Verifying Away Collective Knowledge

Shusuke Ioku[†]

July 23, 2026

Abstract. Generative AI lets influence operations fabricate apparent widespread support cheaply. I develop a theoretical framework for online forums involving politically restricted information to isolate a tradeoff between authenticity and accuracy that generative AI exacerbates. Citizens collectively choose verification, then decide among anonymous posting, verified posting, and silence. When attackers can flood anonymous forums, citizens adopt the least verification that deters them. Because greater access raises knowledge and exposure, verification exposes better-informed sources to greater punishment risk and, unlike in canonical screening models, excludes the best-informed, hence most strongly motivated to contribute, first. Citizens value collective accuracy, but individual contributions remain externalities; private exposure risk therefore makes the forum more authentic but less informative, allowing the influence operation to deplete collective knowledge without deployment or deception. Detection and per-post costs reduce reliance on identity verification. Evaluations of epistemic security should track both the fakes excluded and the composition of who remains.

Word Count: 9,887

[†]Ph.D. Candidate, Department of Political Science, University of Rochester.
Email: iokushusuke@gmail.com.

Introduction

Online disinformation operations have become a consequential instrument of political competition. Russia's Internet Research Agency deployed coordinated personas across major platforms around the 2016 United States election, and social bots accelerated the spread of low-credibility content (Bail et al. 2020; Shao et al. 2018). Generative AI creates a new threat by automating convincing influence content and scaling both political persuasion and personalized appeals (Bai et al. 2025; Goldstein et al. 2023; Matz et al. 2024). An AI swarm can turn that capacity into a large population of coordinated synthetic personas.

Identity and personhood verification are leading defenses against AI swarms, alongside behavioral and network detection (Goldstein et al. 2023; Schroeder et al. 2026). Their appeal grows as AI-generated personas evade content and coordination checks. Yet defensive measures can also backfire politically: state-sanctioned tools may suppress dissent or amplify incumbents (Schroeder et al. 2026). Identity verification poses a specific danger because it can expose dissidents, activists, and whistleblowers who rely on anonymity to speak safely. Such cases motivate the key tradeoff studied here. When claims turn on *politically restricted information*, including nonpublic evidence of election manipulation, concealed official misconduct, or classified state activity, greater knowledge can increase identification and punishment risk: verification makes fabrication costlier but can silence better-informed sources.

I develop a theoretical framework for online forums involving politically restricted information to isolate a tradeoff between authenticity and accuracy that generative AI exacerbates. Citizens receive private signals, audiences treat support volume as evidence, and an operator can buy synthetic voices for a false claim. Verification raises fabrication costs but can drive out better-informed genuine sources because greater access both sharpens signals and increases exposure. The model asks which verification levels deter a swarm and which genuine citizens keep posting.

At a fixed policy, costly fabrication makes the anonymous forum too expensive to swamp, so it remains informative without exposing anyone. As fabrication cheapens, deterrence instead requires verification to supply an offsetting forgery premium. Citizen utility includes both private posting payoffs and a common payoff from the accuracy of the public belief. Verification is chosen before citizens learn their information and exposure types, so they share the same ex-ante ranking: the welfare optimum is their unanimous democratic choice. In the pure-strategy benchmark, society chooses no verification while anonymity is self-defending and the shallowest deterring credential once it is not. The chosen stringency therefore rises as fabrication cheapens. No synthetic voice appears along this selected path; the threat operates through the defense it induces.

The central result arises once the selected credential passes the silencing threshold. For claims involving politically restricted information, greater access raises both knowledge and punishment risk, so the best-informed sources face the highest authentication cost. They exit first as verification deepens, despite having the strongest private motive to post. Citizens also value public accuracy, but an atomless poster cannot change it alone: she bears attribution risk privately, while a positive mass of exits reduces the common informational benefit. Along the selected path, authenticity remains perfect, yet the surviving voices become less precise and the public signal less accurate. The operator benefits from a less informative forum without spending or deceiving anyone. This backfire does not arise if moderate verification already deters, retaliation stakes are too low to induce exit, or information and exposure are not positively aligned.

The model also accommodates other defenses against fabricated participation. Ideal detection raises the attacker's effective cost and weakly lowers the identity stringency citizens choose. I then ask whether a per-post cost can ease the tradeoff when combined with verification. By making synthetic volume more expensive without increasing identity exposure, the cost lowers the verification needed for deterrence and can preserve informed participation. Genuine posters also pay it, however, so welfare improves only when the resulting exposure and accuracy gains outweigh that added burden.

The paper extends theories of political information manipulation by endogenizing the defense triggered by cheaper fabrication. Empirical studies document state-directed posting and astroturfing that make coordinated speech appear to come from many independent citizens (Keller et al. 2020; King, Pan, and Roberts 2017). Formal political models endogenize fraud or false reporting and rational audience responses (Kang and Sheen 2025; Little 2012; 2014; Zakharov 2026). Economic models likewise endogenize signal distortion or anonymous promotional volume (Dellarocas 2006; Edmond 2013; Edmond and Lu 2021; Mayzlin 2006). Across these models, the terms on which genuine and fabricated voices enter the public signal remain fixed. This paper makes authentication responsive to fabrication cost. As synthetic speakers become cheaper, citizens tighten authentication until the operator is deterred; no synthetic voice appears along the selected path. Manipulation technology can therefore reduce public information entirely through the defense it induces, while manipulation itself remains off path.

The paper also identifies a new source-selection mechanism among genuine citizens. Models of protest, petitions, and voting endogenize participation, pivotality, or receiver response while holding fixed the identity and eligibility rules governing access to the channel (Battaglini 2017; Feddersen and Pesendorfer 1997; Lohmann 1993). With heterogeneous expertise, strategic abstention can improve selection because less-informed citizens defer to better-informed peers (McMurray 2013). Cascades and herding lose information when earlier actions cause later actors to disregard private signals (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1992). Preference falsification does so when social penalties induce citizens to conceal genuine views (Kuran 1991). A nonpolitical benchmark shows the conventional disciplining pattern: real-name verification reduced participation in Chinese investment forums but made aggregate sentiment more informative about firms' earnings and returns (Huang et al. 2026). For claims involving politically restricted information, authentication can reverse that selection. By making entry into the public signal costly in proportion to exposure, it removes the best-informed sources first. Every surviving source is authentic, yet survivor precision and public accuracy fall. The contribution is

authentication-induced adverse selection, not the familiar result that costly participation reduces turnout.

Institutionally, citizens themselves choose anti-manipulation policy. Existing formal models locate policy in platform, regulator, or planner problems: an engagement-maximizing platform chooses a sharing network and a learning-oriented regulator evaluates countermeasures (Acemoglu, Ozdaglar, and Siderius 2023), while users decide whether to verify content and a planner evaluates the efficient verification allocation (Cisternas and Vasquez 2022). These models endogenize platform or user responses, but assess policy under a specified institutional or planner objective rather than derive citizens' collective choice. Work on proof of personhood and AI swarms recognizes privacy, inclusion, dissident safety, and democratic accountability, but maps design principles and governance tradeoffs rather than deriving an adoption equilibrium (Ford 2020; Schroeder et al. 2026). Here citizens choose linkable verification before learning their information and exposure types. Because they face the same lottery, the welfare optimum is the verification level they would unanimously choose ex ante. The informational loss can therefore arise even when society itself selects the optimal feasible defense, rather than from platform misalignment, censorship, or policy error.

The Model

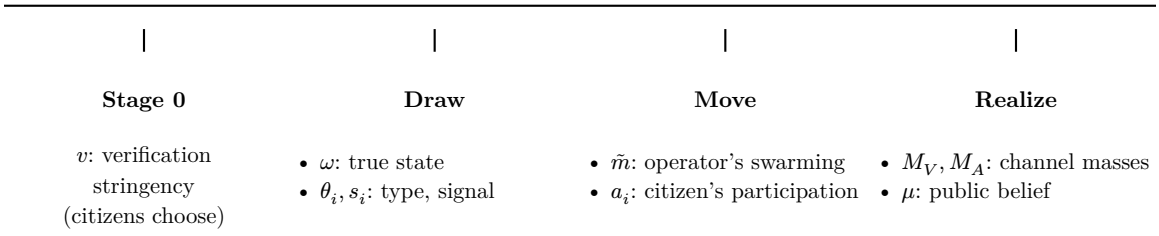

Figure 1: Sequence of the Game

Note: The verification stringency v is fixed at stage 0; within the posting game everyone moves at once. $a_i \in \{V, A, \varnothing\}$ is citizen i's choice of the verified channel, the anonymous channel, or silence.

The model starts from one contested claim whose support is summarized by a count of favorable contributions. This observation structure describes social-media threads and hashtag campaigns, public-comment dockets, petitions, endorsement lists, review systems, and public allegation forums in which the volume of repeated support is evidence. The claim can concern official wrongdoing, electoral administration, or a policy proposal. Each contribution supplies one unit of support rather than message-specific evidence, so the model isolates the volume of participation from the content of individual messages. Figure 1 provides a preview of the sequence of the model.

The model has two moving parts. Verification stringency sets a forgery premium on every synthetic voice and an exposure cost on every genuine poster. Politically restricted information motivates the leading case: the same role-based access that improves a citizen's signal also narrows the possible source set if verification links her post to her identity. The result is a composition problem as well as an authenticity problem. A continuum of citizens and a resourced operator choose simultaneously whether to add voices to the forum after the polity fixes verification stringency.

Citizens and Signals

The environment has a binary state $\omega \in \{0, 1\}$ with prior $\Pr(\omega = 1) = \frac{1}{2}$: the claim is true when $\omega = 1$ and false when $\omega = 0$. A continuum of genuine citizens of measure one is indexed by exposure type $\theta \sim U[0, 1]$, which measures what a citizen risks by coming forward. Low types are bystanders with little to lose; high types are insiders whose position or access makes them identifiable. Each citizen receives a private signal $s_i \in \{0, 1\}$ with precision

$$q(\theta) = q_0 + \delta\theta, \quad q_0 \in \left(\frac{1}{2}, 1\right), \quad \delta \in \left(\frac{1}{2} - q_0, 1 - q_0\right) \tag{1}$$

independently across citizens conditional on ω. Motivated by politically restricted information, the leading case, $\delta > 0$, is *positive exposure–precision dependence*: citizens with greater costs of exposure also receive more precise signals. The cases $\delta = 0$ and $\delta < 0$ identify composition-neutral and reversed-selection benchmarks.

Posting, Verification, and Payoffs

A citizen chooses the verified channel, the anonymous channel, or silence: $a_i \in \{V, A, \varnothing\}$. Posting adds one comment, reply, or signed submission to the observable mass supporting position 1 in the chosen channel. For simplicity, no action records opposition to the claim; allowing symmetric opposition leaves the underlying selection logic unchanged, as §Comments on the Model explains. Silence's action-dependent payoff is zero; regardless of her action, she also receives the common public-accuracy payoff defined below.

Citizens are truth-seeking rather than directional: they do not prefer the claim to be true or false, and they value posting only when it matches the state. Backing position 1 in a heeded channel yields b_C when the claim is true and $-b_C$ when it is false; an unheeded channel yields no contribution benefit. This reduced form captures professional, civic, reputational, or moral value from publicly backing the correct claim in a venue that matters. A verified poster also pays exposure $\theta_i v$. The action-dependent part of her realized payoff from channel $\ell \in \{V, A\}$ is

$$u_i^R(\ell, \omega) = b_C\,(2\omega - 1)\,\rho_\ell - \theta_i\, v\, \mathbb{1}\{\ell = V\}$$

Set $u_i^R(\varnothing, \omega) = 0$.

Citizens also value more accurate collective belief directly. Every citizen receives the common public-accuracy payoff $-\gamma(\omega - \mu)^2$, so her full realized utility from action $a_i \in \{V, A, \varnothing\}$ is

$$U_i^R(a_i, \omega, \mu) = u_i^R(a_i, \omega) - \gamma(\omega - \mu)^2, \quad \gamma > 0.$$

Because $\mathbb{E}[2\omega - 1 \mid s_i, \theta_i] = (2s_i - 1)(2q(\theta_i) - 1)$, the action-dependent part of her interim expected utility is

$$u_i(\ell) = b_C\,(2s_i - 1)\,(2q(\theta_i) - 1)\,\rho_\ell - \theta_i\, v\, \mathbb{1}\{\ell = V\}, \quad b_C > 0 \tag{2}$$

A citizen's interim expected utility consists of (2) plus her expected common accuracy payoff. Because one atomless citizen cannot change μ, the latter is the same whether she

posts or remains silent and therefore cancels from that choice. It still enters stage-0 policy choice because verification changes aggregate participation and public accuracy.

The factor $(2s_i - 1)$ changes the sign of the expected contribution payoff. Because every voice supports position 1, a citizen whose signal favors $\omega = 0$ expects posting to support the wrong position. Only citizens with $s_i = 1$ post in a heeded channel. The coefficient b_C is their private correctness payoff, not the marginal effect of one atomless voice on the posterior and not a share of the common public-accuracy payoff.

Verification stringency $v \in [0, 1)$ measures the identity link surrendered by a poster, from no link to a targetable biometric or legal record. A genuine citizen of type θ_i bears expected exposure $\theta_i v$. A synthetic persona costs $c > 0$ to create or purchase in either channel and pays an additional forgery premium $k(v)$ to acquire an account that passes verified enrollment.[1] The premium is continuous and strictly increasing, with $k(0) = 0$ and $k(v) \to \infty$ as $v \to 1$ (leading form $k(v) = \frac{\bar{k}v}{1-v}$):

$$\kappa_A = c, \qquad \kappa_V(v) = c + k(v) \tag{3}$$

The unbounded forgery premium represents linked systems in which deeper checks make credential lending, reuse, and ban evasion increasingly attributable and sanctionable, contracting the supply of safely rentable credentials. It is a technology assumption, not a consequence of linkability itself. If coercion, rental, or spoofing caps the premium, a sufficiently high-stakes operator cannot be deterred by identity verification alone.

The Operator and the Public Belief

A single *operator* originates the claim and knows the state. Unlike citizens, the operator is directional: it benefits when the public believes the claim in the false state. When $\omega = 0$, it can buy synthetic masses $\tilde{m}_A, \tilde{m}_V \geq 0$ for the anonymous and verified channels at the

[1] Markets for established personal accounts are documented in the Philippines, where dormant Facebook accounts are rented and pages with accumulated audiences are sold for political use (Tandoc 2022). DRAGONBRIDGE, a China-linked influence network, likewise obtained Google Accounts from bulk sellers and reused dormant accounts that appeared to change hands (Butler and Taege 2023).

per-persona costs in (3). Its payoff rises with the public belief that $\omega = 1$ and falls with persona spending:

$$u_O = b_O \, \mathbb{E}[\mu \mid \omega = 0] - c \, \tilde{m}_A - (c + k(v)) \, \tilde{m}_V, \quad b_O > 0 \tag{4}$$

In state $\omega = 1$ the operator is idle. The public observes the pair of channel masses. Each equals the genuine posting mass $m_\ell(\omega)$ plus synthetic mass $\tilde{m}_\ell(\omega)$, which is zero in the true state, and channel-specific organic noise:

$$M_\ell = m_\ell(\omega) + \tilde{m}_\ell(\omega) + \sigma \, \varepsilon_\ell, \quad \ell \in \{V, A\}, \quad (\varepsilon_V, \varepsilon_A) \sim N(0, I_2), \quad \sigma > 0 \tag{5}$$

For each channel ℓ, define net state separation and standing by

$$\Delta_\ell := \mathbb{E}[M_\ell \mid \omega = 1] - \mathbb{E}[M_\ell \mid \omega = 0], \quad \rho_\ell := \mathbb{1}\{\Delta_\ell > 0\}.$$

Thus a channel is *heeded* exactly when $\rho_\ell = 1$.

The two shocks are independent of the state, citizen types and signals, and all behavioral randomization. Within a channel, the public cannot separate genuine from synthetic voices, so an unusual burst of genuine participation is statistically confusable with a swarm. The shared posterior $\mu(M_V, M_A) = \Pr(\omega = 1 \mid M_V, M_A)$ forms by Bayes' rule under equilibrium-consistent conjectures. Independence makes a channel with zero state separation irrelevant to the likelihood ratio: its density is identical in both states and cancels. Flooding such a channel therefore changes no equilibrium belief and only adds cost.

Timing

Figure 1 summarizes the timing. At stage 0, the citizenry chooses verification stringency v through a democratic process, anticipating the participation and flooding behavior it induces. Because types are drawn only afterward, citizen welfare is each citizen's own ex-ante criterion, so its maximizer is the level society would implement rather than merely a planner's recommendation. Participation cannot be mandated, so the chosen policy must be self-enforcing for genuine posters.

Nature then draws ω and every citizen's type–signal pair. The operator observes ω; each citizen observes only her own (θ_i, s_i). They move simultaneously, neither observing the other's action. The aggregate masses then realize and determine the posterior μ. At stage 0, the citizenry selects the smallest welfare-maximizing v. Indifferent citizens choose anonymity over verification and silence over posting in an unheeded channel. At $v = 0$, anonymity labels otherwise identical channel outcomes. These conventions resolve payoff ties without changing any strict comparison.

Comments on the Model

Posting is an aggregate count rather than message content. A voice adds one unit of mass for a position; the public observes channel totals and never individual text. Synthetic text and faces can be difficult to distinguish from genuine ones, which makes volume a plausible target for fabrication (Nightingale and Farid 2022; Spitale, Biller-Andorno, and Germani 2023). The 2017 net-neutrality docket supplies the clearest institutional example: fabricated comments changed the apparent scale of participation in a formal public process (Office of the New York State Attorney General 2021). The benchmark isolates inference from voice counts by assuming that the public observes no other evidence. Fact-checks, provenance records, or independent estimates of synthetic participation would provide additional signals and change the posterior.

The continuum of citizens represents claims that can be corroborated by dispersed observers. Examples include alleged election irregularities across precincts, systemic misconduct across agencies, or implementation of a broadly applied policy. Each observer's access may be locally narrow, but agreement across many independent observations is informative, making synthetic volume a threat. Claims known only to one identifiable witness fall outside this scope, because inference then turns on that witness and her evidence rather than on a support count. The Supporting Information (§E.4.1. Finite Source Pools and Pivotality) shows how the analysis changes with a finite source pool.

Other influence operations work through different registers. Coordinated posting can crowd out unwanted topics through distraction and flooding, or change citizens' beliefs about what others believe and thereby induce self-censorship (King, Pan, and Roberts 2017; Roberts 2018; Ross et al. 2019). Those operations need not make the count for one focal claim informative. The present accuracy results apply to institutions in which the volume of favorable contributions is itself evidence. Models of agenda displacement or perceived-majority effects instead require audiences to respond to topic salience or perceived social norms, not to infer truth from a focal claim's support count.

Posting is unidirectional. Citizens can support the focal claim or remain silent. This restriction isolates how authentication changes the composition of genuine posting while keeping the model tractable. A competing claim would expand each citizen's choice from three actions to five: support or oppose through either channel, or remain silent. Public inference would then depend on four endogenous participation counts, and the operator could divide manipulation between promoting its claim and attacking the counterclaim across both channels. Solving the joint participation cutoffs and this multidimensional flooding problem would remove the closed-form deterrence and policy analysis used here.

