x(x,q,\xi) = p_{succ|i}(x;q,\xi) + x\, p_{i,x}(x;q,\xi) + s(\tau)\, p_{\neg i,x}(x;q,\xi) \geq \underline{h} > 0$ on the relevant domain, so for each q there is a *unique* solution $x^*(q;\xi)$, and the Implicit Function Theorem yields

$$\frac{dx^*}{dq} \;=\; -\frac{H_q(x^*,q,\xi)}{H_x(x^*,q,\xi)} \;=\; -\,\frac{x^*(q;\xi)\, p_{i,q}(x^*;q,\xi) + s(\tau)\, p_{\neg i,q}(x^*;q,\xi)}{p_{succ|i}(x^*;q,\xi) + x^*(q;\xi)\, p_{i,x}(x^*;q,\xi) + s(\tau)\, p_{\neg i,x}(x^*;q,\xi)}. \tag{20}$$

Using $|x^* p_{i,q} + s\, p_{\neg i,q}| \le \bar{p}_q$ and $H_x \ge \underline{h}$, we obtain the bound

$$\left|\frac{dx^*}{dq}\right| \;\le\; \frac{\bar{p}_q}{\underline{h}}. \tag{21}$$

Define the best–response map (Eq. (16))

$$S(q;\xi) \;=\; 1 - F_e\big(x^*(q;\xi) + c(q;\xi)\big),$$

which takes values in $[0,1]$. Differentiating and using the chain rule gives Eq. (18),

$$S'(q;\xi) \;=\; -\,f_e\big(x^*(q;\xi) + c(q;\xi)\big)\left(\frac{dx^*}{dq} + c_q(q;\xi)\right).$$

With $f_e \le \bar{f}$ and $c_q(q;\xi) \in [-\bar{c}(\xi), 0)$, combine (21) to obtain

$$|S'(q;\xi)| \;\le\; \bar{f}\left(\left|\frac{dx^*}{dq}\right| + |c_q(q;\xi)|\right) \;\le\; \bar{f}\left(\frac{\bar{p}_q}{\underline{h}} + \bar{c}(\xi)\right). \tag{22}$$

If the uniform bound

$$\sup_{q\in[0,1]} |S'(q;\xi)| \;<\; 1 \quad \text{holds, i.e.,} \quad \bar{f}\left(\bar{c}(\xi) + \frac{\bar{p}_q}{\underline{h}}\right) \;<\; 1 \;\; \forall\, q,$$

then $S(\cdot;\xi)$ is a strict contraction on the complete metric space $([0,1], |\cdot|)$. By Banach's fixed–point theorem, S has a unique fixed point $q^* \in [0,1]$, which is the unique rational–expectations equilibrium share, and the iterates $q_{t+1} = S(q_t;\xi)$ converge to q^* from any initial condition. This is exactly condition (19) and proves Proposition 11.

Remark.— The lower bound $H_x(x,q,\xi) \ge \underline{h} > 0$ implies that, for each fixed q, the mapping $x \mapsto H(x,q,\xi)$ is strictly increasing, hence the cutoff equation $H(x,q,\xi) = 0$ admits a *unique* solution $x^*(q;\xi)$. Thus the best–response map $S(q;\xi) = 1 - F_e(x^*(q;\xi) + c(q;\xi))$ is well-defined. Moreover, continuity of F_e, $c(\cdot;\xi)$, and $x^*(\cdot;\xi)$ (the latter from the Implicit Function Theorem) implies $S(\cdot;\xi)$ is continuous on $[0,1]$. Therefore a fixed point of $q = S(q;\xi)$ exists by Brouwer's fixed–point theorem even without the contraction condition. Proposition 11 strengthens this to *uniqueness* (and global convergence of iterates) by ensuring $S(\cdot;\xi)$ is a strict contraction on $[0,1]$.

Special case ($\xi = 0$). When $\xi = 0$, $c_q(q;\xi) = 0$ so $S(q;0)$ has no q-feedback and the fixed-point problem (17) collapses to the fixed-cost version of this appendix specialization. The contraction and uniqueness results in this subsection then cease to be informative, while level effects driven by other parameters (e.g., curvature γ that shifts F_x) remain.

