WORKING PAPERS

N° 1714

February 2026

"Public Persuasion with Endogenous Fact-Checking"

Georgy Lukyanov and Samuel Safaryan

Toulouse School of Economics

Public Persuasion with Endogenous Fact-Checking

Georgy Lukyanov* Samuel Safaryan†

Abstract

We study public persuasion when a sender communicates with a large audience that can fact-check at heterogeneous costs. The sender commits to a public information policy before the state is realized, but any verifiable claim she makes after observing the state must be truthful (an ex-post implementability constraint). Receivers observe the public message and then decide whether to verify; this selective verification feeds back into the sender's objective and turns the design problem into a constrained version of Bayesian persuasion. Our main result is a reverse comparative static: when fact-checking becomes cheaper in the population, the sender optimally supplies a strictly less informative public signal. Intuitively, cheaper verification makes bold claims invite scrutiny, so the sender coarsens information to dampen the incentive to verify. We also endogenize two ex-post instruments—continuous falsification and fixed-cost repression—and characterize threshold substitutions from persuasion to manipulation and, ultimately, to repression as monitoring improves. The framework provides testable predictions for how transparency, manipulation, and repression co-move with changes in verification technology.

Keywords: Bayesian persuasion; information design; verifiable evidence; costly verification; public signals; Blackwell informativeness; falsification; repression.
JEL: D72; D82; D83; L14

1 Introduction

We analyze a public persuasion problem in which a sender communicates with a mass audience that can *endogenously verify* the state at heterogeneous costs. The sender commits ex ante to a public information policy (an experiment over messages), observes the state, and then sends a public message. Ex post, any verifiable claim must be *truthful*: given receivers' equilibrium verification, the sender cannot benefit from misreporting verifiable content. This

*Toulouse School of Economics; georgy.lukyanov@tse-fr.eu
†National Research University — Higher School of Economics; sesafaryan@edu.hse.ru

ex-post incentive (truthfulness) constraint (EPIC) serves as an *implementability* restriction on top of Bayes plausibility, yielding a constrained information-design problem in the spirit of Kamenica and Gentzkow (2011) with feasibility frictions.

Receivers observe the public message, form a posterior μ, and then decide whether to incur a private cost (drawn from F) to verify the state. This generates a cutoff rule and a verifying mass $\lambda(\mu; F)$ that depends on the public posterior. Folding selective verification into the sender's objective induces an indirect value $v(\mu; F)$; because actions aggregate across a continuum of receivers, the *distribution* of induced posteriors (not only the mean) matters for outcomes, echoing the role of higher-order uncertainty in public-signal environments. The sender's design problem is therefore a *constrained concavification*: choose a distribution over posteriors subject to Bayes plausibility and EPIC implementability to maximize the expected value of $v(\mu; F)$.

Our first main result is a stark comparative static that reverses the usual logic from single-receiver persuasion with costly learning.

If the verification-cost distribution improves in the first-order stochastic dominance sense (verification becomes cheaper), then the sender's indirect payoff $v(\mu; F)$ becomes more concave in μ. Consequently, every optimal public experiment is (strictly) less informative in the Blackwell order.

Intuitively, when more receivers are willing to verify, extreme public beliefs trigger scrutiny and attenuate the sender's gains; concavity increases, and the optimal policy coarsens—"confusion as strategy." The result hinges on EPIC implementability in a mass-audience environment; without the truthfulness friction or with full commitment to unverifiable claims, cheaper outside information can instead induce *more* precise signaling.

We then endogenize two ex post instruments that many applications feature: *falsification*, a continuous manipulation technology that distorts observed outcomes at convex marginal cost (e.g., fabricated engagement or padded counts), and *repression*, a fixed-cost discrete tool that directly shifts aggregate actions. We characterize threshold substitutions from persuasion to falsification and, once verification is sufficiently cheap, to repression. These results provide a unified, tractable account of how a sender reallocates effort across persuasion and post-message instruments as monitoring improves.

Methodologically, the paper contributes a clean formulation of public persuasion with selective verification as a constrained concavification problem: Bayes-plausible posteriors filtered through EPIC implementability. Substantively, it delivers a reverse comparative static—cheaper verification $\Rightarrow$ optimally *less* informative public signals—and sharp substitution thresholds toward manipulation and repression. The framework complements Bayesian persuasion, mechanisms with evidence, constrained information design, and work emphasizing the role of belief dispersion in public environments, while contrasting with models where cheaper learning induces more transparent signaling.

2 Related literature

Our paper connects and advances four strands: (i) information design and Bayesian persuasion with feasibility/frictions, (ii) persuasion when receivers can acquire (or verify) information at a cost, (iii) communication with verifiable evidence and limited commitment, and (iv) multi-receiver/public-signal environments where belief dispersion matters, with applications to political economy.

We build on the persuasion paradigm of Kamenica and Gentzkow (2011), treating the sender's problem as a concavification over posteriors subject to Bayes plausibility. Within the broader information-design program (see Bergemann and Morris, 2016; Kamenica, 2019), we introduce an explicit *implementability* restriction: after the state is realized, any verifiable claim must pass an *ex post* truthfulness (EPIC) constraint. Methodologically, this complements recent work on *constrained* information design (e.g., Doval and Skreta, 2022), which shows how auxiliary feasibility constraints can be embedded into the design problem and alter the shape of the sender's indirect value. Our contribution is to pin down how an EPIC-style constraint—motivated by hard-evidence disclosure—interacts with the equilibrium verification response of a mass audience, and to derive a sharp comparative static for Blackwell informativeness.

A growing literature endogenizes the receiver's acquisition of outside information within persuasion. In a single-receiver benchmark, Matysková and Montes (2023) show that a sender may *increase* public informativeness to deter costly learning by the receiver; related forces appear when the receiver can verify probabilistically at a cost (e.g., Yang, 2024). Our environment differs in two respects: (i) a *mass audience* self-selects into verification, so the sender cares about the cross-sectional distribution of posteriors that triggers verification thresholds; and (ii) EPIC implementability binds ex post. These features overturn the usual deterrence logic: when the cost distribution improves in the FOSD sense, we show that the induced value function $v(\mu; F)$ becomes more concave and every optimal public experiment becomes strictly less informative in the Blackwell order. In short, cheaper verification leads to *coarser* public information in our setting, in contrast to the single-receiver/full-commitment predictions.[1]

Our EPIC restriction is in the spirit of mechanisms with verifiable evidence. Classic disclosure theory (Grossman, 1981; Milgrom, 1981; Shin, 1994, 2003) shows that, with hard

[1]See also work on persuasion with privately informed or strategic receivers (e.g., Kolotilin et al., 2017; Arieli and Babichenko, 2019), which highlights how informational frictions on the receiver side shape optimal experiments. Our mechanism is distinct: it operates through selective verification by a continuum of receivers and an EPIC friction on the sender.

evidence and no frictions, non-disclosure tends to unravel. In mechanism-design environments with evidence, Ben-Porath et al. (2019) establish powerful commitment and robustness properties, while in persuasion settings Titova (2022) explores how verifiable information restores commitment payoffs under appropriate conditions. We differ in two ways: we study a *public* signal to a mass audience that endogenously verifies, and our EPIC feasibility requirement is coupled with a minimal delivery friction to prevent complete unraveling of silence. Relative to certifiable-message models (e.g., Seidmann and Winter, 1997), we keep the hard-evidence discipline but embed it as an *implementability constraint* inside an information-design problem rather than as a full-blown disclosure game. This generates tractable, testable predictions for how credibility constraints reshape optimal public experiments.

In multi-receiver persuasion, the sender's payoff depends on how posteriors distribute across an audience rather than on a representative posterior; see, e.g., Caillaud and Tirole (2007) for group persuasion and Arieli and Babichenko (2019) for private signaling. Our analysis is closest to the *public*-signal case with independent actions: the verifying mass $\lambda(\mu; F)$ depends on the public posterior, so the sender effectively optimizes over the *distribution of posteriors*. This resonates with the macro/coordination literature's emphasis on the *dispersion* of beliefs under public information (Morris and Shin, 2002). We formalize this within persuasion: selective verification makes $v(\mu; F)$ curvature the pivotal object, and a FOSD drop in verification costs increases that curvature, pushing optimal experiments toward coarser partitions.

Beyond persuasion, we allow the sender to (i) *falsify* outcomes at a convex marginal cost and (ii) use *repression/violence* at a fixed cost. The falsification margin is linked to design with manipulable inputs (e.g., Skreta and Pérez-Richet, 2022), while the repression margin connects to models and evidence on *informational autocracy* and the substitution between propaganda and force (Guriev and Treisman, 2019, 2020; Gehlbach et al., 2021). Our contribution is to embed these instruments *within* a constrained persuasion problem and characterize threshold substitutions as verification becomes cheaper: the sender first coarsens public information, then shifts into falsification, and finally resorts to repression once persuasion and manipulation can no longer sustain desired actions.

Relative to (i) unconstrained persuasion, we show how EPIC feasibility reshapes concavification and flips the key comparative static under improved verification; relative to (ii) persuasion with costly learning, we identify the mass-audience/EPIC channel that yields *less* transparency when verification becomes cheaper; relative to (iii) evidence/limited-commitment models, we use hard-evidence logic as a tractable implementability constraint in a public-signal environment; and relative to (iv) multi-receiver/public-signal work, we make the *distribution* of posteriors the optimizing object through the endogenous verification margin. The unified framework also rationalizes observed substitution patterns between persuasion, manipulation, and repression in political-economy applications.