The restriction does not drive the adverse selection. On either side, the private value of correct posting rises with precision, while the cost of verification rises with exposure. Once exposure rises faster, the best-informed posters on that side exit first. A two-sided model would change the deterrence thresholds and accuracy formulas, but not this participation ordering. The one-sided version also maps directly to comment dockets, petitions, endorsement lists, reviews, and public allegation forums where repeated support is itself evidence.

Why better-informed posters face higher exposure costs. Politically restricted information often consists of different pieces of evidence held by officials or employees about classified or otherwise nonpublic conduct, making corroboration informative. Role-based access sharpens information and, after attribution, narrows the source set. This punishment margin also exists in democracies, which commonly criminalize unauthorized disclosure of

classified information through secrecy, espionage, or defense laws (Government of Japan 2013; Supreme Court of the Australian Capital Territory 2024; U.S. Department of Justice 2018). Once a source is identified, exposure can therefore carry legal sanctions even for truthful disclosure.

Politically restricted information motivates this relationship, but I do not claim it is universal. It is the scope condition for the leading result: expected identification costs rise with signal precision ($\delta > 0$). The one-dimensional type makes this rank alignment exact, so the result applies most directly where the alignment is strong. When the two are unrelated ($\delta = 0$), verification changes volume but not composition; when better-informed posters face lower identification costs ($\delta < 0$), verification improves survivor precision. For politically restricted information, positive dependence is plausible whenever the circumstances that improve knowledge also make attribution more costly. Classified programs, official misconduct, internal fraud, and compliance failures are learned through records, meetings, and decisions tied to employment, legal, or professional stakes. Retaliation is most certain and severe for disclosures of systematic and significant misconduct, consistent with this common source of knowledge and identification costs (Rothschild and Miethe 1999).

The credential is linkable. A pseudonym may conceal a poster from the audience yet leave a link recoverable by a state, court, or employer. The model concerns claims involving politically restricted information for which the holder of that link can punish disclosure even when the post is truthful. Audience anonymity does not remove risk when systems retain names or enrollment records. A design that keeps issuance and later use unlinkable escapes the retaliation channel while still limiting duplicate participation. Pseudonym parties illustrate this de-linked class and face different constraints involving physical attendance, scaling, exclusion, multiple issuance, transfer, and coercion (Ford 2020). Those constraints are not relabeled as identity exposure. The unbounded forgery premium is likewise an assumption about increasingly traceable linked credentials, not a general property of proof-of-personhood systems.

Channel standing is binary. A compromised channel gives posting no private value, while a heeded channel gives each voice the same value regardless of how many citizens remain. This assumption removes a benefit-side reward for keeping the verified forum informative and isolates the exposure channel. It also creates a sharp difference between the live and compromised regimes: once a channel carries no state separation, a contribution yields no private return.

The binary restriction is innocuous for the adverse-selection result. The Supporting Information (§E.3. Continuous Benefit-Side Standing for Proposition 2) instead lets the value of posting rise continuously with the channel's informativeness. Whenever the forum remains in a stable informative equilibrium, deeper verification still reduces participation and accuracy; the feedback makes both declines steeper. Continuous standing can change the global equilibrium structure by creating multiple live cutoffs or eliminating a participation branch. The binary specification avoids those coordination complications without generating the composition effect.

Functional-form scope. Uniform exposure, affine precision, an even prior, and independent equal-variance Gaussian noise deliver the displayed cutoff, constant precision gap, scalar deterrence bound, and closed-form policy thresholds. They also make the public experiment depend on a single discriminability index, which permits the global secant characterization of flooding. These choices are substantive primitives rather than innocuous normalizations.

The adverse-selection sign needs less structure. With any continuous type distribution and increasing signal precision, a low-type single crossing between the private posting benefit and exposure makes greater stringency lower the participation cutoff. Mean survivor precision and genuine state separation then fall. Any observation family ordered by state separation carries that decline into posterior accuracy. The exact cutoff, the constant gap between silenced and heard precision, the per-post-cost threshold, and the numerical location of the deterrence peak belong to the uniform-affine-Gaussian specification. Later

results that require a slope of the deterrence bound impose that sign on the relevant range rather than relying on global numerical uniqueness.

The Equilibrium at Fixed Verification

The solution concept is pure-strategy Bayesian Nash equilibrium, in which no player randomizes, in the simultaneous posting game, completed by Bayes-consistent public updating. Beliefs, citizen behavior, and manipulation are jointly consistent (Edmond 2013). The posterior $\mu^*(\cdot)$ is a contingent rule in the equilibrium profile. Gaussian noise has full support, so Bayes' rule pins that rule down at every observed mass. When checking an operator deviation, citizen strategies and the posterior rule remain fixed; a different flood changes the masses at which the rule is evaluated. This is ordinary unilateral optimality in a simultaneous game.

An equilibrium profile contains a citizen strategy, a flooding rule, a posterior function, and the two channel standings.

> Definition 1 (Equilibrium at fixed v): A profile $(\alpha^*, \tilde{m}^*, \mu^*, \rho^*)$ such that: (i) *citizen optimality:* for every (θ_i, s_i), the action $\alpha^*(\theta_i, s_i)$ maximizes full citizen utility, equivalently the action-dependent payoff (2) because a unilateral action leaves the common accuracy term unchanged, given the standings ρ^*; (ii) *belief consistency:* $\mu^*(M_V, M_A) = \Pr(\omega = 1 \mid M_V, M_A)$ by Bayes' rule under the observation structure (5) evaluated at $(\alpha^*, \tilde{m}^*)$, and $\rho^*_\ell = 1$ if and only if $\Delta^*_\ell > 0$; (iii) *operator optimality:* taking α^* and the contingent posterior rule $\mu^*(\cdot)$ as given, $\tilde{m}^*$ maximizes (4) over $\mathbb{R}^2_+$; a unilateral deviation changes (M_V, M_A), and hence $\mu^*(M_V, M_A)$, but not the other equilibrium strategy functions.

The Supporting Information (§A. The Equilibrium Concept) records how these objects fit together and explains why deviations reduce to the one informative channel.

Two equilibrium issues require separate treatment. First, sufficiently cheap fabrication eliminates every pure-strategy profile in which the channel remains heeded: trust invites a flood, and a pooling-scale flood destroys the state separation that sustained trust. The resulting pure-strategy outcome is the *compromised forum*, with silence and prior-level accuracy. Allowing the operator to randomize restores heeded equilibria. The Supporting Information constructs these equilibria (§C.1. Existence) and bounds their accuracy (§C.2. The Accuracy of the Compromised Channel); the bound converges to the uninformative floor as the persona price vanishes, but the accuracy bound alone does not determine whether citizens prefer such an equilibrium to a verified forum.

Second, a dead forum can be self-confirming. If citizens expect no one to contribute, the channel carries no state separation and receives no standing. Without standing, posting yields no private correctness benefit, so citizens remain silent and confirm the initial expectation. The *Pareto Selection* chooses the equilibrium that every citizen weakly prefers. It removes forums that are dead solely from expectations while retaining deaths caused by the economics of flooding, as shown in the Supporting Information (§D. The Pareto Selection).

Preliminaries

Five lemmas characterize citizen participation, public accuracy, operator flooding, deterrence, and equilibrium at fixed v.

Let $t \in [0,1]$ index a participation cutoff, the set of posting $s_i = 1$ types being $[0,t]$. Since $q(\theta)$ is the probability a type-θ signal matches the true state, the genuine mass $m(\omega)$ at cutoff t (channel subscript ℓ dropped, holding for any fixed channel) and its state-separation are

$$m(1) = \int_0^t q(\theta)\ \mathrm{d}\theta, \quad m(0) = \int_0^t (1 - q(\theta))\ \mathrm{d}\theta,$$

$$\Delta(t) = m(1) - m(0) = (2q_0 - 1)t + \delta t^2$$

with discriminability $d = \frac{\Delta(t)}{\sigma}$. Two verification thresholds organize participation. The *incentive-balance threshold*, $\hat{v} = 2b_C\delta$, is where exposure cost and the private value of correct posting rise equally fast with type. The *silencing threshold*, $\bar{v} = b_C(2q(1) - 1) > \hat{v}$, is where the most exposed and best-informed citizen first becomes unwilling to post.

Consider first the citizen side, at conjectured standings (ρ_V, ρ_A). The decision is a single comparison. A heeded anonymous channel pays the full private correctness benefit at zero exposure, so it absorbs every potential poster. When only the verified channel is heeded, a citizen with $s_i = 1$ posts iff $b_C(2q(\theta_i) - 1)$ covers the exposure cost $\theta_i v$. The private benefit rises in θ at slope $2b_C\delta = \hat{v}$, while the cost rises at slope v, so once v exceeds $\hat{v}$ the comparison resolves into a cutoff in θ. Proofs of all formal results appear in the Supporting Information.

> LEMMA 1 (Citizen Best Responses Given Channel Standings): A citizen whose signal opposes the position never posts in a heeded channel. If the anonymous channel is heeded, every citizen whose signal favors it posts there. If only the verified channel is heeded, she posts exactly when her exposure type is below the cutoff $t(v)$: everyone posts up to the silencing threshold ($v \leq \bar{v}$); beyond it ($v > \bar{v}$), $t(v) = \frac{b_C(2q_0-1)}{v-2b_C\delta}$, strictly falling. If neither is heeded, all are silent.

Verification is a technology citizens use only after the anonymous channel loses standing. Whenever the anonymous channel is heeded, every potential poster chooses it and pays no exposure. In the verified regime, participation remains complete until v passes $\bar{v}$. Beyond that threshold, exposure rises faster with type than the private benefit of correct posting, and the marginal exit at every further increase in stringency is the best-informed voice still posting.

Consider a source deciding whether to support a claim involving politically restricted information in the public forum. At low stringency, her strong signal gives her the greatest private motive to post and verification costs too little to stop her. Once stringency passes

the silencing threshold, the exposure attached to her access outweighs that motive. She becomes silent while a less-informed bystander with less to lose continues posting. The forum therefore empties from the top of the knowledge distribution down.

The public side requires a measure of what the forum contributes to collective belief. Posterior accuracy depends on discriminability d, the difference between a channel's mean volume in the two states measured in units of organic noise. Discriminability is high when genuine participation makes a true claim generate systematically more support than a false one.

> LEMMA 2 (The Accuracy Yardstick): Posterior accuracy $Q(d) = -\mathbb{E}[\mu(1-\mu)]$ is strictly increasing in the discriminability d ($Q'(d) > 0$), from $-\frac{1}{4}$ at zero, where the public keeps its prior, toward 0, full revelation, as d grows without bound.

The accuracy yardstick has a direct interpretation. A forum informs the public only to the extent that truth and falsehood generate different observed volumes. Fabrication closes that gap. When state separation is small relative to ordinary noise, the posterior stays near the prior; when the gap is large, the observed mass nearly reveals the state. The Gaussian experiment makes this ordering exact rather than assuming it.

The operator's problem is how much fabrication a forum can withstand. Write $K(d)$ for the highest per-persona price at which a profitable flood exists against a channel of discriminability d. The bound begins at zero, rises at a positive limiting slope, remains positive at every finite positive d, and returns to zero as discriminability diverges. It therefore reaches at least one interior maximum. A numerical grid for the maintained Gaussian specification finds a single peak near $d = 2.54$; the proof requires only existence of a peak, not numerical uniqueness. Fix a channel with participation cutoff $t > 0$ and persona price $\kappa > 0$, while the other channel carries no state separation.

LEMMA 3 (Jamming and Deterrence): A channel is safe from flooding exactly when the persona price is at least a deterrence bound ($\kappa \geq K(d)$). The bound is continuously differentiable on $(0, \infty)$ and strictly positive for $d > 0$, never exceeds $\frac{b_O d}{4\sigma}$, satisfies $\frac{K(d)}{d} \to \frac{b_O}{4\sigma}$ as $d \to 0$, and converges to zero as $d \to \infty$; hence it attains at least one interior maximum. A deterred channel is heeded, with the accuracy its discriminability delivers; an undeterred one is compromised, and worth only the prior. A channel believed to carry no information is trivially deterred.

Deterrence is global rather than marginal. When flooding is profitable, the operator jumps to a pooling-scale mass that makes the false state resemble the true one. A smaller swarm shifts the posterior too little to cover its persona cost. This discrete incentive explains why the pure-strategy forum collapses below the deterrence threshold instead of sustaining a steady trickle of fakes. The analytical result establishes an interior maximum of K.

The rising side of the deterrence bound has a close precedent: strategic promotion is most attractive when genuine favorable posting is scarce (Mayzlin 2006; Mayzlin, Dover, and Chevalier 2014). The distinct force here is the falling right side. Pooling a highly discriminating channel requires a synthetic mass comparable to its genuine state separation. The required spending then grows faster than the bounded shift in belief the operator can purchase, making a sufficiently informative forum cheaper to defend.

Two derived objects recur. The *anonymous self-defense cost* $\bar{c} = K(d_0)$, with $d_0 = \frac{\Delta(1)}{\sigma}$, is the persona price at which the fully-attended anonymous channel is exactly self-defending. The *deterrence set* $\mathcal{D}(c) = \left\{v : c + k(v) \geq K\left(\frac{\Delta(t(v))}{\sigma}\right)\right\}$ collects the verification stringencies at which the verified channel, populated at its own cutoff $t(v)$, is uncompromised. When $\bar{v} < 1$ and $k(\bar{v}) < \bar{c}$, define the positive crossing cost $\underline{c} := \bar{c} - k(\bar{v})$. Only under those conditions can the minimum deterring depth cross the silencing threshold at a positive fabrication cost.

LEMMA 4 (The Deterrence Set): The deterrence set is closed, grows with the fabrication cost, and contains all sufficiently deep verification stringencies. Its minimum ($\underline{v}(c) = \min \mathcal{D}(c)$) is zero exactly when the anonymous forum defends itself ($c \geq \bar{c}$) and rises as fabrication cheapens. Within the verification region, while participation is full, it is continuous with closed form $\underline{v}(c) = k^{-1}(\bar{c} - c)$. It crosses $\bar{v}$ at the positive cost $\underline{c}$ exactly when $\bar{v} < 1$ and $k(\bar{v}) < \bar{c}$; otherwise it remains below $\bar{v}$ for every $c > 0$.

The anonymous self-defense threshold $\bar{c}$ measures how expensive the fully attended forum is to attack before verification. The minimum deterring stringency $\underline{v}(c)$ measures how much identity exposure is required after fabrication cheapens. While participation remains full, every decline in production cost must be replaced by an equal increase in the forgery premium. Under the two crossing conditions, the minimum deterring policy eventually passes $\bar{v}$ and begins selective exit. If either condition fails, verification can be adopted without reaching that margin.

The fixed-v equilibrium requires citizen behavior, the operator's flooding decision, and the channel standings to be mutually consistent. The pure-strategy restriction resolves the compromised region, and the Pareto Selection resolves self-confirming expectations. Write $d(v) := \frac{\Delta(t(v))}{\sigma}$ for the discriminability produced by the verified participation cutoff.

LEMMA 5 (Regime Structure): In pure strategies, under the Pareto Selection, the equilibrium is one of three regimes. If fabrication is costly enough that the anonymous forum defends itself, everyone with a favoring signal posts anonymously at any stringency, paying no exposure: the *anonymous regime* ($c \geq \bar{c}$). If it is not, but the stringency deters, citizens post verified up to the participation cutoff: the *verified regime* ($c < \bar{c}, v \in \mathcal{D}(c)$). Otherwise all are silent: the *dead forum* ($c < \bar{c}, v \notin \mathcal{D}(c)$).

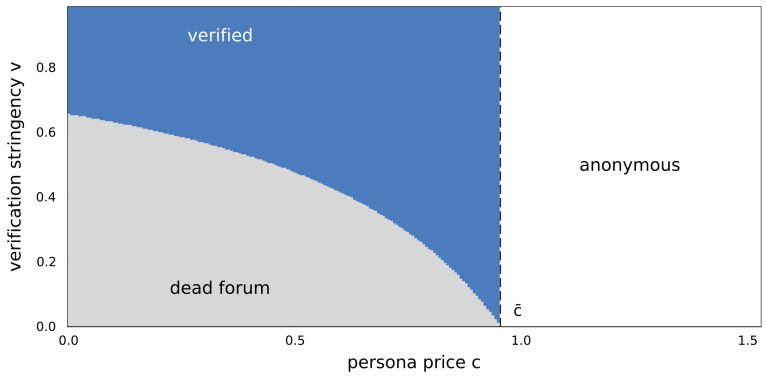

Figure 2: The Regime Map

Note: Regime at each (c, v) in pure strategies under the Pareto Selection (Lemma 5); illustrative parameters $q_0 = 0.75$, $\delta = 0.15$, $b_C = b_O = 1$, $\sigma = 0.30$, $\bar{k} = 0.50$. White: anonymous regime, $c \geq \bar{c}$. Blue: verified regime, $v \in \mathcal{D}(c)$. Grey: dead forum. The verified/dead boundary falls in c (Lemma 4).

Figure 2 shows the resulting three regimes. Costly fabrication sustains the anonymous forum at every offered stringency because citizens never pay for a credential they do not need. Below $\bar{c}$, stringencies above the deterrence boundary sustain a verified forum populated up to its endogenous cutoff. Stringencies below that boundary produce a dead forum: everyone knows the channel can be flooded, the public does not heed it, and genuine citizens therefore remain silent even when the operator posts nothing. As fabrication cheapens, remaining in the live region requires climbing to greater stringency.

The Equilibrium

The first equilibrium result provides the pre-AI benchmark. When fabrication is costly, the anonymous chorus authenticates itself: swamping a fully attended forum costs more than the resulting belief shift is worth. Verification then has no protection to buy and only adds exposure.

Proposition 1 (The Pre-AI Benchmark): When fabrication is costly enough that the anonymous forum defends itself ($c \geq \bar{c}$), the anonymous regime obtains at every verification stringency: no citizen verifies at any stringency on offer, every

citizen with a favoring signal posts without paying exposure, and the public belief attains the highest accuracy available in any regime, unaffected by the stringency.

The operator remains idle because the required swarm costs more than its belief effect is worth. Every citizen with a favorable signal posts anonymously, so the forum collects the largest genuine state separation available in the model. Citizens obtain full participation without exposure, and the public receives the highest attainable accuracy. Offering deeper verification does not change any equilibrium object because the verified channel attracts no poster.

Now let fabrication be cheap enough to compromise the anonymous channel and let verification deter the swarm. Formal work on costly pseudonyms and evidence from transaction-verified review sites motivate the screening intuition that a stricter credential raises the cost of fabricated identities (Friedman and Resnick 2001; Mayzlin, Dover, and Chevalier 2014). That intuition is correct for authenticity everywhere, and the disciplining channel it invokes is general. Where political information is unrestricted, the model leaves only that channel. Consistent with this benchmark, Huang et al. (2026) find a stronger association between aggregate discussion sentiment and subsequent returns after real-name verification even as participation falls; this is not a direct measure of each post's accuracy or of what the voice count reveals. For claims involving politically restricted information, where greater access raises both exposure and knowledge, the added channel removes the most precise signals, the inversion the propositions formalize. For the verified forum, write authenticity as $\pi(v)$, mean poster precision as $\bar{q}(v) = \mathbb{E}[q(\theta) \mid \theta \leq t(v)]$, and accuracy as $Q(d(v))$.