Interface with architect-chosen disclosure. If disclosure feeds back into anticipated participation via a continuation map $\phi : [0,1] \to [0,1]$, then $F_x(\cdot;\phi(\cdot),\xi) = F_e(\cdot + c(\phi(\cdot);\xi))$ shifts right as the public signal improves (since $c_q(\cdot;\xi) < 0$), partially offsetting the direct tightening in the continuation cutoffs described in Online Appendix E.2. Net effects are application-specific and determined by the relative steepness of $-c_q(\cdot;\xi)$ and the sensitivity of p_{succ}.

Robustness: two-type α heterogeneity. Relax the maintained common-α assumption by letting $\alpha_i \in \{\alpha_L, \alpha_H\}$ with $\Pr(\alpha_i = \alpha_H) = \lambda$ and $\Pr(\alpha_i = \alpha_L) = 1 - \lambda$, independent of e_i. Costs are $c_i(q;\xi) = \alpha_i g(q;\xi)$ and net types remain $x_i = e_i - c_i(q;\xi)$. Conditional on anticipated q, the induced distribution of x is the mixture

$$F_x(z;q,\xi) \;=\; (1-\lambda)\,F_e\big(z + \alpha_L g(q;\xi)\big) \;+\; \lambda\,F_e\big(z + \alpha_H g(q;\xi)\big), \tag{23}$$

(with density $f_x(z;q,\xi) = (1-\lambda) f_e(z + \alpha_L g(q;\xi)) + \lambda f_e(z + \alpha_H g(q;\xi))$). The cutoff identity (15) and fixed-point equation (17) are unchanged, except that $p_{succ}(x;q,\xi)$ is computed using (23).

For the contraction/bound argument, differentiating $S(q;\xi) = 1 - F_x(x^*(q;\xi);q,\xi)$ yields

$$S'(q;\xi) = -(1-\lambda) f_e\big(x^* + \alpha_L g(q;\xi)\big)\Big(\frac{dx^*}{dq} + \alpha_L g_q(q;\xi)\Big) - \lambda f_e\big(x^* + \alpha_H g(q;\xi)\big)\Big(\frac{dx^*}{dq} + \alpha_H g_q(q;\xi)\Big).$$

If f_e is bounded by $\bar{f}$ and $\bar{\alpha} \equiv \max\{\alpha_L, \alpha_H\}$, then the bound in (22) continues to hold after replacing the safety-in-numbers term $\bar{c}(\xi) = \alpha\,\xi\,\bar{h}$ by $\bar{c}^{mix}(\xi) \equiv \bar{\alpha}\,\xi\,\bar{h}$ (since $|g_q(q;\xi)| = \xi h(q) \le \xi\bar{h}$). Accordingly, the contraction/uniqueness condition (19) remains sufficient under the same replacement $\bar{c}(\xi) \mapsto \bar{c}^{mix}(\xi)$.

E.2 Equilibrium under architect-chosen disclosure

Main-text bridge. The main text treats architect-chosen disclosure only briefly as a secondary design margin. This appendix subsection records a disclosure extension and technical variants.

Coarsening/noise (design lever). Replacing exact running counts by coarse milestones (or noisy statistics) weakens the sensitivity of perceived success to the signal. Formally, one can work with a Lipschitz bound $|p_q| \le \bar{p}_q^{\delta} \le \bar{p}_q$; making disclosure coarser (smaller $\bar{p}_q^{\delta}$) tightens the contraction inequality in Proposition 11, making uniqueness easier to obtain.

Two-stage "early/late" variant— Let a fraction $\lambda \in (0,1)$ decide in Stage 1 under private progress (baseline x_1^* solves $H(x_1^*, q_1) = 0$ with $q_1 = \lambda\,[1 - F_e(x_1^* + c(\lambda))]$). The realized count $S_1 \sim \text{Binomial}\big(\lambda N,\, 1 - F_e(x_1^* + c(\lambda))\big)$ is disclosed via δ, producing public $y_1 = \delta(S_1)$. Stage 2 agents then solve

$$x_2^*(y_1) \;=\; -\frac{p_{succ|\neg i}\big(x_2^*(y_1);\, S^-(y_1)\big)s(\tau) + \eta}{p_{succ|i}\big(x_2^*(y_1);\, S^-(y_1)\big)}, \quad S^-(y_1) \in [0, S_1] \text{ a lower bound implied by } \delta, \tag{24}$$

and, with safety in numbers, evaluate types under $F_x(\cdot;\phi(y_1/N))$.