Roadmap. Section 3 presents the environment and EPIC implementability. Section 4 derives the sender's indirect value and the reverse comparative static. Section 5 introduces falsification and repression and characterizes substitution thresholds. Section 6 discusses robustness and applications; Section 7 concludes.

3 Model and Implementability under Evidence

3.1 Environment and timing

There is a Sender (S) and a unit mass of Receivers indexed by $i \in [0,1]$. The state is $\theta \in \{0,1\}$ with prior $\Pr(\theta = 1) = \pi \in (0,1)$ (we shall specialize to $\pi = \frac{1}{2}$ when convenient). The Sender's bias is $b \geq 0$.

Each Receiver i chooses an action $a_i \in [0,1]$ after observing a public message and (possibly) verifying the state. Her loss is quadratic, $\ell_i(a_i, \theta) = (a_i - \theta)^2$, and verifying the state costs $c_i \geq 0$. Costs are i.i.d. across Receivers with continuous cdf F on $[0, \bar{c}]$; we use first-order stochastic dominance (FOSD) shifts of F to model cheaper verification in the population. The aggregate action is

$$A = \int_0^1 a_i \, di \in [0,1].$$

The Sender's loss is quadratic in the distance between the aggregate and a bias-shifted state:

$$L_S(A, \theta) = \Big(A - (\theta + b)\Big)^2.$$

(Results below extend to any Sender loss that is convex and symmetric around $\theta + b$.)

Timing is as follows.

1. **Design.** Before learning θ, S commits to a public information policy and to a reporting rule that must satisfy *ex-post truthfulness* (EPIC) once θ is realized.[2]
2. **Realization and reporting.** Nature draws θ. The reporting rule may produce a verifiable disclosure $e \in \{0,1\}$ ("hard evidence") or *silence*. With a small, exogenous probability $\varepsilon \in (0,1)$, verifiable disclosure fails to reach the public (delivery friction).
3. **Verification and actions.** Receivers observe the public message $m \in \{e = 0, e = 1, \text{silence}\}$, form a public posterior $\mu = \Pr(\theta = 1 \mid m)$, and then decide whether to verify the state at private cost c_i. Verifiers learn θ perfectly and then choose $a_i = \theta$; non-verifiers choose $a_i = \mu$ (quadratic loss).

[2]We impose truthfulness only on *verifiable* (hard-evidence) reports. Soft labels are chosen ex ante as part of the public experiment and are not constrained ex post. This limited-commitment friction prevents full-commitment disclosure from collapsing to full revelation and gives bite to the design problem.

3.2 Public experiments and EPIC-implementability

A *public experiment* is a distribution Π over posteriors $\mu \in [0,1]$ that is Bayes-plausible: $\mathbb{E}_\Pi[\mu] = \pi$. We interpret Π as being generated by the triplet $(\delta_0, \delta_1, \mu_s)$ consisting of disclosure probabilities $\delta_\theta \in [0,1]$ in state θ and the posterior $\mu_s \in (0,1)$ induced by silence. With the delivery friction ε[3], Bayes' rule gives

$$\begin{aligned}\mu_s &= \Pr(\theta = 1 \mid \text{silence}) \\ &= \frac{(1-\delta_1+\varepsilon\,\delta_1)\,\pi}{(1-\delta_1+\varepsilon\,\delta_1)\,\pi \;+\; (1-\delta_0+\varepsilon\,\delta_0)\,(1-\pi)}.\end{aligned} \tag{3.1}$$

When verifiable evidence is observed, $\mu = 1$ after $e = 1$ and $\mu = 0$ after $e = 0$.

After observing θ, the Sender must not strictly prefer deviating from the committed reporting rule. With hard evidence, the only feasible deviation is to *withhold* disclosure. Thus EPIC imposes the following inequalities:

$$\mathbb{E}[-L_S(A,\theta) \mid \theta = 1,\ e = 1] \;\geq\; \mathbb{E}[-L_S(A,\theta) \mid \theta = 1,\ \text{silence}]\,; \tag{EPIC-1}$$

$$\mathbb{E}[-L_S(A,\theta) \mid \theta = 0,\ \text{silence}] \;\geq\; \mathbb{E}[-L_S(A,\theta) \mid \theta = 0,\ e = 0]\,. \tag{EPIC-0}$$

Under quadratic losses and the behavior described below, (EPIC-1) binds as a strict inequality so that $\delta_1 = 1$; (EPIC-0) binds at equality when $\delta_0 \in (0,1)$ and pins down the silence posterior μ_s (the object that will respond to F).

3.3 Receivers: verification and actions

Given a public posterior $\mu \in [0,1]$ and quadratic loss, a Receiver who does not verify chooses $a_i = \mu$, yielding expected loss $\mathbb{E}[(\mu-\theta)^2 \mid \mu] = \mu(1-\mu)$. By verifying, she learns θ and sets $a_i = \theta$, yielding loss 0 but paying c_i. Hence the *benefit* of verification at belief μ equals $\mu(1-\mu)$. The equilibrium verification rule is a cutoff:

$$\text{verify} \iff c_i \leq c^*(\mu) := \mu(1-\mu), \tag{3.2}$$

$$\lambda(\mu;F) := \Pr(c_i \leq c^*(\mu)) = F\Big(\mu(1-\mu)\Big). \tag{3.3}$$

We call $\lambda(\mu;F)$ the *verifying mass* at belief μ. A FOSD reduction of F increases $\lambda(\mu;F)$ pointwise for all $\mu \in (0,1)$.

[3]Interpret ε as a vanishing probability that verifiable evidence fails to reach the public (outages, censorship glitches). We keep $\varepsilon > 0$ only to avoid full unraveling of silence and take $\varepsilon \to 0$ when convenient; none of the comparative statics hinge on its exact value.

Given μ and $\lambda(\mu; F)$, the realized aggregate action is

$$A(\mu, \theta; F) \;=\; (1 - \lambda(\mu; F))\,\mu \;+\; \lambda(\mu; F)\,\theta. \tag{3.4}$$

Note that $A(\mu, \theta; F)$ is random through θ even conditional on μ.

3.4 The Sender's indirect payoff at belief μ

Define the Sender's expected *indirect value* at belief μ as

$$\begin{aligned} v(\mu; F) \;&:=\; -\mathbb{E}\Big[L_S\Big(A(\mu, \theta; F), \theta\Big) \,\Big|\, \mu\Big] \\ &=\; -\Big(b^2 + (1 - \lambda(\mu; F))^2\,\mu(1 - \mu)\Big). \end{aligned} \tag{3.5}$$

The first term, $-b^2$, is the loss from the bias; the second term increases in the posterior variance $\mu(1 - \mu)$ and decreases with the verifying mass. When F FOSD-decreases, $\lambda(\mu; F)$ increases and $v(\mu; F)$ becomes *more concave* in μ.

3.5 Implementable experiments and EPIC in closed form

Because $\delta_1 = 1$ under (EPIC-1), the only nontrivial implementability restriction is (EPIC-0). Under quadratic losses we have

$$\mathbb{E}[-L_S(A, \theta) \mid \theta = 0,\, e = 0] \;=\; -\,b^2, \tag{3.6}$$

$$\begin{aligned} \mathbb{E}[-L_S(A, \theta) \mid \theta = 0, \text{ silence}] \;&=\; -\,\Big(A(\mu_s, 0; F) - b\Big)^2 \\ &=\; -\,\Big((1 - \lambda(\mu_s; F))\,\mu_s - b\Big)^2. \end{aligned} \tag{3.7}$$

Therefore the (EPIC-0) condition binds at the Sender's optimum whenever $\delta_0 \in (0, 1)$ and delivers the *indifference equation*[4]

$$(1 - \lambda(\mu_s; F))\,\mu_s \;=\; 2b. \tag{3.8}$$

Equation (3.8) is the key implementability restriction that determines the equilibrium silence posterior μ_s as a function of b and F.

[4] This equates the sender's continuation loss under silence in $\theta = 0$, $\big((1 - \lambda(\mu_s; F))\,\mu_s - b\big)^2$, to the loss under truthful disclosure, b^2. It implies $\mu_s \geq \frac{1}{2}$ whenever $b > 0$.

Bayes plausibility. The disclosure probabilities (δ_0, δ_1) and the silence posterior μ_s must satisfy $\mathbb{E}_\Pi[\mu] = \pi$, that is

$$\pi \;=\; (1-\varepsilon)\,\delta_1 \cdot 1 \;+\; (1-\varepsilon)\,\delta_0 \cdot 0 \;+\; \Big(1-(1-\varepsilon)(\delta_0+\delta_1)\Big)\,\mu_s. \tag{3.9}$$

Given π and μ_s that solve (3.8), (3.9) pins down the set of feasible (δ_0, δ_1); under (EPIC-1) we take $\delta_1 = 1$ and solve for $\delta_0 \in [0,1]$ (feasible for all sufficiently small ε).

3.6 Equilibrium

An *implementable public persuasion equilibrium* is a triplet $(\delta_0^*, \delta_1^*, \mu_s^*)$ and a verification rule $\lambda(\cdot; F)$ such that (i) Receivers use the cutoff rule (3.2); (ii) posteriors are given by Bayes' rule (3.1); (iii) EPIC holds, with $\delta_1^* = 1$ and μ_s^* solving (3.8); and (iv) the Sender's experiment (i.e., the choice of μ_s and thus the distribution Π) maximizes $\mathbb{E}_\Pi[v(\mu; F)]$ subject to Bayes plausibility and (3.8).