Proposition 2 (Composition of the Certified Forum): Suppose the cost of exposure rises with knowledge ($\delta > 0$) and some feasible stringency silences ($\bar{v} < 1$). In the verified regime, authenticity is pinned at its maximum ($\pi(v) = 1$); yet past the silencing threshold ($v > \bar{v}$), a stricter credential strictly lowers the mean precision of the heard voices, the evidentiary weight each carries, and the accuracy

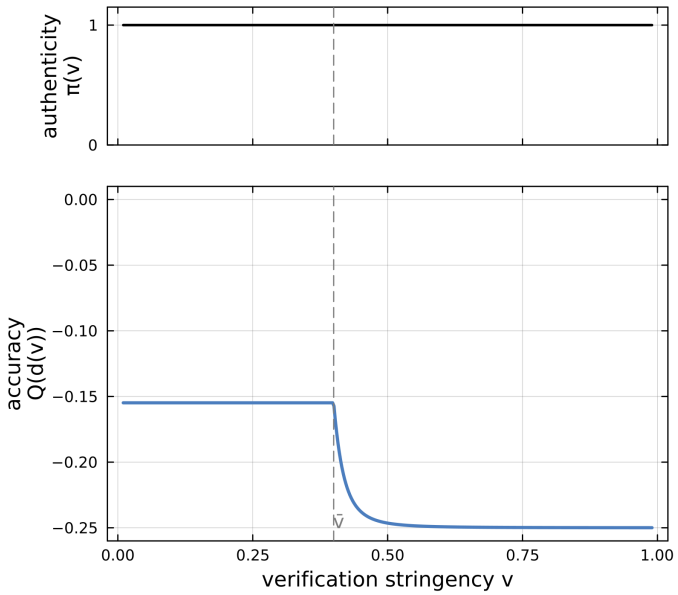

Figure 3: Authenticity versus Accuracy

Note: Authenticity and accuracy across the verified branch; $q_0 = 0.55$, $\delta = 0.35$, $b_C = 0.5$, $\sigma = 0.30$, chosen for visibility, not calibration. Top: authenticity $\pi(v)$, pinned at 1 (Proposition 2). Bottom: accuracy $Q(d(v))$, flat while $v \leq \bar{v}$, then falling as v deepens.

of the public belief. On average the silenced know more than the heard, by the constant margin $\frac{\delta}{2}$: zero when exposure cost is unrelated to knowledge, reversed when the two are opposed.

Two forces rise with insidership. Better information strengthens the private motive to post, while identifiability raises the cost of verified participation. At low stringency the motive dominates for every type, so authentication is compositionally free. Past $\bar{v}$, the exposure gradient becomes steeper than the posting motive, and the first voice lost at each additional depth is the best-informed one still present. Figure 3 shows the divergence: authenticity remains fixed at one across the verified branch, while accuracy is flat before silencing and falls afterward.

Three objects therefore move differently with stringency. Authenticity asks whether a heard voice is genuine; deterrence fixes that share at one throughout the verified regime. Mean precision asks how informative the surviving genuine poster is; it falls because a lower

participation cutoff removes the highest types. State separation combines composition with volume; it falls both because fewer favorable-signal citizens post and because the missing voices carry the strongest signals. The public observes this last object, not authenticity alone. A platform can therefore report perfect identity integrity while the experiment available to its audience becomes strictly worse. The result is not that genuine voices become less truthful, but that authentication changes which genuine signals enter the count.

This is inverted screening. A conventional costly credential separates high-quality participants because quality makes the credential easier to bear or more valuable to acquire. Here the credential screens on safety from identification. In the leading case, safety is negatively associated with the quality relevant to the forum: bystanders can authenticate cheaply, while insiders purchase the same credential by accepting a greater expected loss. Greater stringency therefore makes identity more reliable and participant information less reliable at the same time. The zero-sorting and negative-sorting cases identify the source of the reversal. Verification still reduces volume when precision is independent of exposure, but it does not change mean precision; when the best informed are safest, the remaining pool improves in composition even though total state separation can still fall.

The cutoff poster exits when her private benefit just covers exposure, yet her departure lowers the experiment's state separation. This is the positive externality at the center of the result. Citizens value collective accuracy through the common payoff in their utility. But no atomless citizen changes aggregate accuracy alone, so that term cancels from her posting comparison: she bears the exposure cost without internalizing the marginal public value of her signal, while a positive mass of identical decisions nevertheless reduces accuracy. The Supporting Information formalizes this continuum cancellation (§E.4. Direct Concern for Accuracy in Citizen Utility) and then lets a citizen's own post affect accuracy in a finite source pool. Holding fixed a deterred verified forum and a specified equilibrium path, that analysis (§E.4.1. Finite Source Pools and Pivotality) gives the exact participation condition. A scarce informed poster remains active throughout exactly when her pivotality-adjusted participation threshold stays above the policy along that path ($v \leq v_i^N(v)$ for every $v \in$

$[0, 1)$); if this condition fails, exposure induces her exit at the first violated stringency. When a small informed pool makes pivotality strong enough to keep its members posting, the composition result fails: verification no longer drives out the best informed.[2]

The externality also clarifies why private motivation does not protect the most valuable posting. High-precision citizens already receive the largest correctness benefit, and that force is included in their decision. What they do not capture is the improvement their participation creates in the audience's posterior. They can therefore be the most willing posters in the absence of identification and the first to withdraw once exposure becomes sufficiently steep. A policy that observes only voluntary participation may interpret their exit as low demand for posting even though their information has the highest social value.

Socially Optimal Verification Stringency

The results so far describe the equilibrium at every fixed verification stringency. The analysis now returns to stage 0 to ask what verification level society would implement through democratic choice. The welfare problem evaluates each feasible depth by the equilibrium consequences citizens collectively face.

Welfare is utilitarian over genuine citizens: it sums the posting surplus and the accuracy of the public belief, $Q(v) := Q(d(v)) = -\mathbb{E}[(\omega - \mu)^2] \in \left[-\frac{1}{4}, 0\right]$, the same accuracy object as in Lemma 2 evaluated at the discriminability induced by v, weighted by $\gamma > 0$; because stage 0 precedes the type draw, this welfare also equals each citizen's ex-ante expected utility and identifies the unanimous democratic choice:

[2] Information value may itself rise as the informed pool shrinks (a private order, a closed meeting, or a one-off transaction), making the finite-population case especially relevant. The composition result then fails when the resulting pivotal accuracy benefit keeps informed posters active throughout the relevant verification range. This is a bounded scope condition, not a general reversal: precision and prevalence are distinct, recurring or systemic claims can generate many precise signals (routine accounting manipulation, standing operational directives, or organization-wide compliance failures), and a smaller pool only attenuates the aggregate loss unless pivotality dominates exposure.

$$W(v) = \underbrace{\int_0^1 \mathbb{E}[u_i^*]\ \mathrm{d}\theta}_{\text{posting surplus } S(v)} + \gamma\, Q(v), \quad \gamma > 0 \tag{6}$$

The common accuracy term is part of each citizen's utility, not an external planner add-on. The first term aggregates citizens' private correctness benefits net of exposure; the second values public accuracy. An atomless citizen's decision does not change Q. The objective contains no reward for collecting identity records. In the anonymous regime, welfare is constant at $\frac{b_C}{2}\Delta(1) + \gamma Q(d_0)$. Call $c \geq \bar{c}$ the *privacy region* and $c < \bar{c}$ the *verification region*.

Society therefore faces a simple policy choice: preserve anonymity when it is safe, or impose only as much verification as deterrence requires. Proposition 3 characterizes the verification level selected through democratic choice. When fabrication is costly, society selects no verification. Once fabrication becomes cheap enough to threaten the anonymous forum, it selects the shallowest stringency that deters flooding. Stronger verification only raises exposure and, after silencing begins, reduces public accuracy. As fabrication becomes cheaper, the democratically selected stringency rises and may eventually silence the best-informed posters.

PROPOSITION 3 (The Minimum-Deterrence Optimum): In pure strategies, under the Pareto Selection and with ties broken toward lower verification, the citizen-welfare problem has a unique selected solution $v^*(c)$. (1) In the privacy region ($c \geq \bar{c}$), citizens choose no verification ($v^*(c) = 0$). (2) In the verification region ($c < \bar{c}$), they choose the shallowest deterring stringency ($v^*(c) = \underline{v}(c) > 0$); anything shallower yields the strictly worse dead forum under this selection. (3) The chosen stringency strictly deepens as fabrication cheapens. It passes the silencing threshold at a positive cost exactly under the two crossing conditions in Lemma 4. (4) Welfare is continuous at the boundary between the regions.

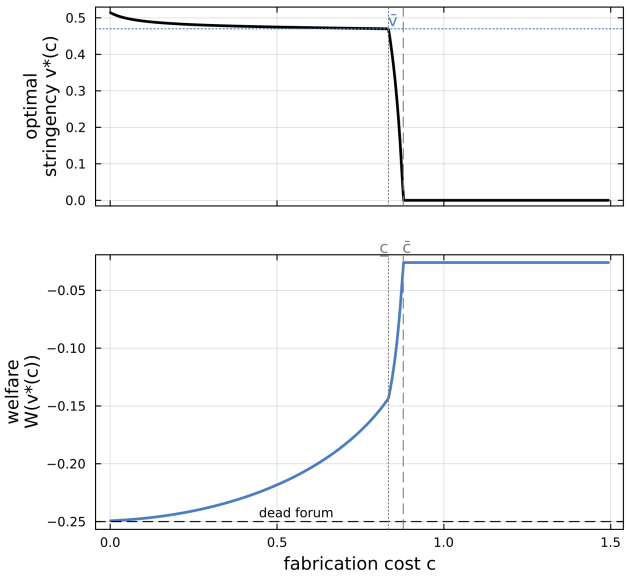

Figure 4: The Optimal Verification Stringency

Note: The selected pure-strategy verification stringency $v^*(c)$ and its welfare $W(v^*(c))$ against fabrication cost c; $q_0 = 0.51$, $\delta = 0.46$, $b_C = 0.5$, $\sigma = 0.30$, $\bar{k} = 0.05$, chosen for visibility, not calibration. Top: $v^*(c) = 0$ above $\bar{c}$, rising as fabrication cheapens and crossing $\bar{v}$ at $\underline{c}$. Bottom: welfare under the same pure-strategy equilibrium selection.

Figure 4 traces the pure-strategy optimum under parameters that satisfy the crossing conditions. Verification is unused in the privacy region because the anonymous forum already deters flooding. Once fabrication crosses the self-defense threshold, democratic choice selects the least costly depth that keeps a credentialed forum alive; greater depth adds exposure without further protection. As fabrication continues to cheapen, the selected depth crosses the silencing threshold and composition begins to change.

Minimum deterrence follows from the ordering of live verified outcomes. Holding the operator out, additional stringency does not improve authenticity: it is already complete. Before silencing, deeper identity only raises exposure. After silencing, it also removes posting surplus and lowers public accuracy. Citizens therefore stop at the first deterring depth. The discontinuity occurs at the boundary with the anonymous regime. Just above the self-defense cost, no identity is needed because organic participation deters the operator;

just below it, a positive credential depth must replace the lost persona cost. Welfare remains continuous because the required depth converges to zero at that boundary.

The pure-strategy restriction matters for the policy result. Once the operator can randomize, a compromised anonymous forum can retain information and private posting surplus. The Supporting Information constructs such equilibria (§C.1. Existence) and shows that their accuracy converges to the uninformative floor as persona costs vanish (§C.2. The Accuracy of the Compromised Channel). Away from that limit, its bound does not select among mixed equilibria or rank their total welfare against verification. Proposition 3 therefore characterizes the benchmark without operator randomization rather than a universal policy threshold across all equilibria of the subsequent posting game.

Allowing mixed strategies can change whether verification is adopted, but not the fixed-policy composition result. If citizens choose a deterring verification level, authenticity remains one; beyond the silencing threshold, stronger linkable verification raises exposure and reduces participation and accuracy.

No synthetic voice appears in the verified equilibrium. The verified channel deters the operator, while the anonymous channel has no standing. For the claims involving politically restricted information studied here, the flooding capability still changes the outcome because citizens anticipate what an unprotected forum would invite and what linkable verification would expose. This prediction rests on a Bayesian audience and an operator paid for belief in one focal claim. Campaigns aimed at distraction or flooding use a different payoff register (King, Pan, and Roberts 2017; Roberts 2018). The benchmark removes direct deception to isolate the verification-induced change in poster composition.

REMARK 1 (When the Backfire Region Is Nonempty): The selected optimal path reaches silencing depths exactly when some feasible verification stringency silences ($\bar{v} < 1$) and the forgery premium at that threshold cannot deter a fully attended forum ($k(\bar{v}) < \bar{c}$). When both hold, the backfire region is $c < \underline{c}$. If either fails, verification silences no one and leaves composition unchanged.

Both crossing conditions have substantive content. The first requires a feasible credential whose exposure cost can outweigh even the strongest private motive to post: identification can cost a citizen her job, legal position, or physical safety. The second requires the forgery technology to remain effective at moderate depth, so deterring a fully attended forum takes more than the premium imposed at the onset of silencing. If retaliation stakes are low, or if modest checks already make personas expensive, the defended forum remains fully populated and exhibits no survivor-composition shift.

The two conditions locate distinct thresholds. The silencing threshold $\bar{v}$ marks where selection begins; the forgery premium $k(\bar{v})$ determines whether deterrence reaches that point while participation is still complete. If $k(\bar{v}) \geq \bar{c}$, the deterrence set intersects the full-participation branch, so verification can deter without crossing into informed exit. If $k(\bar{v}) < \bar{c}$, falling fabrication cost eventually exhausts that branch at $\underline{c}$. Every further decline in cost then requires moving along a verified branch on which participation is already thinning. Backfire is therefore not a generic consequence of adopting identity verification. It begins only when adversarial cost pressure pushes the selected policy past the poster-side participation threshold.

PROPOSITION 4 (The Composition Backfire Along the Selected Policy Path): Suppose the cost of exposure rises with knowledge ($\delta > 0$) and the two crossing conditions in Remark 1 hold. As fabrication cheapens below $\underline{c}$ ($c < \underline{c}$) and the operator does not randomize, every heard voice remains genuine ($\pi(v^*(c)) = 1$),

yet the mean precision of the heard and the accuracy of the public belief strictly fall.

Inside the backfire region, cheaper synthetic production forces the selected policy to use stronger identity checks. Every heard voice remains genuine, yet the survivor pool loses its most precise signals. Figure 5 shows this separation in its top panel: authenticity remains at its maximum while mean survivor precision falls. The lower panel carries the composition change into the corresponding decline in public accuracy.

Proposition 4 composes three mappings. Cheaper fabrication raises the minimum stringency required for deterrence. Greater stringency lowers the participation cutoff after the silencing threshold. A lower cutoff reduces both the precision of the surviving pool and the state separation in observed support, which lowers public accuracy. These steps occur while deterrence keeps the synthetic mass at zero and authenticity at one. The operator benefits in the false state because the public receives a less discriminating experiment, not

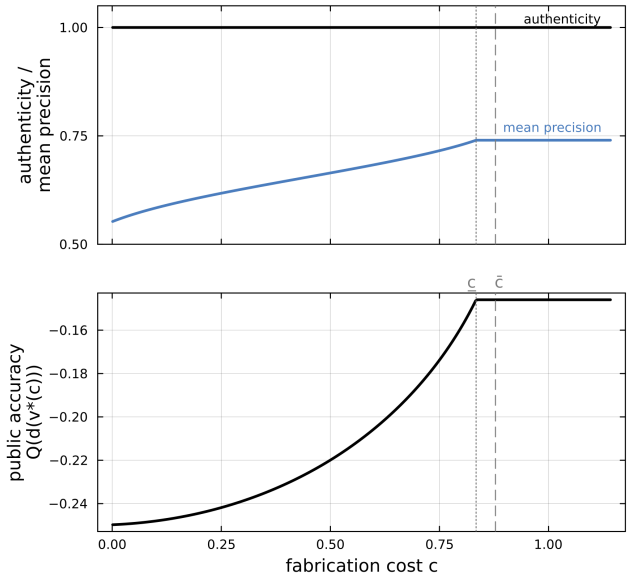

Figure 5: Authenticity, Poster Knowledge, and Public Accuracy

Note: Walking the selected pure-strategy policy $v^*(c)$ as fabrication cost falls; same parameters as Figure 4. Top: authenticity remains 1, while mean heard precision falls once $v^*(c)$ passes $\bar{v}$. Bottom: public accuracy $Q(d(v^*(c)))$ falls over the same region.

because a fabricated message enters that experiment. The threat therefore succeeds by degrading the institution's information supply.

Formally, under the even prior, Proposition 4 expresses the operator's false-state belief payoff as $\mathbb{E}[\mu \mid \omega = 0] = -2Q(d)$. Because Q falls as c falls in the backfire region, the operator's payoff rises even though equilibrium spending remains zero. This equation supplies the final arrow in the chain, not a separate mechanism.

Stage 0 assigns the policy to citizens and therefore excludes a direct institutional benefit from identity records. A state or platform that values those records, or benefits from insider silence, has a different objective. Past $\bar{v}$, each increase in stringency buys additional silence at an accuracy cost the setter need not bear. In that environment, the same comparative static describes deliberate informational control rather than a citizen-chosen defense that backfires (Guriev and Treisman 2019).

Defenses Against Fabricated Participation

The model directly covers, or readily extends to, several defenses against fabricated participation. Detection and content evidence change the effective price of a fake voice or the information observed by the public; crowd verification changes the cost and exposure created by eligibility requirements; and per-post costs burden each contribution without requiring identity linkage. Generative-AI advances can erode detection, content checks, and crowd-verification gates, while per-post costs can suppress genuine participation. These measures can soften but do not solve the model's central dilemma: deterring fabricated voices without silencing informed genuine posters.

Detection and content evidence. Let $p \in [0, 1)$ be the share of fakes removed independently before observation. The benchmark assumes perfect precision, so no genuine voice is removed, and it assumes removal before a persona creates exposure or enters public belief. Detection also precedes verified enrollment, so only surviving personas pay the forgery premium. A surviving anonymous voice then costs $\frac{c}{1-p}$, while a surviving verified voice

also pays $k(v)$. With a continuum of independent personas, the survival share is deterministic and detection changes only the attacker's effective price. Real systems combine account scores, relational campaign evidence, and human review; false positives, correlated takedowns, and delayed removal break this exact equivalence (Shao et al. 2018).

> REMARK 2 (Detection Is a Component of the Persona Price): Under independent, perfect-precision, pre-observation removal at rate p, the model is equivalent to one without detection at the effective fabrication cost ($c_{\text{eff}} = \frac{c}{1-p}$). When the operator does not randomize, the anonymous forum therefore defends itself at lower production costs, and where verification is selected, better detection weakly lowers its optimal stringency.