Multiplicity (diagnostic). If there exists an interior fixed point $\hat{q}$ with $S'(\hat{q}) > 1$, then (generically) there are two additional fixed points, one below and one above $\hat{q}$; the outer two are locally stable and the middle one is unstable (the usual S-shape). Design levers that reduce $|c'(q)|$ at low q (identity escrow until threshold, early-mover insurance, verified aggregate counters) help select the high-participation equilibrium by lowering $S'(q)$ where it is largest.

F Implementation Risks and Mitigations

As emphasized in Section 3, social assurance contracts hinge on conditional revelation and therefore rely on a combination of institutional trust and/or cryptographic enforcement to function as intended. Implementation failures should be analyzed through the lens of confidentiality (preventing premature exposure), integrity (preventing manipulation of eligibility, counts, or release conditions), and availability (ensuring the mechanism can actually reach and execute the threshold decision), and the discussion below maps concrete threats to those three dimensions.

F.1 Sybil Attacks and Ballot-Stuffing

A primary risk is the *ballot-stuffing* or *Sybil* attack (Douceur 2002), where malicious actors generate fake or insincere signatures to falsely achieve the assurance threshold T, leading to premature or misleading publication. Robust identity verification is the key defense. Requiring participants to authenticate via verifiable methods (e.g., organizational email, unique institutional credentials, or digital signatures linked to known identities within the eligible population) significantly hinders the creation of multiple fraudulent entries. Organizational authentication systems like Kerberos offer a model for such verification (Steiner et al. 1988). Additionally, a small, non-monetary stake or commitment, such as a verifiable pledge, could deter frivolous sign-ups.

A distinct challenge arises if legitimate members of the eligible population sign insincerely merely to trigger publication and expose sincere endorsers. Such insincere endorsers might then publicly disavow the statement, claiming they signed only to "out" others. This could be an effective strategy if it is wholly implausible that the insincere endorser could possibly have been sincere. In this case, drawing up a list of likely suspects and excluding them from the contract in the first place would be an effective defense. In the case where it is ambiguous whether an endorser who claims to have signed only to cause the contract to succeed actually did so, this endorser now has their signature indelibly affixed to an allegedly repugnant statement. Whether a *post hoc* claim of insincerity will be remembered as long as an indelible signature on a contract is quite questionable, and should provide a disincentive.

F.2 Insider Attacks and Mechanism Integrity

The integrity of the contract administrator (system or person) is paramount. An insider attack (such as prematurely leaking signatures, falsely claiming T is met, or succumbing to external coercion to reveal identities) or a technical compromise of the platform severely undermines the contract's assurance. Mitigating these risks often involves removing single points of failure and employing robust technical and procedural safeguards.

Distributing trust, for instance, by using multiple independent escrow holders requiring a quorum for action, is one approach. Cryptographic methods, however, should be considered the gold standard for ensuring both secrecy and integrity. Decentralized mechanisms like smart contracts on a blockchain could automate conditional release based on verifiable cryptographic conditions (e.g., T valid digital signatures from institutional certificates), minimizing human intervention (cf. Szabo 1997 for foundational concepts of smart

contracts; for a review, Vacca et al. 2021). Threshold-cryptographic schemes (including threshold signatures) can distribute control so that authentication or release requires cooperation by T participants or by a trustee quorum (Shoup 2000). Furthermore, anonymous credential systems can allow users to prove eligibility without revealing identity to the administrator until publication (Camenisch and Lysyanskaya 2001), and zero-knowledge techniques can, in principle, verify threshold conditions without revealing signer identities (Goldreich and Oren 1994). Related cryptographic literatures offer more decentralized implementation templates. Secure multi-party computation can eliminate the trusted administrator entirely at the cost of heavier communication and computation (Evans et al. 2018), while Open Vote Network-style protocols show how blockchain-based coordination can achieve self-tallying, privacy-preserving voting for small groups (McCorry et al. 2017). While the full development of user-friendly platforms incorporating such advanced cryptographic guarantees for general-purpose social assurance contracts is ongoing, these technologies can offer strong protection against both premature exposure and administrator malfeasance.