3.7 A handy special case and a threshold

Suppose $\pi = \frac{1}{2}$, $\varepsilon \to 0$, and verification costs are uniformly distributed on $[0,1]$, i.e., $F(x) = x$ on $[0,1]$. Then $\lambda(\mu; F) = \mu(1-\mu)$ by (3.2), and (3.8) becomes

$$\Big(1 - \mu_s(1-\mu_s)\Big)\,\mu_s \;=\; 2b. \tag{3.10}$$

The left-hand side of (3.10) is maximized at $\mu_s = \frac{1}{2}$, where it equals $\frac{3}{8}$. Hence:

Proposition 3.1. *In the uniform-cost special case, if $b \geq \frac{3}{16}$*[5] *then the unique solution to (3.10) is $\mu_s^* = \frac{1}{2}$ and the verifying mass at silence equals $\lambda(\mu_s^*) = \frac{1}{4}$. If $0 < b < \frac{3}{16}$, there is a unique interior solution $\mu_s^*(b) \in (\frac{1}{2}, 1)$ characterized by (3.10).*

Proof. The function $g(\mu) := (1-\mu(1-\mu))\mu$ is strictly increasing on $[\frac{1}{2}, 1]$ with $g(\frac{1}{2}) = \frac{3}{8}$ and $g(1) = 1$. For $b \geq \frac{3}{16}$ the equation $g(\mu) = 2b$ has the boundary solution $\mu^* = \frac{1}{2}$; for $0 < b < \frac{3}{16}$ the Intermediate Value Theorem and monotonicity deliver a unique interior root. The expression for $\lambda(\mu^*)$ follows from $\lambda(\mu) = \mu(1-\mu)$. □

Proposition 3.1 reproduces, within our implementability framework, the threshold pattern that motivates the empirical narrative: as the Sender's bias b grows, the silence posterior saturates at $\mu_s = \frac{1}{2}$ and the verifying mass at silence pins at 1/4. Section 4 uses $v(\mu; F)$ to derive the general comparative static (FOSD-improvement of $F \Rightarrow$ optimal coarsening of Π).

[5] With $F(x) = x$, $g(\mu) := (1-\mu(1-\mu))\mu$ attains its maximum $\frac{3}{8}$ at $\mu = \frac{1}{2}$; since EPIC sets $g(\mu_s) = 2b$, the corner binds at $2b = \frac{3}{8}$.

4 Persuasion and the Optimal Public Experiment

This section solves the Sender's ex-ante problem. The object of design is the *distribution of public posteriors* Π (a public experiment). Given Π, Receivers' optimal verification behavior and actions are as in Section 3. The Sender's objective at posterior μ is the indirect value

$$v(\mu; F) \;=\; -\Big(b^2 + (1-\lambda(\mu;F))^2\,\mu(1-\mu)\Big),$$
$$\lambda(\mu;F) = F\Big(\mu(1-\mu)\Big),$$

and the Sender chooses Π to maximize $\mathbb{E}_\Pi[v(\mu;F)]$ subject to Bayes plausibility and implementability.. Until then, "benchmark persuasion" refers to the unconstrained information-design problem in which any Bayes-plausible Π is feasible.

4.1 Concavification benchmark

In the benchmark persuasion problem, the feasible set of posterior laws is

$$\mathcal{X}(\pi) \;=\; \Big\{\Pi \text{ probability law on } [0,1] :\; \mathbb{E}_\Pi[\mu] = \pi\Big\}.$$

The Sender solves

$$V(\pi, F) \;=\; \sup_{\Pi\in\mathcal{X}(\pi)} \;\mathbb{E}_\Pi[v(\mu;F)]. \tag{4.1}$$

Lemma 4.1. *For any F, the function $v(\cdot;F)$ is continuous, symmetric around $\mu = \frac{1}{2}$, and single-peaked with a (global) maximum at $\mu \in \{0, \frac{1}{2}, 1\}$. Moreover, if F FOSD-decreases to F', then $v(\cdot;F')$ is a pointwise* mean-preserving contraction *of $v(\cdot;F)$ in the sense that for any $\mu \in (0,1)$,*

$$v(\mu;F') - v(\tfrac{1}{2};F') \;\le\; v(\mu;F) - v(\tfrac{1}{2};F),$$

with equality at $\mu \in \{0, \frac{1}{2}, 1\}$.

Proof. Symmetry follows from symmetry of $\mu(1-\mu)$ and of $\lambda(\mu;F) = F(\mu(1-\mu))$. Since $x \mapsto (1-F(x))^2 x$ is increasing on $[0, \frac{1}{4}]$, the "penalty term" $(1-\lambda)^2\mu(1-\mu)$ is minimized at the endpoints and (weakly) maximized near $\mu = \frac{1}{2}$, yielding single-peakedness. If $F' \succeq_{\text{FOSD}} F$, then $1-\lambda(\mu;F') \le 1-\lambda(\mu;F)$ for all μ, hence $(1-\lambda(\mu;F'))^2\mu(1-\mu) \le (1-\lambda(\mu;F))^2\mu(1-\mu)$ with equality at $\mu \in \{0,1\}$ (where $\mu(1-\mu) = 0$), establishing the contraction around $\mu = \frac{1}{2}$. ∎

Let $\operatorname{cav} v(\cdot;F)$ denote the concave envelope of $v(\cdot;F)$ on $[0,1]$. By the Kamenica–Gentzkow method, the value in (4.1) equals $(\operatorname{cav} v(\cdot;F))(\pi)$ and an optimal experiment puts probability on at most two posteriors $\mu_L \le \pi \le \mu_H$ with a supporting line to $v(\cdot;F)$ at these points.

Theorem 4.2. *Fix $\pi \in (0,1)$ and let F', F satisfy $F' \succeq_{\text{FOSD}} F$ (verification becomes cheaper). Let Π^*_F and $\Pi^*_{F'}$ be optimal solutions to (4.1) at F and F', respectively. Then there exist binary optimal solutions*

$$\Pi^*_F = \alpha\,\delta_{\mu_H} + (1-\alpha)\,\delta_{\mu_L},$$
$$\Pi^*_{F'} = \alpha'\,\delta_{\mu'_H} + (1-\alpha')\,\delta_{\mu'_L},$$

with $\mu_L \le \mu'_L \le \pi \le \mu'_H \le \mu_H$. In particular, $\Pi^*_{F'}$ is a mean-preserving contraction of Π^*_F and is Blackwell less informative.

Proof sketch. By Lemma 4.1, $v(\cdot;F')$ is a contraction of $v(\cdot;F)$ around $\mu = \frac{1}{2}$. Concavification tightens as the objective gets more concave: the supporting chord at π intersects $v(\cdot;F')$ at posteriors closer to π than under F. Formally, let $\ell(\mu) = \tau\mu + \kappa$ be the common supporting line at μ_L, μ_H for $v(\cdot;F)$; by pointwise contraction, for the same slope τ the contact points with $v(\cdot;F')$ lie weakly inside $[\mu_L, \mu_H]$, yielding μ'_L and μ'_H that satisfy the stated inequalities. Mean preservation pins the weights. Blackwell comparisons follow because binary experiments are ordered by spread when the mean is fixed. □

When fewer people verify, the Sender benefits from sharpening (making the public signal more extreme); when more people verify, the value function becomes more concave in the posterior, so the Sender best responds by coarsening—*creating confusion* by design.

4.2 EPIC-implementable persuasion

We now impose the implementability (EPIC) constraints from Section 3. Write the feasible set as

$$\mathcal{X}_{\text{EPIC}}(\pi; F, b, \varepsilon) \subseteq \mathcal{X}(\pi),$$

the (convex) set of Bayes-plausible posterior laws that arise from some reporting rule and experiment for which, after the state is realized, the Sender prefers truthful reporting to any deviation ("withholding" or misreporting), given Receivers' equilibrium verification behavior. We study two protocols.

Protocol A: Hard evidence + silence. Messages are $e \in \{0,1\}$ (verifiable evidence) or silence. As shown in Section 3, EPIC implies $\delta_1 = 1$ (truthful disclosure in $\theta = 1$) and pins the *silence posterior* μ_s in $\theta = 0$ by the indifference condition

$$(1 - \lambda(\mu_s; F))\,\mu_s \;=\; 2b. \tag{4.2}$$

Hence any implementable Π places mass only on $\{0, \mu_s, 1\}$, with weights determined by Bayes plausibility and ε.

Proposition 4.3. *If $F' \succeq_{\text{FOSD}} F$, then the solution $\mu_s(F)$ of (4.2) satisfies $\mu_s(F') \geq \mu_s(F)$, with strict inequality whenever $b > 0$ and $F' \neq F$.*

Proof. Define $\phi(\mu; F) := (1 - \lambda(\mu; F))\mu = (1 - F(\mu(1-\mu)))\mu$. For fixed $b > 0$, the equation $\phi(\mu; F) = 2b$ defines $\mu_s(F)$; FOSD-shifts of F decrease $\phi(\cdot; F)$ pointwise on $(0,1)$, so the unique solution must move weakly right. □

Under Protocol A, a decrease in verification costs raises the interior posterior μ_s. Because $\delta_1 = 1$, Bayes plausibility forces the probability of silence to adjust. Whether the overall experiment becomes more or less informative (in the Blackwell sense) is *a priori ambiguous*: Π concentrates more weight on the extremes $\{0, 1\}$ (tending to *increase* precision) but also moves the interior point rightward (tending to *decrease* the usefulness of silence). This protocol is therefore too restrictive to guarantee the benchmark coarsening result.