Within the ideal benchmark, detection widens the selected privacy region: its boundary becomes $(1-p)\bar{c}$. On the full-participation verified branch, $\underline{v}(c,p) = k^{-1}\left(\bar{c} - \frac{c}{1-p}\right)$, which falls strictly with detection while positive. Past $\bar{v}$, the same reduction in required depth readmits high-exposure posters. Detection therefore protects composition when it charges only the attacker. A detector that mistakenly removes genuine posters or acts after exposure introduces a second participation cost and requires a different comparison.

The key difference is who bears the deterrence cost. Ideal detection raises only the operator's cost of placing a surviving fake. Linked verification also raises that cost but imposes exposure on every genuine poster. Better detection can therefore reduce the verification needed for deterrence and preserve informed participation.

Content evidence belongs to the same practical family but enters a different model margin. Account or campaign detection removes suspected synthetic voices and changes the effective persona price; fact-checking and provenance add a signal about truth, authenticity, or synthetic mass and therefore change the public's information (Acemoglu, Ozdaglar, and Siderius 2023; Cisternas and Vasquez 2022). Such evidence can improve accuracy without identity linkage, but it cannot reconstruct genuine signals lost through poster exit and

may be costly, slow, or unavailable. The two margins are complementary because both can reduce reliance on linkable verification.

Crowd verification. Crowd-verification systems are partial applications of the model. Eligibility, reputation, consistency, and cross-viewpoint agreement requirements can raise the cost of fielding synthetic raters (Dasgupta and Ghosh 2013; Miller, Resnick, and Zeckhauser 2005; Prelec, Seung, and McCoy 2017; Wojcik et al. 2022). These gates map to the forgery premium $k(v)$, but they need not create the linkable exposure θv borne by genuine raters. The composition result applies only when clearing or using a gate also makes genuine raters identifiable or vulnerable. Generative AI can lower the synthetic cost by emulating diverse subpopulations and coordinated ratings can push lower-quality notes across cross-perspective thresholds (Argyle et al. 2023; Selvam et al. 2026; Wen et al. 2026). Synthetic mass can then fabricate apparent cross-partisan consensus rather than recruit a comparably diverse human coalition, lowering c in the model.

Per-post costs. A uniform cost τ on each post raises the price of synthetic volume without exposing posters. It can represent money, computation, delay, or another resource burden, provided it imposes the same payoff cost on genuine and synthetic voices. When $\delta > 0$, the private correctness benefit rises with type while τ is flat, so the least-informed posters exit first and survivor composition improves. When $\delta = 0$, the resulting thinning is composition-neutral. Unlike a forgery premium, however, the cost also falls on every genuine poster and is bounded by willingness to participate. A sufficiently high-stakes operator can therefore absorb a cost that genuine posters cannot. Combining it with verification can reduce the required identity depth and reaches a cost-only corner when τ alone deters.

Let the verified channel impose τ alongside identity verification. Genuine posters bear both τ and their exposure cost, while each synthetic persona bears τ and the forgery premium. For costs low enough to preserve genuine participation, the selected policy remains the minimum needed to deter the operator. While participation is full, a higher τ reduces the required identity depth until the per-post cost alone deters. Once verification silences, τ also changes who posts, so its welfare effect depends on whether lower identity

exposure readmits more posters than the added participation burden removes. The after-silencing part assumes deterring the operator grows no easier as the forum becomes more informative.

> PROPOSITION 5 (Per-Post Costs and Verification Stringency): Suppose exposure does not fall with knowledge ($\delta \geq 0$), $0 \leq \tau < b_{C(2q_0-1)}$, and the forgery premium $k \in C^1$ satisfies $k'(v) > 0$ on the relevant range. Before silencing ($v^*(\tau, c) \leq \bar{v} - \tau$), a higher per-post cost lowers the identity verification needed to deter manipulation ($\mathrm{d}\frac{v^*}{\mathrm{d}}\tau < 0$) and raises welfare only while the exposure savings exceed the burden on genuine posters; once the cost alone deters ($v^* = 0$), further increases lower welfare. After silencing ($v^*(\tau, c) > \bar{v} - \tau$), provided deterring the operator grows no easier as the forum becomes more informative, the cost still lowers required verification and raises welfare only when lower identity exposure expands participation and the accuracy gain is sufficiently valuable.

Figure 6 illustrates both parts of the result. In the top panel, a higher per-post cost lowers the selected identity depth under both the mild and deep threats. In the bottom panel, this substitution raises welfare under the mild threat because the saved exposure exceeds the burden on genuine posters. Under the deep threat, the same substitution lowers welfare at the plotted accuracy weight because the participation and accuracy gains are too small to offset that burden.

The full-participation branch gives the cleanest substitution. One unit of per-post cost replaces one unit of forgery premium in the deterrence constraint, so the chosen identity depth falls at the slope of k. This lowers exposure but burdens every genuine favorable-signal poster: the cost helps while the saved exposure exceeds that burden. At the cost-only corner, further increases add no privacy and reduce posting surplus directly.

Past the silencing threshold, replacing identity exposure with a common per-post cost also changes who returns. Lower stringency attracts high-exposure, high-precision citizens, while τ discourages those whose private posting benefit is too small. If the forgery

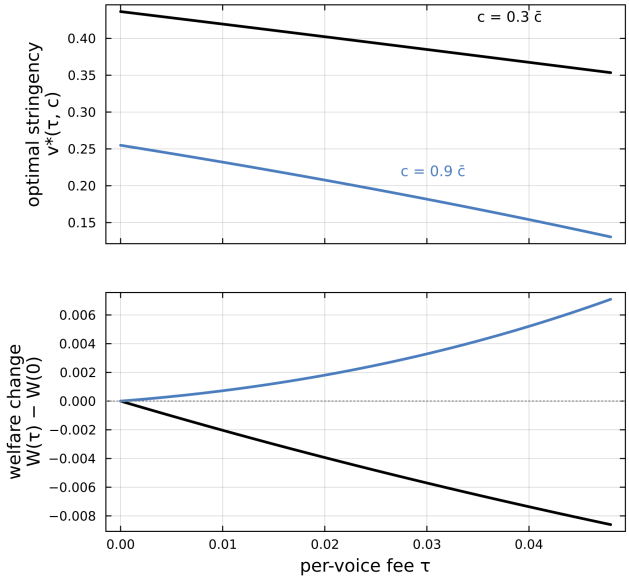

Figure 6: Per-Post Costs as a Supplement to Verification

Note: Optimal verification stringency and welfare as the verified channel adds a per-post cost τ, for a mild threat ($c = 0.9\bar{c}$) and a deep one ($c = 0.3\bar{c}$); $q_0 = 0.55$, $\delta = 0.35$, $b_C = 0.5$, $b_O = 1$, $\sigma = 0.30$, $\bar{k} = 0.25$, and $\gamma = 2$. Top: $v^*(\tau, c)$ falls with the cost over the plotted positive-stringency range; this figure does not reach the cost-only corner. Bottom: the welfare change $W(\tau) - W(0)$ rises with the cost under the mild threat and falls under the deep one.

technology is steep, a small increase in τ sharply lowers required stringency and widens the forum. If it is shallow, the identity reduction cannot offset the added cost, and participation contracts. A wider forum increases state separation, so the welfare sign in the deep branch depends on the accuracy weight rather than on substitution alone.

Per-post costs therefore avoid the identity linkage that generates the main adverse selection but do not provide a free defense. Their maximum sustainable level is bounded by genuine willingness to post, whereas the operator's return from manipulation can be large. They are most attractive as a supplement when a modest burden sharply reduces the identity depth required for deterrence. They are least attractive when the operator can absorb the cost, the forgery premium responds weakly, or genuine posters exit before the cost-only corner.

Conclusion

For claims involving politically restricted information, greater access both sharpens information and raises punishment risk if linkable authentication reveals the source. The paper shows that this alignment creates authentication-induced adverse selection under linkable verification. When the operator does not randomize, cheap fabrication leads democratic choice to select the mildest credential that deters a swarm. Once deterrence requires a silencing depth, the democratically selected defense removes the best-informed posters first and lowers public accuracy. Low retaliation stakes, effective moderate checks, or a sufficiently strong private reward for informational contribution preserve the composition of the forum. The policy comparisons separate instruments by which participants bear a cost. Ideal detection raises the attacker's effective production cost without burdening genuine posters. A per-post cost substitutes a non-identity burden for part of the identity requirement but also burdens genuine participation. Fact-checking and provenance add information to the public experiment. De-linked credentials prevent duplicate participation without creating the recoverable identity link that drives exposure. Evaluating a defense therefore requires tracking both the fabricated voices excluded and the genuine information displaced.

The mechanism yields a testable composition implication inside a treated institution. After a discrete linked-verification change, pre-policy contribution histories can provide observable markers of first-hand access among the surviving pool. A decline in participation without a change in those markers is consistent with composition-neutral exposure. A decline concentrated among contributors with first-hand access is consistent with authentication-induced adverse selection. Off-platform reports are informative substitution outcomes only where the treated forum and its users have a documented bridge to those reporting channels.

Three institutional questions remain. Who should control the identity records created by linkable verification? Should policy deter the acquisition of fabrication capacity when

an off-path threat changes behavior? And how can a jurisdiction impose a persona price on operators beyond its reach? Each question follows from the separation between authenticating a voice and preserving the knowledge supplied by its poster. Institutional design must assign custody, audit access, and limit compelled disclosure alongside the technical task of excluding duplicate identities.

AI Use Disclosure

OpenAI Codex, powered by GPT-5 (OpenAI; used in July 2026), assisted with copyediting and reviewing numerical code. The author independently reviewed, edited, and approved all AI-assisted material and checked the mathematical arguments, code, and reported results. The author assumes full responsibility for the manuscript.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2023. "A Model of Online Misinformation." *Review of Economic Studies* 91(6): 3117–50. doi:10.1093/restud/rdad111.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31(3): 337–51. doi:10.1017/pan.2023.2.

Bai, Hui, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. "LLM-Generated Messages Can Persuade Humans on Policy Issues." *Nature Communications* 16(1). doi:10.1038/s41467-025-61345-5.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. 2020. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences* 117(1): 243–50. doi:10.1073/pnas.1906420116.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics* 107(3): 797–817. doi:10.2307/2118364.

Battaglini, Marco. 2017. "Public Protests and Policy Making." *Quarterly Journal of Economics* 132(1): 485–549. doi:10.1093/qje/qjw039.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, And Cultural Change as Informational Cascades." *Journal of Political Economy* 100(5): 992–1026. doi:10.1086/261849.

Cisternas, Gonzalo, and Jorge Vasquez. 2022. "Misinformation in Social Media: The Role of Verification Incentives." *SSRN Electronic Journal*. doi:10.2139/ssrn.4191785.

Dasgupta, Anirban, and Arpita Ghosh. 2013. "Crowdsourced Judgement Elicitation with Endogenous Proficiency." In *Proceedings of the 22nd International Conference on World Wide Web,* , 319–30. doi:10.1145/2488388.2488417.

Dellarocas, Chrysanthos. 2006. "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms." *Management Science* 52(10): 1577–93. doi:10.1287/mnsc.1060.0567.

Edmond, Chris. 2013. "Information Manipulation, Coordination, And Regime Change." *Review of Economic Studies* 80(4): 1422–58. doi:10.1093/restud/rdt020.

Edmond, Chris, and Yang K. Lu. 2021. "Creating Confusion." *Journal of Economic Theory* 191: 105145. doi:10.1016/j.jet.2020.105145.

Feddersen, Timothy, and Wolfgang Pesendorfer. 1997. "Voting Behavior and Information Aggregation in Elections with Private Information." *Econometrica* 65(5): 1029. doi:10.2307/2171878.

Ford, Bryan. 2020. (arXiv) *Identity and Personhood in Digital Democracy: Evaluating Inclusion, Equality, Security, And Privacy in Pseudonym Parties and Other Proofs of Personhood*. . technical report. doi:10.48550/arXiv.2011.02412.

Friedman, Eric J., and Paul Resnick. 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics and Management Strategy* 10(2): 173–99. doi:10.1111/j.1430-9134.2001.00173.x.

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. (arXiv) *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* . technical report. doi:10.48550/arXiv.2301.04246.

Government of Japan. 2013. "Act on the Protection of Specially Designated Secrets." https://www.japaneselawtranslation.go.jp/en/laws/view/2543.

Guriev, Sergei, and Daniel Treisman. 2019. "Informational Autocrats." *Journal of Economic Perspectives* 33(4): 100–127. doi:10.1257/jep.33.4.100.

Huang, Kanyuan, Yakun Wang, T. J. Wong, and Tianyu Zhang. 2026. "User Anonymity and the Informativeness of Social Media: Evidence from a Natural Experiment." *Review of Accounting Studies* 31(2): 1051–87. doi:10.1007/s11142-026-09953-4.

Kang, Myunghoon, and Greg Chih-Hsin Sheen. 2025. "The Making of the Boy Who Cried Wolf: Fake News and Media Skepticism." *Political Science Research and Methods* 13(2): 465–74. doi:10.1017/psrm.2024.7.

Keller, Franziska B., David Schoch, Sebastian Stier, and JungHwan Yang. 2020. "Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign." *Political Communication* 37(2): 256–80. doi:10.1080/10584609.2019.1661888.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111(3): 484–501. doi:10.1017/s0003055417000144.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(1): 7–48. doi:10.2307/2010422.

Little, Andrew T. 2012. "Elections, Fraud, And Election Monitoring in the Shadow of Revolution." *Quarterly Journal of Political Science* 7(3): 249–83. doi:10.1561/100.00011078.

Little, Andrew T. 2014. "Fraud and Monitoring in Non-Competitive Elections." *Political Science Research and Methods* 3(1): 21–41. doi:10.1017/psrm.2014.9.

Lohmann, Susanne. 1993. "A Signaling Model of Informative and Manipulative Political Action." *American Political Science Review* 87(2): 319–33. doi:10.2307/2939043.

Matz, Sandra C., Jacob D. Teeny, Sumer S. Vaid, Henriette Peters, Gabriella M. Harari, and Moran Cerf. 2024. "The Potential of Generative AI for Personalized Persuasion at Scale." *Scientific Reports* 14(1): 4692. doi:10.1038/s41598-024-53755-0.

Mayzlin, Dina. 2006. "Promotional Chat on the Internet." *Marketing Science* 25(2): 155–63. doi:10.1287/mksc.1050.0137.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review* 104(8): 2421–55. doi:10.1257/aer.104.8.2421.

McMurray, Joseph C. 2013. "Aggregating Information by Voting: The Wisdom of the Experts Versus the Wisdom of the Masses." *The Review of Economic Studies* 80(1): 277–312. doi:10.1093/restud/rds026.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer-Prediction Method." *Management Science* 51(9): 1359–73. doi:10.1287/mnsc.1050.0379.

Nightingale, Sophie J., and Hany Farid. 2022. "AI-Synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy." *Proceedings of the National Academy of Sciences* 119(8). doi:10.1073/pnas.2120481119.

Office of the New York State Attorney General. 2021. (Office of the New York State Attorney General) *Fake Comments: How U.S. Companies & Partisans Hack Democracy.* . technical report. https://ag.ny.gov/sites/default/files/oag-fakecommentsreport.pdf.

Prelec, Dražen, H. Sebastian Seung, and John McCoy. 2017. "A Solution to the Single-Question Crowd Wisdom Problem." *Nature* 541(7638): 532–35. doi:10.1038/nature21054.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion inside China's Great Firewall*. Princeton University Press. doi:10.23943/9781400890057.

Ross, Björn, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. 2019. "Are Social Bots a Real Threat? An Agent-Based Model of the Spiral of Silence to Analyse the Impact of Manipulative Actors in Social Networks." *European Journal of Information Systems* 28(4): 394–412. doi:10.1080/0960085x.2018.1560920.

Rothschild, Joyce, and Terance D. Miethe. 1999. "Whistle-Blower Disclosures and Management Retaliation." *Work and Occupations* 26(1): 107–28. doi:10.1177/0730888499026001006.

Schroeder, Daniel Thilo, Meeyoung Cha, Andrea Baronchelli, Nick Bostrom, Nicholas A. Christakis, David Garcia, Amit Goldenberg, et al. 2026. "How Malicious AI Swarms Can Threaten Democracy." *Science* 391(6783): 354–57. doi:10.1126/science.adz1697.

Selvam, Nikil Roashan, Jay Baxter, Sophie Hilgard, Brad Miller, Keith Coleman, Ellen Vitercik, and Sanmi Koyejo. 2026. (arXiv) *Gaming Consensus: Coordinated Manipulation in Crowdsourced Fact-Checking*. . technical report. doi:10.48550/arXiv.2607.01824.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. "The Spread of Low-Credibility Content by Social Bots." *Nature Communications* 9(1). doi:10.1038/s41467-018-06930-7.

Spitale, Giovanni, Nikola Biller-Andorno, and Federico Germani. 2023. "AI Model GPT-3 (Dis)informs Us Better Than Humans." *Science Advances* 9(26). doi:10.1126/sciadv.adh1850.

Supreme Court of the Australian Capital Territory. 2024. "Judgment Summary." https://www.courts.act.gov.au/___data/assets/pdf_file/0005/2451371/McBride-No-4-Summary.pdf.

U.S. Department of Justice. 2018. "Federal Government Contractor Sentenced for Removing and Transmitting Classified Materials to a News Outlet." https://www.justice.gov/archives/opa/pr/federal-government-contractor-sentenced-removing-and-transmitting-classified-materials-news.

Wen, Changxi, Shuning Zhang, Bohao Chu, Yuwei Chuai, Hui Wang, Dai Shi, Xin Yi, and Hewu Li. 2026. (arXiv) *Towards Multi-Agent-Simulation-Based Community Note Evaluation*. . technical report. doi:10.48550/arXiv.2606.18268.

Wojcik, Stefan, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. 2022. (arXiv) *Birdwatch: Crowd Wisdom and Bridging Algorithms Can Inform Understanding and Reduce the Spread of Misinformation*. . technical report. doi:10.48550/arXiv.2210.15723.

Zakharov, Alexei. 2026. "Propaganda to a Cynical Audience." *Political Science Research and Methods*: 1–12. doi:10.1017/psrm.2025.10081.

Supporting Information

The proofs below are intended for online publication.

Contents

A. The Equilibrium Concept

For reference, the state is $\omega \in \{0,1\}$ with even prior. A unit mass of citizens has exposure type $\theta \in [0,1]$ and binary signal s, whose precision is $q(\theta) = q_0 + \delta\theta > \frac{1}{2}$. Citizens may post in the verified channel V, post in the anonymous channel A, or remain silent. In a heeded channel, a favorable-signal citizen receives private correctness benefit $b_C(2q(\theta)-1)$; verified posting also costs θv. Each synthetic voice costs the operator c in the anonymous channel and $c + k(v)$ in the verified channel. The public observes each channel's genuine mass plus synthetic mass and an independent $N(0,\sigma^2)$ shock, then forms the Bayesian posterior μ. The operator acts only when $\omega = 0$ and values raising that posterior net of persona spending. A channel has standing $\rho_\ell = 1$ exactly when its equilibrium mass has positive state separation.