Beyond purely technical measures, legal agreements imposing significant penalties for malfeasance can provide a deterrent. Centralized platforms must also cultivate a strong reputation for integrity, potentially bolstered by independent third-party audits and open-source software development. Ultimately, participants' trust ($\rho \approx 1$, as discussed in Section B.2) is essential, and this may rely on a combination of technical assurances, reputational trust in the administrator, and clear governance regarding the contract's terms, including the immutability of the statement being endorsed.

F.3 Cryptographic Failure Modes and Operational Mitigations

Even when cryptography is deployed, failures frequently arise from key management and operational missteps rather than from broken primitives. If encryption or signature keys are generated with weak entropy, stored on exposed endpoints, or reused across contexts, confidentiality and integrity degrade rapidly. Mitigations should focus on hardware-backed key storage where feasible, multi-party or threshold custody for administrator secrets, regular rotation, and strict separation of duties so no single operator can unilaterally exfiltrate material. Access controls should be auditable and least-privilege by default, and recovery procedures should be explicit about how lost keys, compromised devices, or personnel departures are handled without creating backdoors. In environments with limited hardware support, a concrete alternative is to bind custody to multiple independent stewards and require quorum-based approvals for any action that could reveal or alter endorsements.

Credential binding and domain separation are also essential. Without explicit binding of signatures or ciphertexts to the precise statement, contract instance, and disclosure conditions, a valid artifact can be replayed or transplanted into a different setting, undermining integrity. Protocols should enforce context strings and structured prefixes that commit to the statement hash, threshold T, and intended disclosure policy; this ensures that endorsements cannot be repurposed to satisfy a different contract. Encryption should use authenticated modes (AEAD) to prevent malleability and undetected tampering, and verification should reject any artifact that does not authenticate to the expected domain. Freshness and revocation mechanisms should be built in to address replay: for example, include nonces or epochs, maintain revocation lists for compromised credentials, and require the administrator (or quorum of trustees) to demonstrate the current validity set before release. Finally, metadata leakage can break confidentiality even when payloads are encrypted, because

timing, volume, or access patterns can reveal who signed or whether the threshold is near. Mitigations include batching submissions, delaying disclosures to fixed intervals, using mix or proxy relays to hide origin metadata, and ensuring that audit logs are privacy-preserving and do not expose signers before the threshold event.

F.4 Blockchain and Smart-Contract Deployment Risks

If a blockchain is used to automate the threshold condition, the contract inherits platform-level risks that can threaten availability and integrity. Transactions can be reordered or censored in the mempool, and maximal-extractable-value (MEV) dynamics can permit front-running, griefing, or strategic delay of endorsements or release transactions. Mitigations include commit–reveal designs that hide sensitive data until finalized, use of private relays or encrypted mempools, and redundancy across multiple relayers to reduce reliance on a single path to inclusion. Censorship resistance should be treated as an availability requirement: if certain endorsers or release transactions can be blocked, the mechanism may never reach a usable threshold even if signers exist. Similarly, chain reorganizations and probabilistic finality can cause a threshold to appear met and then be rolled back, which can create accidental disclosures or inconsistent publication. A robust design should specify a finality rule (e.g., a minimum confirmation depth or a chain finality gadget) and delay any irreversible disclosure until that rule is met.

Administrative and upgrade keys create a separate integrity vulnerability: if a privileged key can pause, upgrade, or withdraw contract state, then trust shifts from the cryptographic protocol back to administrators. These powers should be minimized, protected by multisignature and timelock controls, and documented transparently so participants can assess the residual trust required. Finally, smart-contract bugs can permanently lock funds or release signatures incorrectly. Mitigations include external audits, conservative on-chain logic, extensive testnet deployments, and a clearly defined emergency procedure that trades off safety and liveness in a controlled way. For example, an emergency pause might prevent premature disclosure at the cost of temporary unavailability, while an explicit rollback or migration path could be used to repair critical flaws without silently altering the endorsement threshold or eligibility rules. These operational choices should be stated ex ante so that participants understand how failures will be handled.