Protocol B: Minimal soft layer with evidence. Augment Protocol A with a *soft* public label $m \in \{L, H\}$ that is chosen according to a state-dependent experiment fixed ex ante (probabilities (a, b) with $a = \Pr[m = H \mid \theta = 1]$, $b = \Pr[m = H \mid \theta = 0]$). Messages m are not verifiable; the Sender can always disclose $e \in \{0, 1\}$ truthfully, and EPIC prohibits misreporting (a, b) after the state is realized. Receivers observe (m, e) jointly, update to posteriors $\mu \in [0, 1]$, and then verify optimally. With $\varepsilon > 0$, any posterior $\mu \in (0, 1)$ can be generated with both states occurring with positive probability; the EPIC constraints reduce to statewise no-regret inequalities, which (given quadratic loss and the verification responses) impose:

1. If $\theta = 1$, any interior μ used with positive probability must satisfy

$$(1 - \lambda(\mu; F))(1 - \mu) = 0 \;\Rightarrow\; \mu = 1; \tag{4.3}$$

2. If $\theta = 0$, any interior μ used with positive probability must satisfy

$$(1 - \lambda(\mu; F))\mu = 2b. \tag{4.4}$$

Hence the set of EPIC-admissible posteriors used in $\theta = 0$ is the *interval*

$$\mathcal{M}_0(F, b) \;=\; \Big\{\mu \in [\underline{\mu}(F, b), \overline{\mu}(F, b)]\Big\}$$

with $\overline{\mu}(F, b)$ solving (4.4), and $\underline{\mu}(F, b) \in [0, \frac{1}{2}]$.

Because $\theta = 1$ uses only $\mu = 1$ by (4.3), any implementable experiment under Protocol B has support contained in $\{1\} \cup \mathcal{M}_0(F, b)$ and is convex in the sense of mixtures over $\mathcal{M}_0(F, b)$.

Theorem 4.4. *Fix $\pi \in (0, 1)$ and assume Protocol B. Let $\Pi_F^* \in \mathcal{X}_{\text{EPIC}}(\pi; F, b, \varepsilon)$ be optimal at F. If $F' \succeq_{\text{FOSD}} F$, then there exists an optimal $\Pi_{F'}^* \in \mathcal{X}_{\text{EPIC}}(\pi; F', b, \varepsilon)$ that is a mean-preserving contraction of Π_F^* and hence Blackwell less informative. In particular, if Π_F^* is binary with support $\{\mu_L, 1\}$, then $\Pi_{F'}^*$ is binary with support $\{\mu_L', 1\}$ and $\mu_L' \geq \mu_L$ (the support collapses toward 1).*

Proof sketch. Under Protocol B, the feasible set of posterior laws at a fixed mean π is convex and contains all binary distributions supported on $\{\mu, 1\}$ with $\mu \in \mathcal{M}_0(F, b)$; $\mathcal{M}_0(F', b)$ shifts right as $F' \succeq_{\text{FOSD}} F$ (by the same monotonicity as in Proposition 4.3). By Lemma 4.1, $v(\cdot; F')$ is a contraction of $v(\cdot; F)$, so the concavification at π requires *less* spread in μ once the left endpoint is constrained to lie in a set that shifts right. The binary-support characterization then yields $\mu_L' \geq \mu_L$ and the Blackwell comparison follows. □

The benchmark persuasion comparative static (Theorem 4.2) survives EPIC once we allow a minimal, non-verifiable "label" alongside evidence (Protocol B). By contrast, the pure hard-evidence protocol (Protocol A) is too narrow: it pins a single interior posterior μ_s and may shift the mass of posteriors in ways that do not line up monotonically with Blackwell precision.

4.3 Worked-out symmetric special case

Take $\pi = \frac{1}{2}$ and assume F has a twice continuously differentiable density on $[0, 1]$ with $F(0) = 0$ and $F(1) = 1$. Under Protocol B, the optimal experiment is binary with support $\{\mu_L, 1\}$ and weight $\alpha = \frac{\frac{1}{2} - \mu_L}{1 - \mu_L}$ on $\mu = 1$. The tangency condition reads

$$\begin{aligned} v'(\mu_L; F) &= \frac{v(1; F) - v(\mu_L; F)}{1 - \mu_L} \\ &= \frac{-b^2 + \left(b^2 + (1 - \lambda(\mu_L; F))^2 \mu_L (1 - \mu_L)\right)}{1 - \mu_L}. \end{aligned}$$

As F FOSD-decreases to F', we have $(1 - \lambda(\mu_L; F')) \leq (1 - \lambda(\mu_L; F))$, so the right-hand side contracts and the solution μ_L moves right. Hence the spread $1 - \mu_L$ shrinks and the posterior law coarsens.

Under the uniform-cost example and $\varepsilon \to 0$, (4.2) yields $\mu_s^3 - \mu_s^2 + \mu_s - 2b = 0$ and $\lambda(\mu_s) = \mu_s(1 - \mu_s)$ as in Section 3.7. Under Protocol B, the left endpoint μ_L is constrained by $\mu_L \leq \mu_s$ and moves right as F improves, so the induced spread shrinks.

4.4 What to measure as "precision"

We use the Blackwell (garbling) order to notionally rank informativeness. For binary-support posterior laws at a fixed mean, less spread is less informative. In the tri-point Protocol A, informativeness is not monotone in $(\mu_s, \text{Prob}[\text{silence}])$; by contrast, under Protocol B the one-sided binary support $\{\mu_L, 1\}$ yields a clean monotone relation between μ_L and information (spread $1 - \mu_L$).

In sum, our main comparative static—*cheaper verification* $\Rightarrow$ *less precise public information*—is a robust property of the concavification problem and holds under EPIC as soon as the protocol allows a minimal soft layer alongside evidence.

5 Falsification and Violence

We enrich the baseline with two ex–post instruments that the Sender (government, platform, firm) can deploy *after* the public message has been sent and Receivers have (possibly) verified and acted:

1. **Falsification** allows the Sender to distort the realized aggregate action A to a nearby value A', at a convex cost that grows with the magnitude of the distortion.
2. **Violence** (repression) is a fixed–setup technology that raises the aggregate by a discrete amount (e.g., by silencing or removing a hostile mass of Receivers), at a fixed cost $K > 0$ and possibly a small variable cost.

These instruments operate *after* the state θ is realized and after Receivers choose actions based on the public posterior μ and their verification decisions. Hence the EPIC constraints developed in Sections 3–4 continue to apply to the reporting rule; the new instruments only change the Sender's continuation payoff at each posterior realization.

5.1 Falsification

Let the post–action aggregate under posterior μ be

$$A(\mu, \theta; F) = (1 - \lambda(\mu; F))\, \mu + \lambda(\mu; F)\, \theta,$$
$$\lambda(\mu; F) = F\Big(\mu(1 - \mu)\Big).$$

After observing θ and $A(\mu, \theta; F)$, the Sender can choose a distortion $d \in \mathbb{R}$ to implement $A' = \Pi_{[0,1]}\Big(A(\mu, \theta; F) + d\Big)$ where $\Pi_{[0,1]}$ denotes truncation to $[0, 1]$. The falsification cost is $c_f(d)$, where c_f is convex, even, $c_f(0) = 0$, and $c_f'(0) = 0$. The (state–wise) post–instrument loss is

$$\mathcal{L}_f\Big(d; \mu, \theta\Big) \;=\; \Big(A(\mu, \theta; F) + d - (\theta + b)\Big)^2 + c_f(d),$$

and the Sender chooses d to minimize $\mathcal{L}_f$.

If $c_f(d) = \frac{\kappa}{2}d^2$ with $\kappa > 0$, the unique minimizer is

$$d^*(\mu,\theta) \;=\; \frac{2}{2+\kappa}\Big((\theta+b) - A(\mu,\theta;F)\Big). \tag{5.1}$$

The minimized *state–wise* loss equals

$$\mathcal{L}_f\Big(d^*;\mu,\theta\Big) \;=\; \frac{\kappa}{2+\kappa}\Big((\theta+b) - A(\mu,\theta;F)\Big)^2,$$

so the Sender's *indirect value* at posterior μ becomes

$$\begin{aligned} v^f(\mu;F,\kappa) \;&=\; -\frac{\kappa}{2+\kappa}\,\mathbb{E}\Big[\Big((\theta+b) - A(\mu,\theta;F)\Big)^2 \;\Big|\; \mu\Big] \\ &=\; -\frac{\kappa}{2+\kappa}\Big(b^2 + (1-\lambda(\mu;F))^2\mu(1-\mu)\Big). \end{aligned} \tag{5.2}$$

Up to the multiplicative factor $\kappa/(2+\kappa)$, the shape in μ is the same as in the baseline objective (3.5). Therefore, with *symmetric quadratic* falsification, the Sender's *persuasion choice* (the optimal posterior law Π) is unchanged; falsification simply scales up the continuation value. The instrument is then used purely as an *insurance* device, and there is no substitution margin between persuasion and falsification.

Two empirically plausible frictions break the neutrality:

- *Upward–only falsification:* $d \geq 0$ (bots/upvotes/astroturf raise the aggregate but cannot push it below what the audience organically generated).
- *Capacity constraint:* $|d| \leq \bar{d}$ for some $\bar{d} > 0$ (limited budgets or platform frictions).