Four objects must line up together. A *citizen strategy* says what each citizen does given who she is: it maps her type and signal to one of the three actions, $\alpha : [0,1] \times \{0,1\} \to \{V, A, \varnothing\}$. A *flooding rule* says how many synthetic voices the operator buys in each channel when the truth is against it, $\tilde{m} = (\tilde{m}_A, \tilde{m}_V) \in \mathbb{R}^2_+$ in state $\omega = 0$; the operator is idle in $\omega = 1$. A *posterior function* $\mu : \mathbb{R}^2 \to [0,1]$ says what the public concludes from the two observed masses. And the *standings* $(\rho_V, \rho_A) \in \{0,1\}^2$ record which channels the public's belief actually responds to. Given citizen behavior and flooding, channel ℓ looks different on average depending on whether the position is true or false. Call that difference its *net state-separation*, $\Delta_\ell = [m_\ell(1) + \tilde{m}_\ell(1)] - [m_\ell(0) + \tilde{m}_\ell(0)]$: it is the informativeness that survives the flooding.

Condition (iii) is the standard best-response requirement for the simultaneous game. The public is not a later strategic receiver whose posterior is recomputed after each counterfactual flood; its contingent rule is calibrated on path and, like every other equilibrium strategy function, held fixed during a unilateral-deviation check. Gaussian noise gives every $(M_V, M_A) \in \mathbb{R}^2$ positive density, so Bayes' rule pins that contingent posterior everywhere and leaves no off-path receiver beliefs to refine. The no-average-deception property follows from Bayesian iterated expectations.

Why the one-channel deviation problem is sufficient. Under (5) the conditional joint density factors across channels. Every equilibrium candidate used below has at most one channel with positive conjectured state-separation. If channel j has zero conjectured separation, its two

equilibrium conditional densities coincide, so its likelihood-ratio factor is identically one at every realization. A deviation that adds mass to j therefore leaves $\mu^*(M_V, M_A)$ unchanged, including when combined with a flood in the informative channel, while costing $\kappa_j \tilde{m}_j > 0$. Thus every operator best response sets $\tilde{m}_j = 0$, and the remaining vector deviation reduces exactly to Lemma 3′s one-channel problem. Independence is load-bearing: with correlated shocks an inactive channel could reveal the informative channel's noise, and this reduction would fail.

B. Preliminary Lemmas

B.1. The Compromised Outcome

If $\kappa < K(d)$ at conjectured discriminability d, no pure-strategy profile with $\Delta_\ell > 0$ satisfies the equilibrium definition for channel ℓ: the unique pure-strategy outcome has $\Delta_\ell = 0$ and $\rho_\ell = 0$, with any payoff-equivalent tie broken toward silence.

Proof. Step 1: silence satisfies the equilibrium definition. Take the silence profile in channel ℓ. It satisfies citizen optimality, since with $\rho_\ell = 0$ a voice there earns nothing and staying out is weakly best; belief consistency, since an empty channel has $\Delta_\ell = 0$ and hence $\rho_\ell = 0$; and operator optimality, since a channel with zero conjectured separation is trivially deterred (Lemma 3 (4)).

Step 2: no heeded profile survives. Consider any pure conjecture in which the operator floods state 0 by $\tilde{m}^*$; the channel's effective separation is then $\Delta - \tilde{m}^*$ and its discriminability $d' = \frac{\Delta - \tilde{m}^*}{\sigma}$, and being heeded requires $0 \le \tilde{m}^* < \Delta$. Grade the operator against this conjecture, writing σ_{lg} for the logistic function, φ for the standard normal density, and $G(d', y) = \int \sigma_{\mathrm{lg}}\big(-\frac{d'^2}{2} + y + d'z\big)\varphi(z)\,\mathrm{d}z$ for the state-0 belief it induces at discriminability d' by shifting the receiver's calibrated log-odds score by y, the object whose forward secant defines K in Lemma 3. The operator's marginal value of one more persona is $\partial_y G(d', y) = \int \sigma_{\mathrm{lg}'}\big(-\frac{d'^2}{2} + y + d'z\big)\varphi(z)\,\mathrm{d}z$. This weight is single-peaked in the flood: the value of one more persona first rises, reaches a peak at $y = \frac{d'^2}{2}$, an absolute flood of $\frac{\Delta + \tilde{m}^*}{2}$, and then falls. For any interior conjecture $\tilde{m}^* < \Delta$ that peak lies strictly above $\tilde{m}^*$, so at $\tilde{m}^*$ the marginal value is still rising, and this squeezes the operator's two deviations from opposite sides. Escalating is unprofitable

only if the persona price covers the best forward secant, $\kappa \geq K(d')$, which is Lemma 3′s own deterrence test at the reduced discriminability d'. Abandoning, cutting the flood back toward zero, is unprofitable only if the price stays below the average value of the personas already in place, $\kappa \leq \kappa_{\text{down}}$; and because the marginal value is still rising at $\tilde{m}^*$, that average, and hence κ_{down}, is strictly below the forward secant supremum that defines $K(d')$, so $\kappa_{\text{down}} < K(d')$. No single price is at once above $K(d')$ and below κ_{down}, so at every interior heeded conjecture the operator strictly prefers to escalate or to abandon, and the conjecture is not a fixed point. The endpoints are covered directly: $\tilde{m}^* = 0$ is Lemma 3′s no-flood test at d, which fails by the hypothesis of the compromised region, $\kappa < K(d)$; and $\tilde{m}^* \geq \Delta$ leaves nonpositive separation, so the channel is not heeded and $\tilde{m} = 0$ is strictly optimal (Lemma 3 (4)).

Step 3: uniqueness once ties are broken. The dead forum is therefore the only outcome in which channel ℓ is not heeded. At zero standing a citizen is indifferent between silence and posting in channel ℓ, so many posting patterns coexist, including signal-dependent ones that make the ignored channel anti-informative and so are not literally uninformative; but every such pattern pays exactly what silence pays. The tie-break toward silence, applied in fixing the compromised outcome before the Pareto Selection compares outcomes, forecloses them all, leaving pure silence with $\Delta_\ell = 0$. One limitation of the reduced form deserves note here: the binary standing $\rho_\ell \in \{0, 1\}$ cannot represent an anti-informative channel, since a channel with strictly negative separation would move a Bayes-consistent posterior with a negative slope, a response the standing has no value for. Such configurations unravel on their own terms: against a negative calibrated slope the operator strictly prefers $\tilde{m} = 0$ (Lemma 3 (4)), and the tie-break then empties the channel to $\Delta_\ell = 0$, restoring the case the binary standing does cover. □

B.2. Proof of Lemma 1

Proof. By the citizen payoff a citizen with $s_i = 0$ gets $-b_C(2q(\theta) - 1)\rho_\ell - \theta v \cdot \mathbb{1}\{\ell = V\} \leq 0$ from posting, strictly negative in any heeded channel (since $q > 1/2$ by the bounds on q_0, δ), so she does not post where it matters. Take $s_i = 1$.

(1) With $\rho_A = 1$, anonymous posting pays $b_C(2q(\theta) - 1) > 0$: strictly better than silence, and weakly better than verified posting, whose payoff is at most $b_C(2q(\theta) - 1) - \theta v$ (strictly

worse for $\theta > 0$, any $v > 0$). Equality between channel labels at $v = 0$ is resolved toward anonymity by the convention in Timing.

(2) With $\rho_A = 0$, anonymous posting is worth 0, so the choice is verified-vs-silence with net payoff $\psi(\theta) \coloneqq b_C(2q_0 - 1) + (2b_C\delta - v)\theta$, affine in θ with slope $2b_C\delta - v$ and intercept $b_C(2q_0 - 1) > 0$. *Case* $v \le \hat{v}$*:* slope ≥ 0, so $\psi(\theta) \ge \psi(0) > 0$: all types post, $t = 1$, and claim (i) holds. *Case* $v \in (\hat{v}, \bar{v}]$*:* slope < 0, but $\psi(1) = b_C(2q(1) - 1) - v \ge 0$ by definition of $\bar{v}$: all types still post, $t = 1$, and claim (ii) holds. *Case* $v > \bar{v}$*:* slope < 0 and $\psi(1) < 0$, so ψ has the unique root $t(v) = b_C(2q_0 - 1)/(v - 2b_C\delta) \in (0, 1)$; types $\theta \le t(v)$ post and types above are silent. Differentiating, $t'(v) = -b_C(2q_0 - 1)/(v - 2b_C\delta)^2 < 0$; continuity at $\bar{v}$ is immediate since $t(\bar{v}^+) = b_C(2q_0 - 1)/(\bar{v} - 2b_C\delta) = 1$.

(3) Immediate from the payoffs. With the stated tie-breaks these best responses are unique except for the measure-zero marginal type. □

B.3. Proof of Lemma 2

Proof. By the law of total variance, $\mathbb{E}[(\omega - \mu)^2] = \mathbb{E}[\mathrm{Var}\,(\omega \mid M)] = \mathbb{E}[\mu(1 - \mu)]$, giving the first identity; at $d = 0$ the posterior is the prior $1/2$, so $Q(0) = -1/4$; as $d \to \infty$ the posterior converges to ω a.s. and $Q \to 0$ by bounded convergence.

Strict monotonicity. Take $d' < d$. Any Gaussian binary experiment with discriminability d is equivalent, after an affine map, to observing $Y = \omega d + Z$ with $Z \sim N(0, 1)$; work with this form. Construct

$$Y' = \left(\frac{d'}{d}\right) Y + \sqrt{1 - (d'/d)^2}\, U, \quad U \sim N(0, 1) \text{ independent.}$$

Then $Y' = \omega d' + Z'$ with $Z' \sim N(0, 1)$: the d'-experiment is an exact garbling of the d-experiment, and $\omega \perp Y' \mid Y$. By the law of total variance applied to $\mu(Y) = \mathbb{E}[\omega \mid Y]$ conditional on Y':

$$\mathbb{E}[\mathrm{Var}\,(\omega \mid Y')] = \mathbb{E}[\mathrm{Var}\,(\omega \mid Y)] + \mathbb{E}[\mathrm{Var}\,(\mu(Y) \mid Y')].$$

The second term is strictly positive: given Y', Y remains non-degenerate (the added noise U has positive variance since $d' < d$), and $\mu(Y)$ is strictly increasing in Y (its logit is linear in Y

with positive slope d), so $\mu(Y)$ is not Y'-measurable. Hence $\mathbb{E}[\mathrm{Var}\,(\omega|Y')] > \mathbb{E}[\mathrm{Var}\,(\omega|Y)]$, i.e. $Q(d') < Q(d)$.

For later use (Proposition 5 differentiates through Q), Q is continuously differentiable on $(0,\infty)$ with $Q'(d) > 0$. In the $Y = \omega d + Z$ form, substituting $u = y - d/2$ symmetrizes the posterior-variance integral into

$$\mathbb{E}[\mu(1-\mu)] = \frac{1}{4\sqrt{2\pi}} \int \frac{e^{-(u^2+d^2/4)/2}}{\cosh\left(u\frac{d}{2}\right)}\,\mathrm{d}u,$$

which equals $1/4$ at $d = 0$, as it must. The d-derivative of the integrand is the integrand times $-[d/4 + (u/2)\tanh(ud/2)]$, dominated locally in d by a Gaussian envelope and strictly negative pointwise for $d > 0$, both bracket terms being nonnegative and the first strictly positive. Differentiation under the integral sign therefore gives $Q \in C^1$ with $Q'(d) > 0$. □

B.4. Proof of Lemma 3

Proof. Write $\sigma_{\mathrm{lg}}(x) = \frac{1}{1+e^{-x}}$ and, for the operator's expected induced state-0 posterior when it shifts the receiver's (fixed, candidate-calibrated) log-odds score by $y \geq 0$, $G(d,y) = \int \sigma_{\mathrm{lg}}\left(-\frac{d^2}{2} + y + dz\right)\varphi(z)\,\mathrm{d}z$, with φ the standard normal density; then $K(d) := \left(b_O\frac{d}{\sigma}\right)\sup_{y>0}\frac{[G(d,y)-G(d,0)]}{y}$.

(1) Under the no-flood candidate the receiver's posterior is $\mu(M) = \sigma_{\mathrm{lg}}(L(M))$ with log-odds $L(M) = (\Delta/\sigma^2)(M - (m(1)+m(0))/2)$ (the standard equal-variance Gaussian log-likelihood ratio). If the operator deviates to $\tilde{m} > 0$ in state 0, then $M = m(0) + \tilde{m} + \sigma Z$ with $Z \sim N(0,1)$, so

$$L = \frac{\Delta}{\sigma^2}\left(m(0) + \tilde{m} + \sigma Z - \frac{m(1)+m(0)}{2}\right) = -\frac{d^2}{2} + \frac{d}{\sigma}\tilde{m} + dZ.$$

Writing $y = (d/\sigma)\tilde{m}$ (so $\tilde{m} = \sigma y/d$ and the spend is $\kappa\sigma y/d$), the deviation payoff net of the candidate payoff is

$$b_O[G(d,y) - G(d,0)] - \frac{\kappa\sigma}{d}y.$$

The candidate survives iff this is ≤ 0 for all $y > 0$, which rearranges to $\kappa \geq K(d)$ as stated. Downward deviations do not exist at $\tilde{m} = 0$.

(2) Let $R(d,y) \coloneqq [G(d,y) - G(d,0)]/y$ for $y > 0$. Since $\partial_y G(d,y) = \int \sigma'_{\lg}(-d^2/2 + y + dz)\varphi(z)\,\mathrm{d}z \in (0, 1/4]$, the fundamental theorem of calculus gives

$$R(d,y) = \int_0^1 \partial_y G(d, uy)\,\mathrm{d}u \in (0, 1/4],$$

so R extends continuously to $y = 0$ with $R(d,0) = \partial_y G(d,0) > 0$ and is bounded by $1/4$; hence $K(d) = (b_O d/\sigma)\sup_{y\geq 0} R(d,y)$ is finite, positive, and $\leq b_O d/(4\sigma)$. For continuity: $\partial_y G$ is jointly continuous and bounded (dominated convergence), so R is jointly continuous on $[0,\infty)^2$ (via the integral representation above); moreover $R(d,y) \leq 1/y \to 0$ as $y \to \infty$ uniformly in d, so on any compact set of d's the supremum over y is attained on a common compact interval $[0, \bar{y}]$; the maximum of a jointly continuous function over a fixed compact set is continuous in the parameter (Berge), giving continuity of K.

(3) The first branch restates (1) plus Bayes-consistency: with $\tilde{m}(0) = 0$ conjectured and played, M's state means are Δ apart and the receiver's posterior is exactly the d-experiment posterior, so accuracy is $Q(d)$ by Lemma 2. The second branch is the compromised outcome (no heeded pure-strategy equilibrium exists).

(4) Let the conjectured separation be $\Delta^* \leq 0$, so the channel's calibrated log-odds contribution has slope $\Delta^*/\sigma^2 \leq 0$ in its observation. If $\Delta^* = 0$ the slope is zero: any deviation mass leaves the posterior unchanged and costs $\kappa\tilde{m} > 0$. If $\Delta^* < 0$ the slope is strictly negative: flooding raises M in state 0, which strictly *lowers* the posterior the operator wants to inflate, and still costs $\kappa\tilde{m} > 0$. In both cases $\tilde{m} = 0$ is the strict best response.

(5) *Shape.* Write $S(d) \coloneqq \sup_{y>0} R(d,y)$, so $K(d) = (b_O/\sigma)\,d\,S(d)$. Within the maintained Gaussian specification, $dS(d)$ is a dimensionless curve scaled by $\frac{b_O}{\sigma}$.

Differentiability. Write $f_d(s) \coloneqq \partial_y G(d,s) = \int \sigma_{\lg'}(-d^2/2 + s + dz)\varphi(z)\,\mathrm{d}z$: the convolution of the logistic density $\sigma_{\lg'}$ with the $N(0,d^2)$ density φ_d, evaluated at $s - d^2/2$. Both densities are symmetric and log-concave, and log-concavity is preserved under convolution, so f_d is symmetric about $s = d^2/2$ and single-peaked there (the same convolution structure the compromised-region argument uses). The monotonicity is strict on each side of the peak: for $u > 0$, folding the convolution derivative about zero gives

$$f_{d'}(d^2/2+u) = \int_0^\infty \sigma_{\lg''}(x)[\varphi_d(u-x) - \varphi_d(u+x)]\,\mathrm{d}x < 0$$

strictly, since $\sigma_{\lg''}$ is odd with $\sigma_{\lg''}(x) < 0$ for $x > 0$ while $\varphi_d(u-x) > \varphi_d(u+x)$; symmetry handles $u < 0$. By the integral representation in (2), $R(d,y) = (1/y)\int_0^y f_d(s)\,\mathrm{d}s$ is the running average of f_d. On $(0, d^2/2]$, f_d is strictly increasing, so $R(d,y) < f_d(y)$ and $\partial_y R = (f_d(y) - R)/y > 0$: no critical point there, and the maximizer exceeds $d^2/2$. At any critical point $y > d^2/2$, $\partial_y R = 0$ forces $R = f_d(y)$, whence $\partial_y^2 R = f_{d'}(y)/y < 0$ strictly: every critical point is a strict local maximum, so there is exactly one, the interior maximizer $y^*(d)$ (interior because $R(d,y) \to f_d(0) < \sup R$ as $y \to 0$ and $R \le 1/y \to 0$ as $y \to \infty$). R is jointly C^1 (differentiate $R = \int_0^1 f_d(uy)\,\mathrm{d}u$ under the integral; $\sigma_{\lg'}$ and $\sigma_{\lg''}$ are bounded, so dominated convergence applies), the maximizer is unique and, on a compact neighborhood of any d, confined to a common compact interval (by $R \le 1/y$ and continuity), so Danskin's envelope theorem gives S differentiable with $S'(d) = \partial_d R(d, y^*(d))$, continuous because $y^*(\cdot)$ is continuous (Berge plus uniqueness). Hence $K \in C^1$ on $(0,\infty)$.

Limit at $d \to 0$. Part (2) gives $S(d) \le 1/4$. For the matching lower bound take $y = d$: as $d \to 0$, $f_d(s) \to \sigma_{\lg'}(s)$ uniformly on compacts (dominated convergence), so $R(d,d) = (1/d)\int_0^d f_d(s)\,\mathrm{d}s \to \sigma_{\lg'}(0) = 1/4$. Hence $S(d) \to 1/4$ and $K(d)/d \to b_O/(4\sigma)$: the bound rises from zero at slope $b_O/(4\sigma)$.

Limit at $d \to \infty$. Split the supremum at $y = d^2/4$. For $y \ge d^2/4$: $R(d,y) \le 1/y \le 4/d^2$, so $dR \le 4/d$. For $0 < y < d^2/4$: a running average is at most the supremum of its integrand, and by single-peakedness at $d^2/2$ the supremum of f_d over $[0, d^2/4]$ sits at the right endpoint, so $R(d,y) \le f_d(d^2/4) = \mathbb{E}\big[\sigma_{\lg'}(W)\big]$ with $W \sim N(-d^2/4, d^2)$. Since $\sigma_{\lg'}(w) \le \min\{1/4, e^w\}$,

$$f_d(d^2/4) \le \frac{1}{4}\Pr(W > -d^2/8) + e^{-d^2/8} = \frac{1}{4}\Phi(-d/8) + e^{-d^2/8},$$

so $dR \le d\big[\Phi(-d/8)/4 + e^{-d^2/8}\big] \to 0$. Both pieces vanish, so $K(d) \to 0$: pooling at large d needs a log-odds shift of order d^2, a mass of order σd, whose price outruns the belief it buys.