G Network Topology and the Success of Assurance Contracts

The effectiveness of a social assurance contract can also depend on the underlying social network structure of the N eligible agents. While formal modeling of these network effects is left for future research, this section qualitatively discusses several key influences. The discussion below uses the main text's baseline notation whenever possible: anticipated participation q, expressive benefit e_i, vulnerability α_i, and success probabilities $p_{succ|i}(q;T)$.

G.1 Network Effects, Payoffs, and Safety-in-Numbers

Assurance contracts are inherently networked objects: each additional endorser changes not only the probability of success but the meaning and consequences of success. A basic payoff network effect is that expected utility rises as $p_{succ|i}(q;T)$ rises, which is the coordination externality described in Section 3. Yet social assurance contracts also carry legitimacy and safety-in-numbers effects that are distinct from simple success probability. In many settings, the perceived social cost c_i depends on how many *visible* others are willing to endorse, and on whether endorsers span multiple social circles, so c_i can fall with participation even holding $p_{succ|i}(q;T)$ fixed.

Legitimacy effects are related but distinct. A broad and heterogeneous set of endorsers can shift perceived legitimacy more than a narrow, insular coalition. In terms of the Overton window dynamics in Section 7, cross-cutting coalitions can produce a larger belief update $\mathcal{B}_t \to \mathcal{B}_{t+1}$ by reshaping second-order beliefs about who holds the belief $\mathbf{b}_s$. Dense, homophilous networks can mute this effect, while bridging ties can amplify it even at moderate participation levels.

Safety-in-numbers effects also depend on topology. When a contract succeeds, public expression may require coordinated visibility (simultaneous publication, joint announcements, or synchronized events). Networks with strong local clustering and short paths facilitate that coordination, while fragmented networks can make the same threshold T appear weaker, reducing the protective effect of collective expression. Thus the ex ante choice of T interacts with the ex post capacity to reduce c_i at the moment of revelation.

G.2 Coordination Within Clusters and Across Bridges

The presence of cohesive subgroups or dense clusters within the network may facilitate coordination. Pre-existing trust within such a subgroup can bolster individuals' confidence in the integrity of the contract process and their belief that peers within the subgroup will also choose $a_j = 1$ (Sign). This enhanced trust could translate into a higher subjective probability of success, $p_{succ|i}(q;T)$, or a lower perceived individual cost, c_i. Furthermore, an organizer could leverage existing trust links to quietly recruit a core group of endorsers. If this core group is sufficiently large to approach or meet the assurance threshold T, it could catalyze wider participation from the periphery. Strong intra-group ties, therefore, can be instrumental in overcoming the initial coordination hurdle, suggesting that perceived success probabilities depend not only on the global distribution of supportive types but also on local network characteristics that shape beliefs about others' actions.

Conversely, coordination becomes more challenging if privately supportive agents are dispersed across a fragmented network, for instance, in different departments, distinct social circles, or otherwise structurally isolated parts of the community. Such fragmentation can create information barriers, leaving individuals unaware of the extent of shared sentiment or even of the contract's existence. In these scenarios, the role of trusted individuals or communication channels that act as bridges between disconnected segments, or as central hubs, becomes crucial. These network bridges are essential for disseminating information about the contract and for recruiting a broad base of endorsers; without them, $p_{succ|i}(q;T)$ may be severely underestimated by many potential participants. In highly fragmented networks, it might even be necessary to

employ multiple, parallel assurance contracts, perhaps one for each cluster, potentially with an overarching mechanism to combine successes if each meets a local assurance threshold. The design of the assurance threshold T itself could be adapted, for instance, by requiring a vector of participation $T = (T_1, \ldots, T_k)$ from k different groups for publication, ensuring broad representation.