Write the constrained problem as

$$\begin{aligned} &\min_{d\in D}\;\Big(A(\mu,\theta;F) + d - (\theta+b)\Big)^2 + c_f(d), \\ &D = [0,\infty) \;\text{ or }\; D = [-\bar{d},\bar{d}]. \end{aligned}$$

Let $d_D^*(\mu,\theta)$ denote the optimal constrained distortion, and define

$$v_D^f(\mu;F) \;=\; -\mathbb{E}\Big[\Big(A(\mu,\theta;F) + d_D^*(\mu,\theta) - (\theta+b)\Big)^2 + c_f\Big(d_D^*(\mu,\theta)\Big) \mid \mu\Big].$$

Proposition 5.1. *Suppose falsification is upward–only ($D = [0,\infty)$) and c_f is convex and differentiable. Then for any posterior μ:*

1. *$d_D^*(\mu,1)$ is weakly decreasing in $\lambda(\mu;F)$, while $d_D^*(\mu,0)$ is weakly* increasing *in $\lambda(\mu;F)$.*
2. *If $F' \succeq_{\text{FOSD}} F$, then for any EPIC–implementable experiment the* ex–ante *expected use of falsification (the probability–weighted mass of states in which $d_D^* > 0$) weakly*

increases.

Proof sketch. With $D = [0, \infty)$ and convex c_f, the optimal d is the projection of the unconstrained optimum onto $[0, \infty)$. From (5.1) (or directly by subgradient conditions), d^* is monotone in the shortfall $(\theta + b) - A(\mu, \theta; F)$. For $\theta = 0$, $A(\mu, 0; F) = (1 - \lambda)\mu$ decreases in λ, so the shortfall increases with λ, hence d^* increases. For $\theta = 1$, $A(\mu, 1; F) = (1 - \lambda)\mu + \lambda$ increases in λ, so the shortfall decreases, hence d^* decreases. FOSD improvements raise $\lambda(\mu; F)$ pointwise, implying the second claim after averaging over states and posteriors used in equilibrium. □

With $|d| \leq \bar{d}$, falsification can only partially offset large shortfalls. Let the residual shortfall be

$$\Delta(\mu, \theta; F, \bar{d}) \;=\; \Big|(\theta + b) - A(\mu, \theta; F)\Big| - \bar{d} \quad \text{truncated at } 0.$$

The minimized loss is bounded below by $\Delta(\mu, \theta; F, \bar{d})^2$. As verification becomes cheaper (higher λ), $\Delta(\mu, 0; F, \bar{d})$ weakly increases for any fixed μ, making *persuasion* relatively more valuable (the Sender prefers to coarsen the experiment to move mass away from posteriors that generate large residual shortfalls in state 0).

Corollary 5.2. *If $|d| \leq \bar{d}$ and $F' \succeq_{\text{FOSD}} F$, then the concavified objective in (4.1) with continuation value $v_D^f(\cdot; F)$ becomes strictly more concave in μ on any region where $\Delta(\mu, 0; F, \bar{d}) > 0$, and the optimal posterior law Blackwell–coarsens relative to the case F.*

5.2 Repression

Violence is modeled as a fixed–setup instrument that *jumps* the aggregate upward. Formally, after $A(\mu, \theta; F)$ is realized, the Sender may choose $u \in \{0, 1\}$ at fixed cost $K > 0$, and if $u = 1$ the aggregate becomes

$$A^V(\mu, \theta; F) \;=\; \min\{A(\mu, \theta; F) + \rho,\; 1\},$$

where $\rho \in (0, 1]$ is the repression *reach*. Intuitively, $u = 1$ removes or silences a mass of hostile Receivers, raising the effective average by ρ. Let $c_v(u) = K\,\mathbf{1}\{u = 1\}$.

Given (μ, θ), the Sender compares the best falsification–only continuation loss to the loss after adding violence (and optionally *then* fine–tuning with falsification):

$$L^f(\mu, \theta) := \min_{d \in D} \; \Big(A(\mu, \theta; F) + d - (\theta + b)\Big)^2 + c_f(d) \tag{5.3}$$

$$L^{Vf}(\mu, \theta) := K + \min_{d \in D} \; \Big(\min\{A(\mu, \theta; F) + \rho, 1\} + d - (\theta + b)\Big)^2 + c_f(d). \tag{5.4}$$

Violence is used in (μ, θ) iff $L^{Vf}(\mu, \theta) < L^f(\mu, \theta)$.

Proposition 5.3. *Fix F and D and suppose c_f is convex. There exists a state–contingent* gap function

$$G(\mu,\theta;F,\rho,D) \;:=\; \Big[\,\Big(A_\rho(\mu,\theta;F)-(\theta+b)\Big)^2\,\Big]^{\downarrow f} - \Big[\,\Big(A(\mu,\theta;F)-(\theta+b)\Big)^2\,\Big]^{\downarrow f},$$

where $x^{\downarrow f}$ denotes the value after optimal falsification (projection onto D and payment of c_f), such that violence is used iff $K < G(\mu,\theta;F,\rho,D)$. Moreover:

1. *$G(\mu,1;F,\rho,D)$ is weakly* decreasing *in $\lambda(\mu;F)$, while $G(\mu,0;F,\rho,D)$ is weakly* increasing *in $\lambda(\mu;F)$.*
2. *If $F' \succeq_{\text{FOSD}} F$, the ex–ante region in which violence is optimal (integrating over the equilibrium distribution of posteriors) weakly expands, and strictly expands whenever violence binds in state 0 for some posterior under F.*

Proof sketch. The function G is the violence–induced reduction in the falsification–adjusted squared gap to the target $(\theta+b)$. When $\theta=0$, $A(\mu,0;F)$ decreases with λ, so the pre–violence shortfall to the upward target b increases and the marginal benefit of a ρ–jump rises; conversely in $\theta=1$ the shortfall to $1+b$ decreases with λ. FOSD improvements raise λ pointwise, so the measure of (μ,θ) where $K<G$ weakly increases. □

Violence exhibits a *fixed–cost* threshold K; falsification (with convex c_f) is a *marginal* instrument. Hence, as verification becomes cheaper, the Sender's ex–post adjustments follow a predictable hierarchy: first increase falsification on the margins where upward gaps grow (Proposition 5.1); once the *integrated* benefit G exceeds K on a set of realized posteriors with sufficient probability, trigger violence.

5.3 Interaction with persuasion

Let $V^{\text{pol}}(\pi;F)$ denote the Sender's ex–ante value from persuasion alone (Section 4); let $V^{\text{pol}+f}$ and $V^{\text{pol}+f+V}$ denote the values when falsification and then violence are available ex post. The persuasion problem with these instruments simply replaces $v(\mu;F)$ by the appropriate continuation value:

$$V^{\text{pol}+X}(\pi;F) \;=\; \sup_{\Pi\in\mathcal{X}_{\text{EPIC}}(\pi;F,b,\varepsilon)} \mathbb{E}_\Pi\Big[v^X(\mu;F)\Big],$$

where $X \in \{\ f,\ \ f{+}V\ \}$.

Theorem 5.4. *Assume EPIC–implementable persuasion (Protocol B in Section 4.2). If falsification is upward–only or capacity–limited, then for any $\pi\in(0,1)$ and $F' \succeq_{\text{FOSD}} F$ the optimal posterior law under $X\in\{f,\ f{+}V\}$ Blackwell–coarsens as F improves. Moreover, the* *ex–ante probability of using falsification weakly increases, and—if K is below the threshold in Proposition 5.3 for some (μ, θ) used with positive probability—so does the ex–ante probability of violence.*

Proof sketch. With upward–only or capacity–limited falsification, the continuation value $v^X(\mu; F)$ becomes strictly more concave in μ on regions where the residual shortfall in $\theta = 0$ is positive. The feasible set of posteriors under Protocol B shifts right in $\theta = 0$ as F improves (Proposition 4.3), tightening the concavification at any fixed mean. Standard binary–support arguments then yield Blackwell coarsening. The usage claims follow from Propositions 5.1–5.3 after integrating over the equilibrium posterior distribution. □

Falsification smooths small gaps while persuasion shapes which gaps occur; violence covers large, recurrent gaps but only once a fixed–cost threshold is met. As verification becomes cheaper in the population, the Sender optimally reduces the precision of public information and shifts toward heavier ex–post instruments—first falsification, then violence.

6 Discussion and Applications

This section connects the theory to empirical settings, records robustness and modeling choices, and sketches policy and welfare implications. Throughout, we refer back to the primitives and EPIC implementability in Section 3, the concavification logic in Section 4, and the ex–post instruments in Section 5.

6.1 Applied interpretations and observables

A natural interpretation is propaganda in environments where a government communicates about performance, war progress, or economic conditions to a mass audience that can verify at heterogeneous costs using independent outlets, VPNs, open–source intelligence, or expert reports. In this reading, the public message corresponds to a distribution of posteriors, verification is the equilibrium cutoff response to that message, and ex–post manipulation takes the form of falsification or violence. The central prediction is that when verification becomes cheaper in the population (a first–order stochastic decrease of the cost distribution), the optimal public signal becomes less precise. Confusion is not a by–product but a deliberate equilibrium response to a more disciplined audience.