Continuity, strict positivity, and the endpoint limits prove that K attains at least one interior maximum; they do not prove that K' changes sign only once. In the maintained Gaussian specification, numerical evaluation on 656 points over $(0, 40]$ yields one finite-difference

sign change and a grid-supported peak $\hat{d} \approx 2.54$ (script `16_kd_shape_pilot.jl`). This finite grid cannot rule out sub-grid turns. Every later analytical claim therefore imposes the required sign of K' directly on the full discriminability range it uses. □

B.5. Proof of Lemma 4

Proof. Closedness: $v \mapsto c + k(v) - K(\Delta(t(v))/\sigma)$ is continuous (k continuous by construction; t continuous by Lemma 1 Δ polynomial; K continuous by Lemma 3 (2), and $\Delta(t(v)) > 0$ is bounded away from 0 because $t(v) \geq t_\infty := \min\{1, b_C(2q_0 - 1)/(1 - 2b_C\delta)\} > 0$ for all $v < 1$ when $2b_C\delta < 1$, and $t \equiv 1$ otherwise), so $\mathcal{D}(c)$ is the preimage of $[0, \infty)$, closed in $[0, 1)$.

Nonempty with a neighborhood of 1: $d(v) = \Delta(t(v))/\sigma$ ranges in the compact interval $[\Delta(t_\infty), \Delta(1)]/\sigma \subset (0, \infty)$, on which K is continuous, hence bounded by some $\hat{K}$; since $k(v) \to \infty$ as $v \to 1$, there is $v_{\text{hi}} < 1$ with $c + k(v) \geq \hat{K}$ for all $v \geq v_{\text{hi}}$.

Monotone in c: immediate, c enters only the left side of the defining inequality. Hence $\underline{v}(c) = \min \mathcal{D}(c)$ is nonincreasing.

Membership of 0: $t(0) = 1$, $k(0) = 0$, so the condition at $v = 0$ reads $c \geq K(d_0) = \bar{c}$.

Full-participation form and crossing: for $v \leq \bar{v}$, $t(v) = 1$ (Lemma 1), so membership reads $c + k(v) \geq \bar{c}$. The smallest qualifying depth on this branch is $v_{\text{full}(c)} = k^{-1}((\bar{c} - c)^+)$. If $\bar{v} \geq 1$, every feasible depth lies on the full-participation branch. If instead $\bar{v} < 1$ but $k(\bar{v}) \geq \bar{c}$, then $v_{\text{full}(c)} < \bar{v}$ for every $c > 0$. In either case $\underline{v}(c) = v_{\text{full}(c)}$ throughout the verification region and never crosses $\bar{v}$. Finally suppose $\bar{v} < 1$ and $k(\bar{v}) < \bar{c}$. Then $\underline{c} = \bar{c} - k(\bar{v}) > 0$. For $c \in [\underline{c}, \bar{c}]$, $v_{\text{full}(c)} \leq \bar{v}$ and is the global minimum. For $c < \underline{c}$ no $v \leq \bar{v}$ qualifies, so $\underline{v}(c) > \bar{v}$. Thus the minimum crosses $\bar{v}$ exactly at $\underline{c}$ under the two stated conditions. Continuity and strict monotonicity on the full-participation branch follow from k^{-1}. □

B.6. Proof of Lemma 5

Proof. Step 1: if $c < \bar{c}$*, no equilibrium has* $\rho_A = 1$. Suppose $\rho_A = 1$. By Lemma 1 (1) every $s_i = 1$ citizen posts anonymously, so the anonymous channel carries cutoff $t_A = 1$ and the verified channel is empty (hence uninformative, and the receiver's posterior responds to M_A alone). The no-flood candidate for the anonymous channel then requires $\kappa_A = c \geq K(d_0)$ (Lemma 3 (1)

with $\Delta = \Delta(1)$), which fails: the heeded candidate is broken, so the channel is compromised, the unique pure-strategy outcome (§B.1. The compromised outcome, below), contradicting $\rho_A = 1$. Note anonymous participation is all-or-nothing: the channel is free, so given $\rho_A = 1$ *every* $s_i = 1$ type strictly prefers posting; no partially-populated (hence cheaper-to-defend) anonymous candidate exists.

Step 2: if $c \geq \bar{c}$, the anonymous regime is selected. Candidate: all $s_i = 1$ citizens in the anonymous channel ($t_A = 1$), verified empty, no flooding, $\rho_A = 1$, $\rho_V = 0$. Citizens are optimal by Lemma 1 (1). The anonymous channel is deterred since $c \geq \bar{c} = K(d_0)$; a joint deviation into the zero-separation verified channel cannot affect the posterior and only adds cost, by the factorization argument in §The Equilibrium Concept. Accuracy is $Q(d_0)$. The anonymous regime weakly Pareto-dominates the alternatives. When an all-verified label is exactly payoff-equivalent at $v = 0$, the channel-label convention in Timing selects anonymity.

Step 3: if $c < \bar{c}$, the outcome is the verified regime exactly on $\mathcal{D}(c)$. By Step 1, $\rho_A = 0$ in every equilibrium; anonymous voices, if any, are worthless and carry no information, so the receiver's posterior responds to the verified channel alone. Citizens face Lemma 1 (2)/(3). If $v \in \mathcal{D}(c)$, the populated-verified candidate is consistent and the verified channel is deterred by definition of $\mathcal{D}(c)$. A joint deviation into the zero-separation anonymous channel cannot affect the posterior and only adds cost, by §The Equilibrium Concept. The all-silent outcome is also consistent but Pareto-dominated for citizens, so the Pareto Selection picks the populated one. If $v \notin \mathcal{D}(c)$, the populated-verified candidate is broken, so $\rho_V = 0$, the unique pure-strategy outcome; all citizens are silent and accuracy is $Q(0)$. □

C. The Compromised Region

The main text restricts attention to pure strategies. Wherever no pure-strategy profile with positive net state-separation survives the operator's one-shot-deviation test, the region $\kappa < K(d)$ of Lemma 3, that restriction selects the compromised outcome. This section certifies what the restriction costs: it shows that a mixed-strategy equilibrium exists, bounds the public belief's accuracy in every such equilibrium, and identifies the limit in which setting that accuracy to $Q(0)$ becomes exact. It does not solve the stage-0 policy problem when the operator may mix.

Fix genuine participation at a cutoff t, so the channel carries genuine masses $m(1)$ and $m(0)$ with separation $\Delta = m(1) - m(0) > 0$ and discriminability $d = \frac{\Delta}{\sigma}$. The operator floods only in state $\omega = 0$; write the flood in noise units as $a = \frac{\tilde{m}}{\sigma} \geq 0$, so a flood of mass $\tilde{m}$ shifts the state-0 observation by a standard deviations of the organic noise. Standardize the observation by subtracting the state-0 genuine mass and dividing by the noise scale, $u = \frac{M - m(0)}{\sigma}$. The channel then reads as a unit-variance Gaussian whose mean is d when the position is true and a when it is false and the operator floods by a. Writing φ for the standard normal density, the state-1 density is $p_1(u) = \varphi(u - d)$, and when the operator plays the flood distribution F (a probability distribution over flood sizes on $[0, \infty)$) the state-0 density is the average of the shifted Gaussians,

$$p_0(u) = \int \varphi(u - a) \, \mathrm{d}F(a).$$

The public is calibrated to F: it knows the flood distribution and forms the exact posterior that a rational Bayesian public would hold, the probability that the position is true given the observation,

$$\mu(u) = \frac{p_1(u)}{p_1(u) + p_0(u)},$$

the even prior cancelling. Write $H(a) = \int \mu(u)\varphi(u - a) \, \mathrm{d}u$ for the average posterior the operator induces when it floods by a: this is the belief it manufactures at that flood size. Its expected profit from flooding by a, its stake b_O times the belief it buys, net of what the personas cost at price κ per unit mass, is

$$\Pi(a) = b_O H(a) - \kappa\sigma a.$$

A mixed-strategy equilibrium is a flood distribution the operator is content to randomize over: every flood size it uses must earn the same profit, and no unused size may earn more. Formally $\operatorname{supp}(F) \subseteq \operatorname{argmax}_a \Pi(a)$, with Π evaluated against the posterior μ that F itself calibrates. This is a fixed-point condition, because the belief the operator faces is the one its own randomization teaches the public to expect. Two accounting facts link H to the channel's accuracy. The average posterior in the true state equals H at the genuine separation, $\mathbb{E}[\mu \mid \omega = 1] = H(d)$, because

the true state reads like a flood of size d; and the average posterior in the false state is the operator's own expected induced belief, $\mathbb{E}[\mu \mid \omega = 0] = \int H(a)\,\mathrm{d}F(a)$.

C.1. Existence

A mixing equilibrium exists. The proof truncates the flood, establishes the continuity the fixed-point theorem needs, and applies that theorem to the operator's best-response correspondence; write $\Pi(a, F)$ for the profit of flood a graded against the posterior calibrated to the distribution F.

Truncation. The operator's strategy set can be truncated to floods no larger than $\bar{a} \coloneqq \frac{b_O}{\kappa\sigma}$ without loss: for any larger flood the persona spend $\kappa\sigma a$ exceeds the largest belief the operator could conceivably buy, $b_O \cdot 1$, so its profit falls below the profit of not flooding at all, $\Pi(0, F) = b_O H(0) \geq 0$, and no such flood is ever a best response. Let $\mathcal{F}$ be the set of Borel probability distributions on $[0, \bar{a}]$: convex, compact in the weak-$*$ topology (Prokhorov's theorem on a compact support set), and metrizable there, so continuity can be checked along sequences.

The calibrated posterior varies continuously with the conjecture. Let $F_n \to F$ in $\mathcal{F}$. For each observation u the map $a \mapsto \varphi(u - a)$ is bounded and continuous on $[0, \bar{a}]$, so weak-$*$ convergence tested against exactly this function gives $p_0^n(u) = \int \varphi(u - a)\,\mathrm{d}F_n(a) \to p_0(u)$ pointwise. Since $p_1(u) = \varphi(u - d) > 0$, the posterior $\mu_n(u) = \frac{p_1(u)}{p_1(u) + p_0^n(u)}$ has denominator bounded below by $p_1(u) > 0$ at each u, so $\mu_n(u) \to \mu(u)$ pointwise.

The profit is jointly continuous. Take $(a_n, F_n) \to (a, F)$ in $[0, \bar{a}] \times \mathcal{F}$. The integrand of H converges pointwise, $\mu_n(u)\,\varphi(u - a_n) \to \mu(u)\,\varphi(u - a)$, and is dominated by the integrable envelope $g(u) \coloneqq \sup_{w \in [0, \bar{a}]} \varphi(u - w)$, which is constant at $\varphi(0)$ over $[0, \bar{a}]$ and decays as a Gaussian tail outside it. Dominated convergence gives $H_n(a_n) \to H(a)$, hence $\Pi(a_n, F_n) \to \Pi(a, F)$: joint continuity on $[0, \bar{a}] \times \mathcal{F}$. The same domination yields the uniform version the closed-graph step needs: $\sup_a |H_n(a) - H(a)| \leq \int |\mu_n(u) - \mu(u)|\,g(u)\,\mathrm{d}u \to 0$ (the integrand is bounded by g and vanishes pointwise), so $\Pi(\cdot, F_n) \to \Pi(\cdot, F)$ uniformly on $[0, \bar{a}]$.

Best responses. For each F the argmax set $M(F) = \operatorname{argmax}_{a \in [0, \bar{a}]} \Pi(a, F)$ is nonempty and compact (a continuous function on a compact interval), and by Berge's theorem the value $V(F) = \max_a \Pi(a, F)$ is continuous in F. The best-response correspondence in distributions

is $\beta(F) = \{F' \in \mathcal{F} : F'(M(F)) = 1\}$, equivalently $\{F' \in \mathcal{F} : \int \Pi(a, F)\,\mathrm{d}F'(a) = V(F)\}$, since $\Pi(\cdot, F) \le V(F)$ with equality exactly on the closed set $M(F)$. It is nonempty (any point mass on $M(F)$), convex (the defining condition is linear in F'), and compact (weak-$*$ closed in $\mathcal{F}$ because $\Pi(\cdot, F)$ is continuous and bounded).

Closed graph. Let $F_n \to F$ and $F_n' \to F'$ with $F_n' \in \beta(F_n)$. Split

$$\int \Pi(a, F_n)\,\mathrm{d}F_n' = \int \Pi(a, F)\,\mathrm{d}F_n' + \int [\Pi(a, F_n) - \Pi(a, F)]\,\mathrm{d}F_n'.$$

The first term converges to $\int \Pi(a, F)\,\mathrm{d}F'$, because $\Pi(\cdot, F)$ is a bounded continuous test function for the weak-$*$ convergence $F_n' \to F'$; the second is bounded in absolute value by $\sup_a |\Pi(a, F_n) - \Pi(a, F)| \to 0$ by the uniform convergence above. Since $F_n' \in \beta(F_n)$ means the left side equals $V(F_n)$, and $V(F_n) \to V(F)$ by Berge continuity, the limit reads $\int \Pi(a, F)\,\mathrm{d}F' = V(F)$: $F' \in \beta(F)$, and the graph is closed.

Fixed point. $\mathcal{F}$ is a nonempty, convex, compact subset of the space of finite signed measures on $[0, \bar{a}]$, a locally convex space under the weak-$*$ topology, and β is a nonempty-, convex-, compact-valued correspondence with closed graph. The Kakutani-Fan-Glicksberg theorem delivers $F^* \in \beta(F^*)$: a flood distribution supported on the argmax of the profit graded against the posterior that the distribution itself calibrates. That fixed point is a mixing equilibrium. The numerical companion (script `14_mixing_equilibrium_pilot.jl`) constructs one and verifies the equal-profit condition directly, with the largest profitable deviation below 10^{-5}.

C.2. The Accuracy of the Compromised Channel

Accuracy is the negative expected posterior variance, $Q = -\mathbb{E}[\mu(1-\mu)]$, with uninformative floor $Q(0) = -\frac{1}{4}$ (Lemma 2). Because $\mu(1-\mu) \le \frac{1}{4}$ always, accuracy never drops below that floor, so $Q \ge Q(0)$ in every equilibrium. The distance above the floor takes a clean form. Since $\frac{1}{4} - \mu(1-\mu) = \left(\mu - \frac{1}{2}\right)^2$, the gap $Q - Q(0)$ is the average squared distance of the public belief from the coin-flip prior $\frac{1}{2}$; substituting $\mu = \frac{p_1}{p_1 + p_0}$ and simplifying rewrites that average as a quarter of the belief's spread between the two states,

$$Q - Q(0) = \frac{1}{4}(\mathbb{E}[\mu \mid \omega = 1] - \mathbb{E}[\mu \mid \omega = 0]) = \frac{1}{4}\left(H(d) - \int H(a)\,\mathrm{d}F(a)\right).$$

The operator's equilibrium condition bounds this spread. Write Π^* for the equilibrium profit, the common value Π takes at every flood size in the support. Averaging the equality $\Pi(a) = \Pi^*$ over F and reading off the definition of Π gives $b_O \int H(a)\,\mathrm{d}F(a) = \Pi^* + \kappa\sigma \int a\,\mathrm{d}F(a)$. The particular flood $a = d$ is available to the operator but need not be one it chooses, so it can only earn weakly less, $\Pi(d) \le \Pi^*$, that is $b_O H(d) \le \Pi^* + \kappa\sigma d$. Subtracting the averaged equality from this inequality cancels the equilibrium profit and leaves

$$b_O\left(H(d) - \int H(a)\,\mathrm{d}F(a)\right) \le \kappa\sigma\left(d - \int a\,\mathrm{d}F(a)\right) \le \kappa\sigma d,$$

the last step dropping the operator's mean flood $\int a\,\mathrm{d}F(a)$, which is nonnegative. Dividing by $4b_O$ and using $\sigma d = \Delta$, the accuracy gap obeys, in every equilibrium,

$$0 \le Q - Q(0) \le \frac{\kappa\Delta}{4b_O}.$$

This is the bound the main text quotes. Two consequences sharpen it. First, the operator's mean flood is strictly below the genuine separation. The displayed inequality initially gives $\int a\,\mathrm{d}F(a) \le d$. Equality would force $Q = Q(0)$, hence a posterior equal to $\frac{1}{2}$ almost surely and identical state densities $p_1 = p_0$. Identification of Gaussian location mixtures then requires $F = \delta_d$: dividing the two characteristic functions by the nonzero Gaussian characteristic function leaves the characteristic function of a point mass at d. But against the constant posterior generated by δ_d, flooding at $a = 0$ buys the same belief as $a = d$ at strictly lower cost when $\kappa > 0$, so δ_d cannot be a best response to itself. Therefore $\int a\,\mathrm{d}F(a) < d$, net state-separation is strictly positive, and the channel is heeded. The fixed citizen cutoff is consequently optimal under its conjectured standing, closing the full equilibrium. Second, when the genuine separation d lies in the support of the equilibrium flood, so that flooding by d is itself one of the operator's optimal randomizations, the inequality $\Pi(d) \le \Pi^*$ holds with equality and the bound becomes an identity,

$$Q - Q(0) = \frac{\kappa\Delta}{4b_O}\left(1 - \frac{1}{d}\int a\,\mathrm{d}F(a)\right),$$

valid whenever $d \in \operatorname{supp}(F)$. The numerical companion confirms this premise across the interior of the compromised region and shows it failing only at the extreme upper edge $\kappa \to K(d)^-$,

where the equilibrium flood collapses to an atom at zero (the operator almost never floods) and d drops out of the support; there the identity slightly overstates the gap and only the inequality binds.

The bound vanishes as the persona price falls, $Q - Q(0) \leq \kappa \frac{\Delta}{4b_O} \to 0$ as $\kappa \to 0$. Equivalently the compromised channel's accuracy converges to the uninformative floor,

$$\lim_{\kappa \to 0} Q = Q(0) = -\frac{1}{4},$$

so setting the compromised channel's accuracy to $Q(0)$ is asymptotically exact in the cheap-fabrication limit. Away from that limit the bound is only an upper envelope. Near the boundary $\kappa \to K(d)^-$ it can be as large as $K(d)\frac{\Delta}{4b_O}$, and the Lemma 3 ceiling $K(d) \leq b_O \frac{d}{4\sigma}$ shows this can reach $\frac{d^2}{16}$, comparable to the whole distance from the floor to full revelation; there the compromised channel can keep substantial accuracy. This does not make the main-text policy comparison conservative in a signed way. A heeded mixed-strategy flooding equilibrium also retains private posting surplus, and the bound neither selects among mixed equilibria nor establishes a monotone welfare crossing. It therefore cannot locate when verification begins once the operator may mix. Proposition 3 confines its policy claims to the pure-strategy equilibrium selection.