The social insularity of a group also plays a role, which may come down to the question: who has the power to hurt me? In an insular group, the opinions of outsiders may be of little importance. This could mean the people who might potentially levy costs c_i upon contract success are limited in number and therefore easier to reason about. In the case where a group is not insular, and has more links to society at large rather than to each other, we might expect that the assurance threshold T must be very high since "the rest of society" provides a rather large number of people who could contribute to c_i. In insular groups with strong, monolithic cultures or those that heavily censor dissent, individuals who disagree with the prevailing norm may be completely isolated, with no easy way to find like-minded others or gauge latent support. Launching an assurance contract in such environments might necessitate external intervention or anonymous broadcasts, as any known organizer could be targeted before reaching T. In contrast, in more open insular groups where social ties cut across clusters of beliefs, information about a contract could diffuse more freely. A successful contract in such an environment may be more likely to draw endorsers from a diverse cross-section of beliefs, lending greater credibility and broader impact to the revealed consensus and subsequent belief updates.

G.3 Marketing, Seeding, and Common Knowledge

Network structure shapes not only the dynamics of signing but also the marketing and coordination required to get a contract off the ground. A key practical step is *seeding*: recruiting a small set of credible early endorsers who are willing to commit and positioned to diffuse information. In clustered networks, it may be optimal to seed one or two dense communities first, while in modular networks a cross-cluster seed can prevent the contract from being framed as parochial. Early endorsers also shape legitimacy and safety-in-numbers expectations, which can lower c_i even before T is reached.

Coordination failures often arise from deficits of *common knowledge*. It is not enough for individuals to believe that others privately agree; they must believe that others know that others know, and so on. Public signals and shared observability are therefore central. Platform design and outreach that raise the visibility of the contract, and of its progress, can create the layered beliefs required for coordinated action. This links back to Section 3, where individual decisions depend on perceived success probabilities and costs.

Outreach also interacts with risk. In highly adversarial environments, public marketing can expose organizers or early endorsers to retaliation before the contract reaches the safety threshold. In such cases, private invitations within trusted subgroups may be necessary early on, with public disclosure delayed until a robust coalition is secured. Conversely, in environments where stigma is moderate but coordination is hard, broad public outreach and credible third-party amplification may be the dominant factor in overcoming inertia.

G.4 Progress Signals and Coordination Devices

Progress signals are another critical coordination device. Visible counters, progress bars, or periodic updates do more than convey information; they create a feedback loop in which observed momentum increases expected success and reduces perceived risk. In a network context, the value of such signals depends on how widely and credibly they diffuse. A progress bar visible only within a small cluster may amplify participation locally but fail to mobilize other clusters, while a widely visible signal can unify expectations across the network.

The choice between coarse signals (e.g., "50% funded") and granular signals (e.g., the exact number of endorsers) also matters. Granular signals may accelerate growth in dense networks by enabling inference, but they can be counterproductive in sparse or polarized networks if they enable targeted pressure on early supporters or reveal that endorsement is concentrated within a stigmatized subgroup. Progress signals can also be staged to engineer a cascade; releasing endorsements in batches can create multiple coordination points, but excessive control can reduce credibility. Thus, disclosure policy is a strategic choice that interacts with topology and organizer reputation.

From a network diffusion perspective, as explored in models like Watts (2002), a social assurance contract essentially seeks to trigger a cascade of endorsement once the critical mass T is achieved. One might consider the role of "k-cores" within the network: if a subset of at least T mutually connected or trusting individuals can be convinced to sign (perhaps because they have high e_i or low α_i), their collective private commitments can then persuade others on their periphery to join. However, such cascades might be halted at the boundaries of network modules if bridging ties are weak. Overall, a network exhibiting moderate connectivity—neither overly fragmented nor excessively centralized around a few vulnerable hubs—is likely most conducive to the success of a social assurance contract. Such a structure supports the formation of an initial critical mass and the subsequent widespread propagation of the revealed stance, potentially leading to a more significant and group-wide update of beliefs. The perceived $p_{succ|i}(q;T)$ for any individual would then be influenced by their network position and the local signals of support they observe.

H Outline of a Model Extension with Strategic Opposition

Section 5.2 incorporates a reduced-form durability-floor version of opposition into the main text. This appendix outlines a more comprehensive model of the interaction between the Overton window and a social assurance contract. It incorporates additional realistic features, primarily by endogenizing social costs through the strategic actions of individuals who may oppose the success of the contract, and by allowing for uncertainty regarding the prevalence of supportive types within the eligible population. Formal analysis of this richer extension is left for future research.