The same logic travels to public health and consumer safety. Health authorities or political actors communicate about epidemic risk while citizens can verify through expert channels, lab tests, or third–party dashboards. Firms release performance or safety statistics while consumers can pay to test or rely on professional reviews. Falsification here corresponds to review manipulation or synthetic engagement, while violence corresponds to discrete suppression technologies such as takedowns or delistings.

Several observables map to the primitives. Empirical proxies for the verification–cost distribution include adoption of VPNs, consumption of independent media, engagement with fact checks, data prices, outages, and local expertise density. Message precision can be proxied by dispersion or entropy in official communications, inconsistency rates across outlets, retraction frequencies, or the share of ambiguous frames. Falsification intensity may be proxied by bot and astroturf diagnostics or synthetic engagement, and repression by arrests, platform shutdowns, or event cancellations.

6.2 Testable predictions

The model yields a sequence of qualitative predictions.

First, when verification becomes cheaper, the optimal public experiment coarsens. In the benchmark of Section 4, the support of the posterior distribution contracts toward the prior; under the EPIC–implementable protocol with a minimal soft layer, the low support point shifts right, reducing spread in a Blackwell sense.

Second, cheaper verification increases the use of falsification in precisely those states and posteriors where the upward gap to the sender's target grows; once the integrated benefit of a discrete upward jump exceeds its fixed cost, violence appears, so the hazard of repression rises with cheaper verification.

Third, capacity limits on falsification strengthen the coarsening of persuasion: when falsification cannot cover large residual shortfalls, the sender anticipates those regions and reduces the likelihood of landing there by supplying a coarser public signal.

Fourth, policy shocks that exogenously reduce verification costs—fact–checking rollouts, censorship breaks, or data–price collapses—should be followed by noisier public messages, higher falsification where upward gaps expand, and, where fixed costs are low, discrete onsets of repression.

Finally, cross–sectional heterogeneity matters: groups with cheaper verification should experience both more coarsened messaging and higher marginal use of falsification relative to high–cost groups; reverse shocks have the opposite pattern.

6.3 Robustness and modeling choices

The sender objective was specified as a quadratic distance between the aggregate action and a state–shifted target to connect cleanly to your original file. The comparative static relies only on two properties: verification selectively attenuates the posterior–risk term in the sender's indirect value, and improvements in the cost distribution make that value more concave in belief. These properties extend to any objective that is convex in the aggregate with a state–dependent target and that loads positively on posterior risk; maximizing the aggregate (e.g., a turnout objective) fits by embedding the action rule into the same indirect value.

Quadratic receiver loss makes the private value of verification equal to the posterior variance. With any strictly proper scoring rule or Bregman loss, the gain from verification is the reduction in Bayes risk, which is single–peaked in belief and preserves a monotone cutoff. Hence the verifying mass remains increasing under a mean–preserving improvement of costs, and the sender's indirect value becomes more concave when costs fall.

Protocol choices matter only to the extent they affect implementability and the convexity of the feasible posterior set. The minimal delivery friction in Section 3 prevents silence from becoming fully revealing under hard evidence. The coarsening result under EPIC requires only a minimal non–verifiable label alongside evidence so that posteriors can vary on the $\theta = 0$ branch while $\theta = 1$ is disclosed; any protocol that delivers a convex feasible set with a left endpoint that moves right as verification becomes cheaper yields the same conclusion.

Private messages would expand posterior heterogeneity but leave the concavification logic intact. If the sender can tailor noise privately, coarsening appears as reduced within–group informational content when verification becomes cheaper for that group. In dynamic settings with slow–moving costs, the sender's best response tracks the cost process: sustained declines in verification costs imply gradual coarsening and increasing reliance on falsification, punctuated by threshold spikes into violence when accumulated gaps make the fixed cost worthwhile. If the sender can also raise verification costs through censorship, the comparative statics reverse: sharper messages, lower falsification, and less violence.

6.4 Welfare and policy

Lower verification costs improve private accuracy but induce the sender to coarsen the public signal and to substitute into falsification and, when thresholds are met, violence. Policy that subsidizes fact–checking or access should therefore anticipate strategic responses and pair access with enforcement on falsification (bot detection, audit trails) and credible costs on repression.

Platform design can affect the sender's calculus by reducing the value of ambiguity (for example, through forced claim–review flows or friction on mass reposts), but if falsification capacity is unconstrained, enforcement on that margin must come first.

Because verification incentives are most sensitive near intermediate beliefs, marginal reductions in verification costs for swing populations yield the largest design effects; targeting subsidies or access to those segments can therefore generate outsized welfare gains net of the sender's response.

6.5 Limitations and next steps

The binary state is chosen for clarity. In a continuous state, verification remains a local risk–reduction device, the sender's indirect value becomes more concave as verification costs fall, and concavification again delivers coarsening. Heterogeneous priors can be handled by treating the experiment as a distribution over group–specific posteriors and applying the argument groupwise.

Alternative aggregation technologies, such as thresholds or nonlinear payoffs, change the exact formula for the indirect value but preserve the key feature that verification reduces posterior risk where it matters most; the predictions survive whenever the continuation value is convex in risk.

Network and coordination effects would likely amplify the importance of posterior dispersion and therefore strengthen the case for coarsening. On the empirical side, event–study designs around plausibly exogenous verification–cost shocks—platform rollouts, outages, censorship breaks, or data–price changes—can trace the predicted triple response: coarser public signals, higher falsification, and threshold increases in repression where fixed costs are low.

Taken together, the theory recommends reading public confusion not as noise but as a designed response to disciplined audiences. Policies that lower verification costs increase private accuracy yet push adversarial senders toward coarser signals and heavier ex–post instruments; effective interventions must therefore integrate access, enforcement against falsification, and credible costs on repression.

7 Conclusion

This paper develops a model of public persuasion in which a sender faces a large audience that can verify at heterogeneous costs and the sender must satisfy *ex–post truthfulness*. We cast the problem as information design with evidence and treat the public posterior as a random variable—the natural object when many receivers react to a single public signal. Two forces shape the sender's choice of experiment: verification selectively attenuates posterior risk, and the implementability constraint ties the sender's hands after the state is realized. Together they deliver a simple comparative static: when verification becomes cheaper in the population, the sender's indirect value becomes more concave in the posterior, and the optimal public signal coarsens. Confusion is, in this sense, a strategy.

We give a constructive implementability condition under hard evidence with a minimal delivery friction and show how it pins the interior posterior used in the unfavorable state. A small soft layer layered on evidence yields a convex feasible set of posteriors and restores a clean concavification logic under EPIC. In a useful benchmark with uniform verification costs, the model reproduces the threshold pattern that motivated our empirical reading: for sufficiently large bias, the silence posterior saturates and the verifying mass at silence pins at a quarter of the population.

We extend the analysis to two ex–post instruments. Falsification continuously distorts the observed aggregate at convex marginal cost; violence (repression) induces a discrete upward jump at a fixed cost. As verification costs fall, the sender first responds on the persuasion margin by supplying a less precise public signal, then substitutes toward falsification in precisely those regions where upward gaps expand, and finally—once fixed–cost thresholds are crossed—resorts to violence. This hierarchy lines up with contemporary accounts of information control.

The framework speaks to measurement and policy. On the measurement side, it suggests observable mappings from verification costs (access to independent information, fact–checking frictions) to messaging precision, falsification intensity, and repression hazards. On the policy side, lowering verification costs improves private accuracy but induces strategic responses; effective interventions therefore pair access with enforcement against falsification and credible costs on repression.

Several extensions are natural. A continuous state preserves the logic that cheaper verification steepens curvature in the sender's value and pushes toward coarser experiments. Private messages, heterogeneous priors, richer aggregation technologies, and dynamics can be folded in at modest additional cost; the central mechanism survives whenever verification reduces posterior risk where the sender's payoff is locally convex. Endogenizing verification costs and allowing joint design of censorship and propaganda would yield a fuller theory of information control. Empirically, quasi–experimental shocks to verification costs offer a way to trace the predicted triple response—coarser signals, more falsification, and threshold repression.

Taken together, the results recommend a simple organizing principle for settings in which senders face disciplined audiences: when verification becomes cheaper, precision is optimally sacrificed. Public confusion is not mere noise but the predictable outcome of implementable persuasion in the shadow of fact–checking.

A Proofs and Constructive Implementation

This appendix collects formal proofs and a constructive implementation of the EPIC–feasible public experiments used in the main text. Throughout, we maintain the notation of Sections 3–5: the state is $\theta \in \{0, 1\}$ with prior $\pi \in (0, 1)$; the Sender's bias is $b \geq 0$; Receiver i's loss is $(a_i - \theta)^2$ and her private verification cost c_i is drawn i.i.d. from a continuous cdf F on

$[0,\bar{c}]$, normalized so that F is evaluated at $\mu(1-\mu) \in [0, \frac{1}{4}]$. The verifying mass at public posterior μ is $\lambda(\mu; F) = F(\mu(1-\mu))$, the aggregate action is

$$A(\mu, \theta; F) = (1 - \lambda(\mu; F))\, \mu + \lambda(\mu; F)\, \theta,$$

and the Sender's per–posterior indirect value (before any ex–post instruments) is

$$v(\mu; F) = -\Big(b^2 + (1 - \lambda(\mu; F))^2\, \mu(1-\mu)\Big).$$

Ex–post truthfulness (EPIC) is imposed on *verifiable evidence* only. A vanishing "delivery friction" $\varepsilon \in (0,1)$ guarantees that silence occurs with small probability in any state and prevents full unraveling.