D. The Pareto Selection

The main text selects among self-confirming outcomes by unanimity and asserts that the selection is well-defined and conservative. This section records both claims.

The ranking is unanimous, so the selection is well-defined. Every citizen ranks the candidate outcomes the same way, anonymous over verified over dead: the private correctness payoff is realized only in a heeded venue, and anonymity avoids exposure. The ranking is unchanged if citizens also value the accuracy of the public belief, which falls in the same order across the candidates. At $v = 0$, the channel-label convention resolves the payoff-equivalent anonymous and verified labels. Its behavioral footing is that forums are voluntary venues, and audiences discount venues known to be cheaply floodable (Mayzlin 2006).

The selection is conservative in what it removes. It discards only forums that are dead from expectations alone. When the operator's flooding is what kills a forum, that death stands, so the welfare losses in the analysis, the dead forum and the thinned verified forum, are floors produced by the swarm's economics, not artifacts of how the equilibrium was chosen. Within the compromised region, the tie-break toward silence is applied in fixing the compromised outcome before the Pareto Selection compares outcomes, foreclosing the payoff-equivalent posting patterns that indifferent citizens could otherwise adopt (§B.1. The compromised outcome).

E. The Benchmark and the Degradation

E.1. Proof of Proposition 1

Proof. Immediate from Lemma 5 (1): at $c \geq \bar{c}$ the anonymous regime obtains at every v, so every $s_i = 1$ citizen posts anonymously with no exposure paid, both channels are deterred, and accuracy is $Q(d_0)$. For the maximality claim: in any regime the informative channel carries a cutoff $t \leq 1$, and accuracy is $Q(\Delta(t)/\sigma)$ with Δ strictly increasing in t (its derivative is $2q(t) - 1 > 0$) and Q strictly increasing (Lemma 2), so $Q(d_0)$ at $t = 1$ is the maximum attainable in any regime; the verified regime attains at most $t(v) \leq 1$ and the dead forum gives $Q(0) < Q(d_0)$. Since no equilibrium object in the anonymous regime depends on v (Lemma 5 (1)), the outcome, and in particular this accuracy value, is constant in v. □

E.2. Proof of Proposition 2

Proof. (1) On $\mathcal{D}(c)$ the verified channel's operator is deterred (Lemma 3 (3), via Lemma 5 (2)), so the heard mass is $m(\omega) + \sigma\varepsilon_V$ with no synthetic component: $\pi(v) = 1$.

(2) By Lemma 1 (2), on $(\bar{v}, 1)$ the cutoff $t(v)$ is interior and strictly decreasing, and since q is increasing, the types leaving as v rises, those at the cutoff, have the highest precision $q(t(v))$ among posters. For $\delta > 0$: $\bar{q}(v) = q_0 + \delta t(v)/2$ is strictly increasing in t, hence strictly decreasing in v; likewise $\Delta(t)/t = (2q_0 - 1) + \delta t$; and $Q(d(v))$ falls because $\Delta(t) = (2q_0 - 1)t + \delta t^2$ is strictly increasing in t (its derivative $2q(t) - 1 > 0$), t falls, and Q is strictly increasing in d (Lemma 2).

(3) With $\theta \sim U[0,1]$ and q affine: $\mathbb{E}[q \mid \theta \le t] = q_0 + \delta t/2$ and $\mathbb{E}[q \mid \theta > t] = q_0 + \delta(1+t)/2$; subtracting gives the constant gap $\delta/2$, independent of t and hence of v. A silenced type $\theta > t(v)$ has $b_C(2q(\theta) - 1) > 0$, so with $\rho_A = 1$ she would strictly prefer anonymous posting to silence; but $\rho_A = 0$ in every equilibrium at $c < \bar{c}$ (Lemma 5, Step 1), and on the equilibrium path the operator spends nothing, deterrence in the verified channel and the anonymous channel's death are both off-path threats.

(4) Set $\delta = 0$: $\bar{q} \equiv q_0$, $\Delta(t)/t \equiv 2q_0 - 1$, and the gap in (3) is 0; the cutoff still falls, but composition is precision-neutral, and Q falls only through lost volume. For $\delta < 0$, the survivor-composition comparisons reverse: $\bar{q}$ and $\Delta(t)/t$ rise as v rises, and the precision gap is $\frac{\delta}{2} < 0$. Total state-separation and accuracy nevertheless continue to fall through lost volume, since $\Delta'(t) = 2q(t) - 1 > 0$. □

E.3. Continuous Benefit-Side Standing for Proposition 2

This remark records what becomes of Proposition 2 when the poster's benefit scales continuously with how informative the forum is, rather than switching between zero and full at the standing threshold. It is a robustness check on the citizen side only: the benefit multiplier $r(d) \in [0,1]$ is a posited reduced form, increasing and differentiable in the discriminability d, and the receiver is held at the paper's binary standing. A type-θ citizen with $s_i = 1$ now posts in the deterred verified channel iff $b_C(2q(\theta) - 1)r(d) - \theta v \ge 0$, and because r moves with $d = \frac{\Delta(t)}{\sigma}$, which moves with the cutoff t, the cutoff is a fixed point rather than a closed form:

$$t = T(t) := \frac{b_C(2q_0 - 1)\, r(d(t))}{v - \hat{v}\, r(d(t))}, \quad d(t) = \frac{\Delta(t)}{\sigma}.$$

Write $F(t, v) := (v - \hat{v}\, r(d(t)))\, t - b_C(2q_0 - 1)r(d(t))$, so a live cutoff solves $F = 0$. Factoring gives the exact identity $F(t, v) = (v - \hat{v}r)(t - T(t))$, hence at a fixed point $\partial \frac{F}{\partial} t = (v - \hat{v}r)(1 - T'(t))$. A fixed point is stable, meaning small perturbations of who posts die out, exactly when $T'(t) < 1$, which is exactly when $\partial \frac{F}{\partial} t > 0$. Differentiating $F = 0$ then gives, for every increasing r,

$$\frac{\mathrm{d}t^*}{\mathrm{d}v} = -\frac{t}{\frac{\partial F}{\partial t}} < 0 :$$

at any stable participation equilibrium the cutoff still falls in v, so the mean heard precision $\bar{q}(v)$, the per-voice weight $\frac{\Delta(t)}{t}$, and the accuracy $Q(d(t))$ all still decline past the participation boundary, exactly as in Proposition 2. The feedback moreover steepens the decline: since $\partial \frac{F}{\partial} t = (v - \hat{v}r) - r'(d)\, d'(t)[\hat{v}t + b_C(2q_0 - 1)]$, with $d'(t) = \Delta' \frac{t}{\sigma} > 0$, is smaller than its flat-standing value $v - \hat{v}r$ whenever $r' > 0$, the magnitude $|\, \mathrm{d} \frac{t^\star}{\mathrm{d}} v|$ is larger than under binary standing. The binary benefit is thus the conservative choice for the degradation claim.

The derivative above is local to a regular stable live root. Continuous standing can change global existence: $r(0) = 0$ supports a dead fixed point, and nonlinear increasing r can generate multiple live roots or saddle-node disappearance. Increasing and differentiable standing alone therefore does not guarantee a unique live cutoff or a positive participation floor. Whenever the selected live root is regular and locally stable, the signed comparative static above applies; no broader existence or uniqueness claim is made.

E.4. Direct Concern for Accuracy in Citizen Utility

The common public-accuracy payoff is distinct from private correctness. Under an action profile a that generates discriminability $d(a)$, its expectation is $\gamma Q(d(a))$. A unilateral deviation by an atomless citizen leaves $d(a)$ unchanged. The accuracy term therefore cancels from every posting-versus-silence comparison in Lemma 1, leaving best responses and the equilibria in Lemma 5 unchanged. With population mass one, aggregating private payoffs and adding the common term yields welfare $W = S + \gamma Q$.

E.4.1. Finite Source Pools and Pivotality

Fix a deterred verified channel and replace the continuum by N citizens, each of whose voice adds $\frac{1}{N}$ to the observed mass. Let M_{-i} denote the mass generated by every source except citizen i, including organic noise, and let μ_N be the equilibrium-calibrated posterior. Suppose citizens directly value public accuracy through the realized payoff $-\lambda(\omega - \mu_N)^2$, with $\lambda \geq 0$. For a citizen with $s_i = 1$, define the pivotal accuracy benefit of posting rather than remaining silent as

$$P_i^N(a_{-i}) := \mathbb{E}\left[(\omega - \mu_N(M_{-i}))^2 - \left(\omega - \mu_N\left(M_{-i} + \frac{1}{N}\right)\right)^2 \;\middle|\; s_i = 1, \theta_i\right]$$

At any fixed strategy profile of the other citizens, her exact posting condition is

$$b_C(2q(\theta_i) - 1) + \lambda P_i^N(a_{-i}) \geq \theta_i v$$

Thus scarcity changes incentives rather than mechanically overturning them. For $\theta_i > 0$, define the profile-specific participation threshold

$$v_i^N(a_{-i}) := \frac{b_C(2q(\theta_i) - 1) + \lambda P_i^N(a_{-i})}{\theta_i}$$

Citizen i posts at the candidate profile exactly when $v \leq v_i^N(a_{-i})$. Along any specified equilibrium path $a_{-i}(v)$, write $v_i^N(v) := v_i^N(a_{-i}(v))$. The exact scope condition for her to remain active at every feasible stringency is

$$v \leq v_i^N(v) \quad \text{for every } v \in [0, 1)$$

If the pivotal benefit is invariant along the path, this condition reduces to $v_i^N \geq 1$. A uniquely informative source can have a large enough P_i^N to satisfy the condition, but being unique does not guarantee it. When the inequality fails, she exits at the first stringency at which the policy exceeds her threshold.

The pivotal benefit is bounded. Because squared error lies in $[0, 1]$, $\left|P_i^N(a_{-i})\right| \leq 1$, so $\theta_i v > b_C(2q(\theta_i) - 1) + \lambda$ is sufficient for silence even under the largest possible accuracy gain. Conversely, for fixed $\sigma > 0$ and a regular posterior, Gaussian smoothing makes μ_N differentiable, while one action shifts the observation by only $\frac{1}{N}$; the mean-value theorem then gives $P_i^N(a_{-i}) = O\left(\frac{1}{N}\right)$. The atomless benchmark is the limit in which this pivotal term vanishes.

The finite-population selection condition is also transparent. At any candidate profile, order citizens by exposure and define their thresholds $v_i^N(a_{-i})$ as above. The best-informed citizens exit first when these thresholds fall with exposure over the relevant upper tail. A small informed pool can raise their $P_i^N(a_{-i})$ and flatten or reverse that ordering. The continuum result therefore describes large count-based forums, while the finite extension identifies exactly when a pivotal source is protected by the private value of her informational contribution.

This extension is participation-side. It fixes a deterred verified channel and characterizes citizen best responses along a specified finite-population path. When a small informed pool makes pivotality strong enough to keep the best-informed citizens posting or reverse their exit ordering, the composition result fails. The extension does not re-solve the operator's flooding problem or the stage-0 policy choice with finitely many citizens.

For a population-free sufficient condition, let $\eta \geq 0$ and let $h : [0,1] \to \mathbb{R}_+$ be continuously differentiable and increasing. This reduced form includes a bounded reward for one's own contribution without solving the finite game at every strategy profile. Replace (2) by

$$u_i^\eta(\ell) = u_i(\ell) + \eta(2s_i - 1)h(\theta_i)\rho_\ell$$

The added reward is largest for an informed favorable-signal poster and is negative for a contribution she expects to be incorrect. Thus citizens with $s_i = 0$ remain silent. Write $w(\theta) := 2q(\theta) - 1$ and $\bar{h}' := \max_{\theta \in [0,1]} h'(\theta)$. A citizen with a favorable signal gains

$$g_\eta(\theta, v) = b_C w(\theta) + \eta h(\theta) - \theta v$$

from posting in the heeded verified channel. Since $w'(\theta) = 2\delta$ and $\hat{v} = 2b_C\delta$,

$$\partial_\theta g_\eta(\theta, v) = \hat{v} + \eta h'(\theta) - v \leq \hat{v} + \eta\bar{h}' - v$$

The gain is therefore strictly decreasing in type whenever

$$v > \hat{v} + \eta\bar{h}'$$

Moreover, $g_\eta(0, v) > 0$, while $g_\eta(1, v) < 0$ whenever

$$v > \bar{v} + \eta h(1)$$

If a feasible stringency satisfies both inequalities, there is a unique interior cutoff: citizens below it post and the best-informed citizens exit first. The two conditions identify the opposing forces. The slope condition preserves low-type single crossing; the endpoint condition ensures that exposure eventually exceeds even the strongest information reward.

When the private reward is proportional to a poster's contribution to state separation, $h(\theta) = w(\theta)$, the extension is especially transparent. It replaces b_C by $b_C + \eta$, so the incentive-balance and silencing thresholds become $2(b_C + \eta)\delta$ and $(b_C + \eta)w(1)$. A larger reward delays

selective exit but does not change its ordering. If no feasible stringency exceeds these thresholds, informed posters remain; the adverse-selection result does not apply. Thus the mechanism survives a bounded private return to informational contribution, but not an arbitrarily strong subsidy for producing public accuracy.

E.5. Stage 0 as Collective Choice

This subsection establishes the democratic interpretation of the welfare optimum: it is each citizen's own criterion, first ex ante and then under interim knowledge of her own exposure type. The result identifies the verification level society implements through collective choice, not merely a planner's recommendation. Throughout, γ is the common direct concern for public accuracy defined in the main text and analyzed in the preceding subsection (§E.4. Direct Concern for Accuracy in Citizen Utility): its realized contribution is $-\gamma(\omega-\mu)^2$, and its expectation is $\gamma Q(d(v))$.

Proof. Ex ante. Types are drawn after stage 0, so before the draw every citizen faces the same lottery: her exposure type is a fresh draw $\theta \sim U[0,1]$, her signal matches the state with probability $q(\theta)$, and she then plays the equilibrium selected at stringency v. Her ex-ante expected utility separates into the private posting payoff and the common accuracy term,

$$\mathbb{E}[u_i] = \int_0^1 \mathbb{E}[u_i^* \mid \theta]\,\mathrm{d}\theta + \gamma Q(d(v)) = S(v) + \gamma Q(v) = W(v),$$

where the first integral is the cross-sectional posting surplus $S(v)$ of (6), because a single citizen's type is distributed exactly as the population cross-section (i.i.d. $U[0,1]$). This value is identical for every citizen, so all citizens share the objective W and therefore the same ideal stringency $v^*(c)$ (Proposition 3). Any anonymous voting rule over stage-0 policies selects $v^*(c)$ unanimously.

Interim. Now let each citizen know her own exposure type θ before the choice but not yet her signal. Write $U_{\theta(v)}$ for her interim expected utility, averaging over her signal, the state, and the other citizens. Across the three regimes of Lemma 5 her interim utility is

$$U_\theta^{\text{anon}} = \frac{1}{2} b_C (2q(\theta) - 1) + \gamma Q(d_0), \quad U_\theta^{\text{dead}} = \gamma Q(0),$$

$$U_\theta^{\text{ver}}(v) = \frac{1}{2}[b_C(2q(\theta) - 1) - \theta v]\mathbb{1}\{\theta \le t(v)\} + \gamma Q(d(v)),$$

since in the anonymous regime every favorable signal posts with no exposure, in the deterred verified regime a favorable signal below the cutoff posts and pays θv while every other type is silent, and the dead forum silences all and leaves the prior accuracy $Q(0)$.

On the verified branch, differentiate in v. For a posting type ($\theta \le t(v)$) on the interior branch, the exposure term and the common accuracy term both fall,

$$\frac{\mathrm{d}}{\mathrm{d}v}U_\theta^{\text{ver}}(v) = -\frac{\theta}{2} + \gamma Q'(d(v))d'(v) < 0,$$

negative because past the silencing threshold $d'(v) < 0$ while $Q' > 0$ (Lemma 1 (2) and Lemma 2), and below it the accuracy term is flat while the private term still falls at rate $\theta/2$. For a silenced type the private term is absent and $U_\theta^{\text{ver}} = \gamma Q(d(v))$ is nonincreasing, strictly so past silencing whenever $\gamma > 0$. Hence U_θ^{ver} is nonincreasing in v for every θ, strict except for a $\gamma = 0$ silenced type, so every type's ideal on the deterred branch is its minimum element, the minimum deterring stringency $\underline{v}(c)$ (Lemma 4).

It remains to rank $\underline{v}(c)$ against shallower policies. A stringency below $\underline{v}(c)$ lies outside the deterrence set, so the selected pure-strategy equilibrium there is the dead forum (Lemma 5). For a poster the gap is

$$U_\theta^{\text{ver}}(\underline{v}(c)) - U_\theta^{\text{dead}} = \frac{1}{2}[b_C(2q(\theta) - 1) - \theta\underline{v}(c)] + \gamma[Q(d(\underline{v}(c))) - Q(0)] \ge 0,$$

with both terms nonnegative: the bracket is nonnegative for $\theta \le t(\underline{v}(c))$, vanishing only at the cutoff, and $Q(d(\underline{v}(c))) > Q(0)$ because the deterred discriminability is positive (Lemma 2). For a silenced type only the accuracy term survives, again nonnegative. The gap is strict for every poster and strict for every type whenever $\gamma > 0$.

Therefore, under the pure-strategy equilibrium selected at each stringency, every type weakly prefers $\underline{v}(c)$ to any shallower policy, which yields the dead forum, and to any deeper policy, since U_θ^{ver} is nonincreasing in v. So $v^*(c) = \underline{v}(c)$ is a Condorcet winner: it is a weak majority, indeed unanimous, winner against every alternative. In the privacy region ($c \ge \bar{c}$)

the anonymous regime obtains at every stringency, so U_θ is independent of v, all types are indifferent, and the lower-verification tie-break selects $v = 0$.

Disagreement requires an equilibrium in which some shallower policy does not produce the dead forum. The only such case is $v = 0$ under a mixed-strategy flooding equilibrium in which the compromised anonymous forum retains value above $\gamma Q(0)$; its accuracy bound (§C.2. The Accuracy of the Compromised Channel) shows only that the accuracy gain vanishes as the persona price falls. Private posting surplus need not vanish, so Proposition 3 leaves that adoption margin open. □

F. The Optimal Verification Stringency

F.1. Proof of Proposition 3

Proof. By Lemma 5 the outcome at (v, c) is the anonymous regime ($c \geq \bar{c}$, any v), the verified regime ($c < \bar{c}$, $v \in \mathcal{D}(c)$), or the dead forum ($c < \bar{c}$, $v \notin \mathcal{D}(c)$). Write the corresponding welfare values

$$W_{\text{anon}} = \frac{b_C}{2}\Delta(1) + \gamma Q(d_0), \quad W_{\text{ver}}(v) = S(v) + \gamma Q(d(v)), \quad W_{\text{dead}} = \gamma Q(0),$$

where, in the verified regime (each citizen posts with ex-ante probability $1/2$, exposure paid only by posters),

$$S(v) = \frac{1}{2}\int_0^{t(v)} [b_C(2q(\theta) - 1) - \theta v]\,\mathrm{d}\theta = \frac{1}{2}\left[b_C\Delta(t(v)) - v\frac{t(v)^2}{2}\right],$$

and the dead forum has zero posting surplus (all silent, Lemma 5 (3)).