H.1 Model Components in the Extension Framework

1. Population Segmentation and Preferences— Let $\mathcal{P}$ be the total underlying population and let $\mathcal{E} \subseteq \mathcal{P}$ denote the eligible population for the contract. Each individual $j \in \mathcal{P}$ possesses an individual-specific valuation v_j

concerning the public revelation and potential normalization of the statement endorsed by the contract. This valuation v_j can be positive, zero, or negative, leading to a natural segmentation of $\mathcal{P}$:

- **Supporters ($\mathcal{S}$):** The set of eligible individuals for whom $v_j > 0$, so $\mathcal{S} \subseteq \mathcal{E}$. Let $K = |\mathcal{S}|$ be the true number of such supportive types in the eligible population. In this extension, unlike the main model, the prevalence of supportive types is not summarized only through a known common distribution; agents may instead form subjective beliefs about the realized support base. Concretely, each agent $k \in \mathcal{P}$ may form a subjective belief or estimate, $\hat{K}_k$ (which could be a point estimate or a probability distribution), about the true value of K. Each supporter $i \in \mathcal{S}$ also has heterogeneous private esteem e_i, cost c_i (which will now be endogenous), and warm-glow w_i (which I assume is heterogeneous here) associated with choosing $a_i = 1$ (Endorse).

- **Indifferents ($\mathcal{N}_0$):** The set of individuals for whom $v_j = 0$. The size of this set, $|\mathcal{N}_0|$, may also be uncertain.

- **Opposers ($\mathcal{O}$):** The set of individuals for whom $v_j < 0$. These individuals experience disutility if the endorsed statement is successfully publicized and enters the Overton window. The size of this set, $|\mathcal{O}|$, may also be uncertain.

2. Endogenous Social Costs (c_i) for Supporters— For a supporter $i \in \mathcal{S}$, the private cost c_i incurred if they endorse ($a_i = 1$) and the contract succeeds ($M \geq T$) is now explicitly a function of the aggregate punishment enacted by opposers. Let P_{total} be the total punishment enacted by the set of opposers $\mathcal{O}$. Then $c_i = C_i(P_{total}, M, \text{own characteristics}_i)$, where a higher P_{total} or a lower M (number of endorsers sharing the burden) would typically increase c_i.

3. Opposers' Characteristics and Strategic Behavior— Each opposer $j \in \mathcal{O}$ is characterized by their negative valuation $v_j < 0$ and an "influence" or "power" parameter $i_j \geq 0$. Their capacity to contribute to the punishment pool is $p_j(v_j, i_j)$, where higher $|v_j|$ (stronger opposition) and higher i_j lead to greater p_j. Let $\mathbf{p} = \{p_j\}_{j \in \mathcal{O}}$ be the vector of punishment potentials. If the contract succeeds ($M \geq T$) and the endorsed statement is published, opposers then decide whether to enact punishment. Crucially, I assume enacting punishment is *costless* to the opposers themselves ($k_p = 0$). However, they only choose to punish if they believe their collective punishment will be *effective*. Punishment is effective when it prevents the endorsed statement from durably entering or remaining in the Overton window. If the total punishment $\sum p_j$ is sufficiently high relative to the number of endorsers M (i.e., if T was set too low by the administrator, leading to a small M), the resulting c_k for endorsers might be so large that expressing the endorsed statement remains prohibitively costly overall, negating the contract's impact on the Overton window. In this case, ex-post utility for a significant number of revealed endorsers is negative: $v_k + e_k - c_k(M, \sum p_j) + w_k < 0$ for many $k \in \{i \mid a_i = 1\}$. Opposers would not punish if they estimate that M is so large (potentially influenced by their own estimate of K and of how supporters react to that) that their pooled punishment $\sum p_j$ would be too diluted to effectively increase c_k to a deterrent level, meaning the underlying belief $\mathbf{b}_s$ would enter the Overton window regardless of their efforts.