A.1. Receiver best responses and verifying mass

Lemma A.1. *At any public posterior $\mu \in [0,1]$, a Receiver who does not verify chooses $a_i = \mu$ and suffers expected loss $\mu(1-\mu)$. If she verifies, she learns θ and chooses $a_i = \theta$, pays c_i, and suffers loss c_i. Hence the unique optimal verification rule is a cutoff:*

$$\textit{verify} \iff c_i \le c^*(\mu) := \mu(1-\mu), \qquad \lambda(\mu; F) = F\Big(\mu(1-\mu)\Big).$$

Proof. Quadratic loss implies that the Bayes action absent verification is $a_i = \mu$ with Bayes risk $\mu(1-\mu)$. Verification yields risk 0 and cost c_i. The comparison is $\mu(1-\mu) \ge c_i$, which is monotone in c_i. Aggregating over i yields the verifying mass $\lambda(\mu; F)$. □

A.2. Properties of the indirect value and the benchmark design problem

Lemma A.2. *For any F, the function $v(\cdot; F)$ is continuous on $[0,1]$, symmetric around $\mu = \frac{1}{2}$, and single–peaked with (global) maximum at some $\mu^\star \in \{0, \frac{1}{2}, 1\}$. If $F' \succeq_{\text{FOSD}} F$, then for all $\mu \in (0,1)$,*

$$v(\mu; F') - v(\tfrac{1}{2}; F') \;\le\; v(\mu; F) - v(\tfrac{1}{2}; F),$$

with equality at $\mu \in \{0,1\}$.

Proof. Symmetry follows because $\mu(1-\mu) = (1-\mu)\mu$ and $F(\mu(1-\mu))$ depends on μ only via $\mu(1-\mu)$, which is symmetric around $\frac{1}{2}$. Since $x \mapsto (1-F(x))^2 x$ is weakly increasing on $[0, \frac{1}{4}]$ when F is a cdf, the penalty term $(1-\lambda(\mu;F))^2 \mu(1-\mu)$ is minimized at the endpoints and maximized near $\frac{1}{2}$, delivering single–peakedness. If $F' \succeq_{\text{FOSD}} F$ then $1 - \lambda(\mu; F') \le 1 - \lambda(\mu; F)$

for all μ, so $(1-\lambda(\mu;F'))^2\mu(1-\mu) \le (1-\lambda(\mu;F))^2\mu(1-\mu)$, with equality when $\mu(1-\mu)=0$ (i.e., $\mu \in \{0,1\}$). Subtracting the common constant $-b^2$ and comparing to the value at $\frac{1}{2}$ yields the stated contraction. □

Let $\operatorname{cav} v(\cdot;F)$ denote the concave envelope on $[0,1]$.

Proposition A.3. *In the unconstrained persuasion problem with prior $\pi \in (0,1)$,*

$$\sup_{\Pi:\ \mathbb{E}_\Pi[\mu]=\pi} \mathbb{E}_\Pi[v(\mu;F)] \;=\; (\operatorname{cav} v(\cdot;F))(\pi),$$

and an optimal experiment has support of size at most two, $\Pi^* = \alpha\,\delta_{\mu_H} + (1-\alpha)\,\delta_{\mu_L}$ with $\mu_L \le \pi \le \mu_H$ on an exposed face of $\operatorname{cav} v(\cdot;F)$.

Proof. This is the standard concavification characterization (one–dimensional posterior). Carathéodory's theorem implies support size at most two. □

Theorem A.4. *Fix $\pi \in (0,1)$ and let $F' \succeq_{\mathrm{FOSD}} F$. There exist optimal binary experiments*

$$\Pi^*_F = \alpha\,\delta_{\mu_H} + (1-\alpha)\,\delta_{\mu_L}, \qquad \Pi^*_{F'} = \alpha'\,\delta_{\mu'_H} + (1-\alpha')\,\delta_{\mu'_L},$$

with $\mu_L \le \mu'_L \le \pi \le \mu'_H \le \mu_H$. Hence $\Pi^*_{F'}$ is a mean–preserving contraction of Π^*_F and Blackwell less informative.

Proof. By Lemma A.2, $v(\cdot;F')$ is a pointwise contraction of $v(\cdot;F)$ around $\frac{1}{2}$. Supporting chords at π to $v(\cdot;F')$ touch at points weakly closer to π than those for $v(\cdot;F)$. Mean preservation pins the weights. For binary experiments with fixed mean, smaller spread corresponds to a Blackwell garbling. □

A.3. EPIC under hard evidence and the silence posterior

We consider Protocol A (hard evidence $e \in \{0,1\}$ plus silence). With delivery friction $\varepsilon \in (0,1)$, the probability of *silence* in state θ is $(1-\delta_\theta)+\varepsilon\delta_\theta$, where δ_θ is the disclosure probability in state θ. When evidence is disclosed, posteriors are $\mu=1$ after $e=1$ and $\mu=0$ after $e=0$. When silence occurs, the posterior is

$$\mu_s = \Pr(\theta=1 \mid \text{silence}) = \frac{(1-\delta_1+\varepsilon\delta_1)\,\pi}{(1-\delta_1+\varepsilon\delta_1)\,\pi + (1-\delta_0+\varepsilon\delta_0)\,(1-\pi)}. \tag{A.1}$$

Lemma A.5. *Let the Sender's loss be $(A-(\theta+b))^2$. Then: (i) the EPIC constraint in $\theta=1$ implies $\delta_1=1$ at any optimum; (ii) if $\delta_0 \in (0,1)$, the EPIC constraint in $\theta=0$ binds and pins μ_s by*

$$(1-\lambda(\mu_s;F))\,\mu_s = 2b. \tag{A.2}$$

Proof. In $\theta = 1$, truthful disclosure sets $\mu = 1$ and the aggregate is $A(1, 1; F) = (1 - \lambda(1; F)) \cdot 1 + \lambda(1; F) \cdot 1 = 1$, so the loss is $(1-(1+b))^2 = b^2$. Withholding induces silence and $\mu_s \in (0, 1)$, which yields $A(\mu_s, 1; F) = (1 - \lambda(\mu_s; F))\mu_s + \lambda(\mu_s; F)$ and loss $\big((1 - \lambda)\mu_s + \lambda - (1+b)\big)^2 \geq b^2$ with strict inequality whenever $\mu_s < 1$ (since $(1 - \lambda)\mu_s + \lambda \leq 1$). Hence $\delta_1 = 1$.

In $\theta = 0$, disclosure yields $\mu = 0$ and $A(0, 0; F) = 0$, loss b^2. Withholding yields loss $\big((1 - \lambda(\mu_s; F))\,\mu_s - b\big)^2$. Indifference ($\delta_0 \in (0, 1)$) gives $(1 - \lambda(\mu_s; F))\,\mu_s = 2b$. □

Proposition A.6. *Let $F' \succeq_{\text{FOSD}} F$. If $\mu_s(F)$ solves (A.2) for F, then $\mu_s(F') \geq \mu_s(F)$, with strict inequality when $b > 0$ and $F' \neq F$.*

Proof. Define $\phi(\mu; F) = (1 - \lambda(\mu; F))\mu = (1 - F(\mu(1 - \mu)))\mu$. If $F' \succeq_{\text{FOSD}} F$, then $\phi(\mu; F') \leq \phi(\mu; F)$ for all $\mu \in (0, 1)$. Since $\phi(\cdot; F)$ is strictly increasing on $[\frac{1}{2}, 1]$ (because μ increases and $1 - F(\mu(1 - \mu))$ weakly decreases), the unique solution to $\phi(\mu; F) = 2b$ moves weakly right as F improves. □

A.4. EPIC with a minimal soft layer and concavification

Protocol B augments hard evidence with a non–verifiable public label $m \in \{L, H\}$, chosen according to state–dependent probabilities (a, b) fixed ex ante: $a = \Pr[m = H \mid \theta = 1]$, $b = \Pr[m = H \mid \theta = 0]$. The Sender still faces EPIC on verifiable evidence as in Lemma A.5; the soft layer is a standard commitment device chosen before θ is realized.

When evidence arrives ($e \in \{0, 1\}$), it pins the posterior at $\mu \in \{0, 1\}$ in either protocol. When silence occurs, Bayes' rule with the delivery friction ε gives, for $x \in \{L, H\}$,

$$\Pr(\theta = 1 \mid \text{silence}, m = x) = \frac{\pi(1 - \delta_1 + \varepsilon\delta_1)\Pr(m = x \mid \theta = 1)}{\sum_{\theta' \in \{0,1\}} \Pr(\theta')\,(1 - \delta_{\theta'} + \varepsilon\delta_{\theta'})\,\Pr(m = x \mid \theta')}. \tag{A.3}$$

Because $\delta_1 = 1$ by Lemma A.5, the numerator has factor ε; by choosing δ_0 close to 1, the denominator acquires the same factor, and the posterior becomes sensitive to the ratio a/b. This allows us to generate interior posteriors on the $\theta = 0$ branch while keeping $\theta = 1$ at $\mu = 1$ whenever evidence is delivered.

Lemma A.7. *Under Protocol B, any posterior used with positive probability in state $\theta = 1$ must be $\mu = 1$. Any interior posterior $\mu \in (0, 1)$ used with positive probability in state $\theta = 0$ must satisfy*

$$(1 - \lambda(\mu; F))\,\mu = 2b.$$

Hence the set of EPIC–admissible interior posteriors under $\theta = 0$ equals the solution set of the above equation; in the uniform special case $F(x) = x$ there is a unique solution $\mu_s \in (\frac{1}{2}, 1]$.