Step 1: case $c \geq \bar{c}$. By Lemma 5 (1) the outcome is the anonymous regime at every v: every $s_i = 1$ citizen posts with no exposure paid and accuracy is $Q(d_0)$ (Proposition 1). Since each citizen posts with ex-ante probability $1/2$ (equal prior and symmetric signal give $\Pr(s_i = 1) = 1/2$ for every type) and each poster earns $b_C(2q(\theta) - 1)$ with no exposure cost, the posting surplus is $S_{\text{anon}} = 1/2\int_0^1 b_C(2q(\theta) - 1)\,\mathrm{d}\theta = b_C/2\Delta(1)$, giving $W(v) \equiv W_{\text{anon}}$, constant since no equilibrium object varies with v in the anonymous regime. v is payoff-irrelevant; ties are broken toward lower verification, selecting $v^* = 0$. This proves (1).

Step 2: W_{ver} is strictly decreasing on $[0,1)$. For $v < \bar{v}$: $t \equiv 1$, so $\mathrm{d}S/\,\mathrm{d}v = -1/4$ and the accuracy term is constant; hence $\mathrm{d}W_{\text{ver}}/\,\mathrm{d}v = -1/4 < 0$. For $v > \bar{v}$: the cutoff type is indifferent, $b_C(2q(t)-1) = tv$, so the envelope terms cancel:

$$\frac{\mathrm{d}S}{\mathrm{d}v} = \frac{1}{2}\left[b_C(2q(t)-1)t' - \frac{t^2}{2} - vtt'\right] = \frac{1}{2}\left[t'(b_C(2q(t)-1) - vt) - \frac{t^2}{2}\right] = -\frac{t^2}{4} < 0,$$

and the accuracy term falls as well: $t' < 0$ (Lemma 1 (2)), $\Delta'(t) = 2q(t) - 1 > 0$, and Q strictly increasing (Lemma 2). W_{ver} is continuous (all components are), so strictly decreasing.

Step 3: case $c < \bar{c}$. For $v \in \mathcal{D}(c)$ the outcome is worth $W_{\text{ver}}(v)$; by Step 2 the best such verification stringency is the minimum element $\underline{v}(c)$ of the closed set $\mathcal{D}(c)$ (Lemma 4), and $\underline{v}(c) > 0$ since $0 \notin \mathcal{D}(c)$. For $v \notin \mathcal{D}(c)$ the outcome is worth W_{dead}. The verified regime strictly dominates:

$$W_{\text{ver}}(\underline{v}(c)) - W_{\text{dead}} = S(\underline{v}(c)) + \gamma[Q(d(t(\underline{v}(c)))) - Q(0)] > 0,$$

because both terms are strictly positive, $S(v) > 0$ for every $v < 1$ (the integrand $b_C(2q(\theta) - 1) - \theta v$ is strictly positive on a neighborhood of $\theta = 0$ and nonnegative up to the cutoff, and $t(v) > 0$ always since the intercept $b_C(2q_0 - 1) > 0$), and $Q(d) > Q(0)$ for $d > 0$ (Lemma 2). So $v^*(c) = \underline{v}(c)$, uniquely by Step 2's strict monotonicity. This proves (2).

Step 4: monotone deepening and the crossing. At $\underline{v}(c) > 0$ the deterrence constraint binds; otherwise continuity would permit a slightly smaller deterring depth. If $0 < c' < c < \bar{c}$, every $v < \underline{v}(c)$ already fails at c and hence at c', while $\underline{v}(c)$ itself fails at c' because its price falls by $c - c'$. Thus $\underline{v}(c') > \underline{v}(c)$: the selected depth strictly rises as fabrication cheapens. Lemma 4 gives the rest. If either crossing condition fails, the selected depth remains below $\bar{v}$ for every $c > 0$. If both hold, $\underline{c} = \bar{c} - k(\bar{v}) > 0$, the full-participation closed form reaches $\bar{v}$ there, and the minimum exceeds $\bar{v}$ below it. This proves (3).

Step 5: strictness and continuity at the threshold. Strictness below $\bar{c}$ is Step 3; above, all verification stringencies tie at W_{anon} and the tie-break is stated in the proposition. Continuity of the value: as $c \uparrow \bar{c}$, $v^*(c) \le k^{-1}(\bar{c} - c) \to 0$ and W_{ver} is continuous with $W_{\text{ver}}(0) = \frac{b_C}{2}\Delta(1) + \gamma Q(d_0) = W_{\text{anon}}$, so $W(v^*(c)) \to W_{\text{anon}}$: the same value the privacy region delivers. The regime,

who posts, in which channel, at what price, changes at $\bar{c}$; the welfare level does not jump. This proves (4). □

F.2. Proof of Proposition 4

Proof. Strict decrease of $v^* = \underline{v}(c)$. Let $0 < c' < c < \bar{c}$. At $\underline{v}(c) > 0$ the deterrence constraint binds with equality: if $c + k(\underline{v}(c)) > K(d(t(\underline{v}(c))))$ strictly, then by continuity of all components (Lemma 4′s closedness argument) a slightly smaller verification stringency would also lie in $\mathcal{D}(c)$, contradicting minimality. Hence $c' + k(\underline{v}(c)) < c + k(\underline{v}(c)) = K(d(t(\underline{v}(c))))$, so $\underline{v}(c) \notin \mathcal{D}(c')$ and $\underline{v}(c') > \underline{v}(c)$.

Composition. For $c < \underline{c}$, $\underline{v}(c) > \bar{v}$ (Lemma 4), so $c' < c < \underline{c}$ gives $\bar{v} < v^*(c) < v^*(c')$, and Lemma 1 yields $t(v^*(c')) < t(v^*(c))$ strictly. Then $\bar{q} = q_0 + \delta t/2$ falls strictly (for $\delta > 0$), and $Q(d)$ falls strictly since $\Delta(t)$ is strictly increasing in t and Q strictly increasing in d (Lemma 2). $\pi = 1$ on the deterred branch is Proposition 2 (1).

Operator payoff. On the equilibrium path the operator spends nothing in any regime: the verified channel is deterred on $\mathcal{D}(c)$ (Lemma 3 (3)) and a channel with zero conjectured separation attracts no flood (Lemma 3 (4)), so by (4) its realized payoff is $b_O \mathbb{E}[\mu \mid \omega = 0]$. With prior $1/2$, iterated expectations give $\mathbb{E}[\mu] = \mathbb{E}[\,\mathbb{E}[\omega \mid M]\,] = \Pr(\omega = 1) = 1/2$ and $\mathbb{E}[\mu\,\mathbb{1}\{\omega = 1\}] = \mathbb{E}[\,\mu\,\mathbb{E}[\mathbb{1}\{\omega = 1\} \mid M]\,] = \mathbb{E}[\mu^2]$, hence

$$\mathbb{E}[\mu \mid \omega = 0] = \frac{\mathbb{E}[\mu] - \mathbb{E}[\mu^2]}{\Pr(\omega = 0)} = 2\,\mathbb{E}[\mu(1-\mu)] = -2Q(d),$$

where d is the discriminability of the informative channel ($d = 0$ in the dead forum). The identity holds in every regime, so the anonymous benchmark pays the operator $-2b_O Q(d_0)$ and the optimally-defended verified regime pays $-2b_O Q(d(v^*(c)))$. By the composition step, $Q(d(v^*(c'))) < Q(d(v^*(c))) < Q(d_0)$ for $c' < c < \underline{c}$ (the last inequality because $t(v^*(c)) < 1$ gives $d(v^*(c)) < d_0$, with Lemma 2); multiplying by $-2b_O < 0$ reverses all three: the operator's payoff strictly rises as c falls and strictly exceeds its anonymous-benchmark value. □

G. Defenses Against Fabricated Participation

G.1. Proof of Proposition 5

Proof. The proof is direct, in five steps, under $\delta \geq 0$. Throughout, the verified channel charges the pair (τ, v): a poster of type θ pays $\tau + \theta v$, each synthetic persona costs $c + \tau + k(v)$, and the operator's problem is Lemma 3′s at persona price $\kappa = c + \tau + k(v)$. The anonymous channel is unchanged, and $c < \bar{c}$, so Step 1 of Lemma 5′s proof applies: the anonymous channel is compromised in every pure-strategy equilibrium, and the outcome is the verified regime where the verified channel is deterred and the dead forum elsewhere.

Step 1: participation and the cap. A citizen with $s_i = 1$ posts in the heeded verified channel iff

$$g(\theta) := b_C(2q(\theta) - 1) - \tau - \theta v = (b_C(2q_0 - 1) - \tau) + (\hat{v} - v)\,\theta \geq 0,$$

an affine function of θ by (2) and (1). The least-exposed types post iff $g(0) > 0$, i.e. $\tau < b_C(2q_0 - 1)$; for $\tau \geq b_C(2q_0 - 1)$ and any $v \geq \hat{v}$, g is nonpositive at $\theta = 0$ and nonincreasing, so no configuration of the form $[0, t]$ is populated, which is the cap. Assume $\tau < b_C(2q_0 - 1)$ henceforth. Since $b_C(2q_0 - 1) + \hat{v} = b_C(2q(1) - 1)$, we have $g(1) \geq 0$ iff $v \leq \bar{v} - \tau$: participation is full ($t = 1$) for $v \leq \bar{v} - \tau$, and for $v > \bar{v} - \tau$ (necessarily $v > \hat{v}$) the cutoff solves $g(t) = 0$,

$$t(\tau, v) = \frac{b_C(2q_0 - 1) - \tau}{v - \hat{v}} \in (0, 1),$$

strictly decreasing in both arguments, with $t(0, v)$ Lemma 1′s cutoff.

Step 2: the optimum is the minimum that deters. Define $\mathcal{D}(c, \tau) = \left\{v : c + \tau + k(v) \geq K\left(\frac{\Delta(t(\tau,v))}{\sigma}\right)\right\}$. Closedness follows from continuity. For nonemptiness, if $\hat{v} \geq 1$, then $\bar{v} - \tau = \hat{v} + b_C(2q_0 - 1) - \tau > 1$, so participation is full at every feasible v. If $\hat{v} < 1$, the interior cutoff is bounded away from zero as $v \to 1$ by $\frac{b_C(2q_0-1)-\tau}{1-\hat{v}} > 0$. In either case K is bounded on the attainable range while $k(v) \to \infty$, so all sufficiently deep verification stringencies deter. In the verified regime the posting surplus is $S(\tau, v) = \frac{1}{2}\int_0^t (b_C(2q(\theta) - 1) - \tau - \theta v)\,\mathrm{d}\theta$. On the full-participation branch, $\mathrm{d}\frac{S}{\mathrm{d}}v = -\frac{1}{4}$ and accuracy is constant at $Q(d_0)$; past it the cutoff type is indifferent, so $\mathrm{d}\frac{S}{\mathrm{d}}v = -\frac{t^2}{4} < 0$ and accuracy falls. Welfare in the verified regime is therefore

strictly decreasing in v, and the verified regime strictly beats the dead forum. Hence $v^*(\tau, c) = \min \mathcal{D}(c, \tau) =: \underline{v}(\tau, c)$ under the pure-strategy equilibrium selected at each stringency.

Step 3: strict decrease in the per-post cost, and the closed form. On the full-participation branch the deterrence condition is $c + \tau + k(v) \geq \bar{c}$. Since $k(0) = 0$ and $v \geq 0$, its minimum is

$$\underline{v}(\tau, c) = k^{-1}((\bar{c} - c - \tau)^+).$$

Write $\tau_0(c) = \bar{c} - c$. For $\tau < \tau_0(c)$ the optimum is positive and the constraint binds, so the implicit function theorem gives $\mathrm{d}\underline{v}\frac{\tau, c}{\mathrm{d}}\tau = -\frac{1}{k'}(\underline{v}(\tau, c)) < 0$. At $\tau_0(c)$ it reaches zero and remains zero thereafter. The cost-only corner lies in the proposition's populated, full-participation domain whenever $\tau_0(c) < b_C(2q_0 - 1)$; the proposition does not extend the low-exposure cutoff analysis beyond $\tau < b_C(2q_0 - 1)$.

For part (2), assume directly that $K'(d) \geq 0$ at every on- and off-path discriminability used below. Equivalently, $\kappa_t \geq 0$ in Step 5′s notation throughout the comparison. Write the residual $r(v; \tau) = c + \tau + k(v) - K\left(\frac{\Delta(t(\tau, v))}{\sigma}\right)$. Under the range condition it is nondecreasing in v and strictly increasing in τ: the per-post cost raises the price directly, and its reduction of the cutoff weakly lowers K. By the definition of the minimum, $r(v; \tau) < 0$ for every $v < \underline{v}(\tau, c)$. Let $\tau' > \tau$ and suppose $\underline{v}(\tau, c) > 0$. For $v < \underline{v}(\tau, c)$ close enough to it, continuity and binding at the minimum give $r(v; \tau) > -(\tau' - \tau)$, hence $r(v; \tau') > 0$. Such v lies in $\mathcal{D}(c, \tau')$, so $\underline{v}(\tau', c) < \underline{v}(\tau, c)$ strictly.

Step 4: the welfare margin. On the full-participation branch, $S(\tau, v) = \left(\frac{b_C}{2}\right)\Delta(1) - \frac{\tau}{2} - \frac{v}{4}$ and accuracy is $Q(d_0)$, so

$$W(\tau) = \frac{b_C}{2}\Delta(1) - \frac{\tau}{2} - \frac{\underline{v}(\tau, c)}{4} + \gamma Q(d_0).$$

For $\tau < \tau_0(c)$, $\mathrm{d}\frac{W}{\mathrm{d}}\tau = -\frac{1}{2} + \frac{1}{4k'(\underline{v}(\tau, c))}$, which is positive iff $k'(\underline{v}(\tau, c)) < \frac{1}{2}$. For $\tau > \tau_0(c)$, $\underline{v}(\tau, c) = 0$ and $\mathrm{d}\frac{W}{\mathrm{d}}\tau = -\frac{1}{2}$. Welfare has the corresponding kink at $\tau_0(c)$; only its left derivative obeys the $k' < \frac{1}{2}$ test. For the leading form $k(v) = \bar{k}\frac{v}{1-v}$, the positive-stringency test is $v < 1 - \sqrt{2\bar{k}}$, a nonvacuous region iff $\bar{k} < \frac{1}{2}$.

Step 5: the interior branch and the accuracy weight. Past full participation the cutoff is interior, $t(\tau, v) = \frac{a}{v - \hat{v}}$ with $a := b_C(2q_0 - 1) - \tau$, and the binding deterrence constraint $c + \tau +$

$k(v) = K\left(\frac{\Delta(t(\tau,v))}{\sigma}\right)$ pins the optimal $v^*(\tau, c)$ implicitly, since the deterrence bound now moves with the cutoff. Write κ_t for the rate at which that bound rises with the cutoff, $\kappa_t \coloneqq K'(d)\,\Delta'\frac{t}{\sigma}$, evaluated at $d = \frac{\Delta(t)}{\sigma}$ and with $\Delta'(t) = 2q(t) - 1 > 0$; K' exists by Lemma 3 (5), and part (2)'s tameness hypothesis gives $\kappa_t \geq 0$ at every attainable configuration (Step 3). Differentiating the constraint through both $k(v)$ and the cutoff, and using the cutoff partials $t_\tau = -\frac{t}{a}$ and $t_v = -\frac{t^2}{a}$,

$$\frac{\mathrm{d}v^*}{\mathrm{d}\tau} = -\frac{a + \kappa_t t}{a\,k'(v^*) + \kappa_t t^2} < 0,$$

where $\kappa_t \geq 0$ signs both pieces, the numerator at least a and the denominator at least $a\,k'(v^*) > 0$: a per-post cost still lowers the required verification stringency. The total change in the cutoff along the path, $\mathrm{d}\frac{t}{\mathrm{d}}\tau = t_\tau + t_v\left(\mathrm{d}\frac{v^*}{\mathrm{d}}\tau\right)$, simplifies after the κ_t terms cancel to

$$\frac{\mathrm{d}t}{\mathrm{d}\tau} = \frac{t(t - k'(v^*))}{a\,D_v}, \quad D_v \coloneqq k'(v^*) + \kappa_t\frac{t^2}{a} \geq k'(v^*) > 0,$$

whose sign is that of $t - k'(v^*)$: a higher per-post cost widens the cutoff, readmitting the posters at the margin, exactly when the marginal forgery slope $k'(v^*)$ sits below the cutoff. Welfare along the path is $W = S + \gamma Q(d(t))$. The posting surplus obeys $\mathrm{d}\frac{S}{\mathrm{d}}\tau = -\frac{t}{2} - \left(\frac{t^2}{4}\right)\left(\mathrm{d}\frac{v^*}{\mathrm{d}}\tau\right)$, the envelope terms cancelling because the cutoff type is indifferent; the accuracy term contributes $\gamma\,Q'(d)\left(\Delta'\frac{t}{\sigma}\right)\left(\mathrm{d}\frac{t}{\mathrm{d}}\tau\right)$, with accuracy slope $Q'(d) > 0$ (Lemma 2), whose sign is again that of $t - k'(v^*)$. Collecting,

$$\frac{\mathrm{d}W}{\mathrm{d}\tau} = \frac{t}{D_v}\left[\frac{1}{4}\left(t - 2k'(v^*) - \frac{\kappa_t t^2}{a}\right) + \gamma\frac{Q'(d)\Delta'(t)}{\sigma a}(t - k'(v^*))\right].$$

When $k'(v^*) < t$ the accuracy term's coefficient is positive, so $\mathrm{d}\frac{W}{\mathrm{d}}\tau > 0$ if and only if the accuracy weight exceeds the threshold that sets the bracket to zero,

$$\bar{\gamma}(\tau, c) = -\frac{\sigma a}{4Q'(d)\Delta'(t)}\frac{t - 2k'(v^*) - \kappa_t\frac{t^2}{a}}{t - k'(v^*)}.$$

When instead $k'(v^*) \geq t$, both terms of the bracket are nonpositive and at least one is strictly negative: the posting term satisfies $t - 2k'(v^*) - \kappa_t\frac{t^2}{a} \leq -t < 0$, and the accuracy term carries the sign of $t - k'(v^*) \leq 0$. Hence $\mathrm{d}\frac{W}{\mathrm{d}}\tau < 0$ at every $\gamma > 0$. On the positive-stringency full-

participation segment, the cutoff is pinned at $t = 1$, every κ_t term drops out, and the derivative reduces to Step 4′s $-\frac{1}{2} + \frac{1}{4k'(v^*)}$. At the cost-only corner, Step 4′s right derivative $-\frac{1}{2}$ applies.

When the primitive range condition fails, $\kappa_t < 0$ is possible; the numerator, the denominator D_v, and Step 3′s residual monotonicity are then unsigned, and none of the interior-branch conclusions is claimed. With K decreasing at the relevant discriminability, a per-post cost that thins the forum can move d toward a harder-to-deter region and raise the required verification stringency. The proposition conditions on the range for exactly this reason. □