4. Supporters' Decision with Endogenous Costs— Supporters $i \in \mathcal{S}$ must now form expectations that are conditional on their individual estimate of the total number of supportive types, $\hat{K}_i$. These expectations

cover: (a) the likely number of actual endorsers M who will sign (which depends on $\hat{K}_i$, the threshold T, the distribution of private types (v_j, e_j, c_j, w_j) among the estimated supportive agents, and their strategic responses); (b) the potential collective punishment P_{total} from O if the contract succeeds with M endorsers; and (c) how this punishment translates into their individual $c_i(M, P_{total})$. Their marginal decision to sign depends on the incremental expressive payoff $e_i - c_i(M, P_{total}) + w_i$, together with any pivotal and stigma terms, rather than on the full total-success utility $v_i + e_i - c_i(M, P_{total}) + w_i$. The latter remains the relevant ex-post durability object once success has occurred. The central strategic problem is therefore to forecast whether the punishment subgame will leave the marginal signing gain positive in expectation.

5. Architect's Objective— The contract architect sets T. Their objective might be to maximize the probability that $\mathbf{b}_s$ *durably* enters the Overton window. This now explicitly involves choosing a T considering not only the likely distribution of types but also the fact that potential supporters (and opposers) operate with potentially heterogeneous estimates $\hat{K}_k$ of the true number of supportive types K. An optimal T would need to be robust enough or be perceived as achievable given these beliefs, such that if $M = T$ endorsers are revealed, this number is sufficient to signal to opposers that punishment would be futile, thus ensuring c_i remains manageable for supporters.

H.2 Straightforward Implications and Complexities of the Model with Strategic Opposition

Introducing strategic, (personally) costless but efficacy-dependent punishment by opposers, coupled with uncertainty about the total number of supportive types K, has several key implications for the social assurance contract model. Firstly, the concept of "safety" for endorsers becomes endogenously determined and subject to more layers of uncertainty. It is no longer solely about being one of $M \geq T$ individuals, but more critically about M being sufficiently large (relative to individual beliefs $\hat{K}_i$ and the true support base K) to deter or significantly dilute punishment from the opposing group O. If the administrator sets the mechanism's threshold T too low, a social assurance contract might formally "succeed" (i.e., achieve $M \geq T$ endorsements), yet this M could be insufficient to deter opposers. In such a scenario, the resulting costs c_i imposed on endorsers might become so high as to render their ex-post utility negative. Consequently, the endorsed belief $\mathbf{b}_s$ might fail to durably enter the Overton window because association with the endorsed statement remains prohibitively costly. This implies the existence of an effective impact threshold, $M_{Overton}$, representing the level of observed endorsement M at which opposers would likely deem punishment ineffective and stand down. The administrator's optimal choice for T would then ideally be at or above this $M_{Overton}$, or at least set to maximize the probability $P(M \geq M_{Overton})$, considering the distribution of beliefs about K. Second, this framework significantly increases the complexity of strategic calculations for all involved. Supporters must now forecast not only the participation decisions of other potential supporters (from a pool whose size $\hat{K}_i$ they estimate) but also the strategic response of the entire group of opposers, which influences their expected c_i. Opposers, in turn, must forecast the likely M resulting from the social assurance contract to determine if their potential punishment would be worthwhile. Third, this introduces new potential failure modes for the contract. Beyond the simple failure of $M < T$ (resulting in no publication), a social assurance contract could achieve $M \geq T$ only to be met with effective punishment. Such an outcome would make endorsers regret their decision and could prevent the intended shift in the Overton window, representing a pyrrhic victory for the

initial publication. Finally, the success of a social assurance contract in durably shifting norms and achieving its objectives will depend on the characteristics and coordination of opposers, and now also on the collective beliefs and estimation accuracy regarding the latent support base. The aggregate punishment potential, $\sum p_j$, and the ability of opposers to coordinate their punishment strategy become paramount factors. Analyzing these rich interactions would indeed require advanced game-theoretic tools.

A full analysis of this extension is a substantial undertaking well beyond the scope of the present paper. However, outlining it serves to highlight the broader context in which social assurance contracts operate. It emphasizes that social costs (c_i) faced by supporters are not merely exogenous parameters but can be the result of active, strategic opposition, and that the realized size of the supportive faction within a known eligible population may itself be subject to uncertainty. The simpler model analyzed in the main body of this paper provides a foundation by elucidating the core coordination mechanism among agents in a known population with uncertain private types. On this foundation, future research can model these more complex inter-group dynamics and informational and strategic uncertainties.