Proof. The first claim follows as in Lemma A.5: because disclosure in $\theta = 1$ is strictly preferred to any silence–induced interior posterior, an interior posterior cannot be used with positive probability in $\theta = 1$. For $\theta = 0$, the no–regret condition is the same indifference as in Lemma A.5, since only hard evidence is constrained ex post. Uniqueness in the uniform case follows from strict monotonicity of $g(\mu) = (1 - \mu(1 - \mu))\mu$ on $[\frac{1}{2}, 1]$. □

Theorem A.8. *Fix $\pi \in (0, 1)$. Under Protocol B, the set of EPIC–implementable posterior laws with mean π is convex and contains all binary laws supported on $\{\mu_L, 1\}$ where μ_L solves $(1 - \lambda(\mu_L; F))\mu_L = 2b$. If $F' \succeq_{\text{FOSD}} F$, there exists an optimal law $\Pi^*_{F'}$ that is a mean–preserving contraction of an optimal law Π^*_F.*

Proof. Convexity follows from the ability to mix m across realizations of silence in $\theta = 0$ while $\theta = 1$ is pinned at $\mu = 1$ by evidence. The admissible μ_L is determined by Lemma A.7. The objective $v(\cdot; F)$ becomes more concave as F improves (Lemma A.2); binary support and a left endpoint that moves weakly right (Proposition A.6) imply a tighter supporting chord and a mean–preserving contraction of the optimal posterior law. □

A.5. Constructive implementation under EPIC

This subsection provides an explicit recipe to implement the binary law $\Pi^* = \alpha\, \delta_1 + (1-\alpha)\, \delta_{\mu_L}$ that arises in the main text. We treat Protocol A first to pin the silence posterior and then Protocol B to realize it on the $\theta = 0$ branch with a soft label.

A.5.1. Protocol A (hard evidence + silence)

Choose $\delta_1 = 1$. Let μ_L be a solution to $(1 - \lambda(\mu; F))\mu = 2b$ (Lemma A.5) and set $\mu_s := \mu_L$. Select any $\delta_0 \in (0, 1)$; the probability of silence in $\theta = 0$ is $s_0 := 1 - \delta_0 + \varepsilon\delta_0$, and in $\theta = 1$ it is $s_1 := \varepsilon$. The Bayes mean imposed by (A.1) is

$$\pi = (1 - \varepsilon) \cdot 1 + \Big(1 - (1 - \varepsilon)(1 + \delta_0)\Big)\mu_s,$$

which is a linear equation in δ_0. For any $\varepsilon \in (0, 1)$ and $\mu_s \in (0, 1)$ there exists $\delta_0 \in (0, 1)$ solving it (indeed, as δ_0 varies the r.h.s. spans an interval containing π). The resulting posterior law has support $\{0, \mu_s, 1\}$, with the mass at 0 proportional to $(1 - \varepsilon)\delta_0(1 - \pi)$; EPIC holds by Lemma A.5.

A.5.2. Protocol B (adding a minimal soft label)

We now collapse the interior support to a single μ_L on the $\theta = 0$ branch while keeping $\theta = 1$ at $\mu = 1$ whenever evidence arrives. Let $\delta_1 = 1$. Choose $\delta_0 \in (0, 1)$ *close to* 1 so that $s_0 := 1 - \delta_0 + \varepsilon\delta_0$ is of order ε. Fix $a_1 := \Pr[m = H \mid \theta = 1] = 1$ and choose

$a_0 := \Pr[m = H \mid \theta = 0] \in (0,1)$ to make the posterior after $(\text{silence}, m = H)$ equal to the target $\mu_L \in (\frac{1}{2}, 1)$:

$$\mu_L = \frac{\pi\, s_1\, a_1}{\pi\, s_1\, a_1 + (1-\pi)\, s_0\, a_0} \qquad \Longleftrightarrow \qquad a_0 = \frac{\pi s_1}{(1-\pi) s_0}\left(\frac{1}{\mu_L} - 1\right), \qquad s_1 = \varepsilon. \tag{A.4}$$

As $\delta_0 \to 1$ we have $s_0 \to \varepsilon$, and the right–hand side tends to $\frac{\pi}{1-\pi}\left(\frac{1}{\mu_L} - 1\right)$. For $\pi \le \mu_L \le 1$ this limit lies in $[0,1]$; hence, by continuity, there exists δ_0 sufficiently close to 1 such that $a_0 \in (0,1)$. Under this choice, all silence+$m = H$ events induce μ_L. Silence+$m = L$ can be made measure–zero by setting $\Pr[m = L \mid \theta = 0, \text{silence}] = 0$ (which is compatible with (A.4) since only $m = H$ is used). Bayes plausibility pins the mass α on $\mu = 1$ via the mean constraint $\alpha \cdot 1 + (1-\alpha)\mu_L = \pi$. EPIC holds since only $\mu = 1$ is used with positive probability in $\theta = 1$, and μ_L satisfies $(1 - \lambda(\mu_L; F))\mu_L = 2b$.

A.6. Extensions: falsification and violence

This subsection derives the properties used in Section 5. Let D denote the feasible set for the distortion d (either $\mathbb{R}$, $[0,\infty)$, or $[-\bar{d}, \bar{d}]$), and let c_f be convex, even, $c_f(0) = 0$.

A.6.1. Quadratic falsification cost

If $c_f(d) = \frac{\kappa}{2} d^2$, the statewise problem

$$\min_{d \in \mathbb{R}} \Big(A(\mu, \theta; F) + d - (\theta + b)\Big)^2 + \tfrac{\kappa}{2} d^2$$

has unique minimizer $d^* = \frac{2}{2+\kappa}\Big((\theta + b) - A(\mu, \theta; F)\Big)$ and minimized loss $\frac{\kappa}{2+\kappa}\Big((\theta + b) - A(\mu, \theta; F)\Big)^2$. Hence

$$v^f_{\mathbb{R}}(\mu; F, \kappa) = -\frac{\kappa}{2+\kappa}\Big(b^2 + (1 - \lambda(\mu; F))^2\, \mu(1-\mu)\Big),$$

which is a positive multiple of $v(\mu; F)$. The concavification and thus the optimal experiment are unchanged.

A.6.2. One–sided and capacity–limited falsification

When $D = [0, \infty)$ (upward–only manipulation), the optimal d^*_D is the projection of the unconstrained minimizer onto $[0, \infty)$. Writing $\Delta(\mu, \theta; F) := (\theta + b) - A(\mu, \theta; F)$, we have $d^*_D = \max\{0, \tilde{d}(\mu, \theta)\}$ where $\tilde{d}$ is monotone in Δ. Since $\partial_\lambda A(\mu, 0; F) = -\mu < 0$ and $\partial_\lambda A(\mu, 1; F) = 1 - \mu > 0$, it follows that $\Delta(\mu, 0; F)$ is increasing and $\Delta(\mu, 1; F)$ is decreasing in $\lambda(\mu; F)$. As $F' \succeq_{\text{FOSD}} F$ raises $\lambda(\mu; F)$ pointwise, the probability and expected size of upward falsification weakly increase in $\theta = 0$ and weakly decrease in $\theta = 1$. Integrating over the equilibrium posterior law yields the monotonicity in Proposition 5.1 in the main text.

With a capacity $|d| \leq \bar{d}$, the minimized loss is bounded below by the squared residual shortfall

$$\Delta_{\bar{d}}(\mu, \theta; F) \;:=\; \max\Big\{0,\; \Big|(\theta + b) - A(\mu, \theta; F)\Big| - \bar{d}\Big\},$$

loss $\;\geq\; \Delta_{\bar{d}}(\mu, \theta; F)^2$.

As F improves, $\Delta_{\bar{d}}(\mu, 0; F)$ weakly increases for any fixed μ while $\Delta_{\bar{d}}(\mu, 1; F)$ weakly decreases, making the continuation value strictly more concave in μ on regions where $\Delta_{\bar{d}}(\mu, 0; F) > 0$. The corollary in Section 5.1 follows by the same concavification argument as in Theorem A.4.

A.6.3. Violence (fixed–cost upward jump)

Let $u \in \{0, 1\}$ denote the violence decision, with cost $K > 0$ when $u = 1$ and shift $A \mapsto A_\rho := \min\{A + \rho, 1\}$. Define the post–violence falsification–adjusted values

$$L^f(\mu, \theta) = \min_{d \in D} \Big(A(\mu, \theta; F) + d - (\theta + b)\Big)^2 + c_f(d),$$

$$L^{Vf}(\mu, \theta) = K + \min_{d \in D} \Big(A_\rho(\mu, \theta; F) + d - (\theta + b)\Big)^2 + c_f(d).$$

Violence is optimal iff $K < G(\mu, \theta; F, \rho, D)$ where

$$G(\mu, \theta; F, \rho, D) := \Big[\Big(A_\rho(\mu, \theta; F) - (\theta + b)\Big)^2\Big]^{\downarrow f} - \Big[\Big(A(\mu, \theta; F) - (\theta + b)\Big)^2\Big]^{\downarrow f},$$

and $x^{\downarrow f}$ denotes the value after optimal falsification subject to D. Since $\partial_\lambda A(\mu, 0; F) = -\mu < 0$, the pre–violence gap to the upward target b increases with λ in $\theta = 0$, while the gap to $(1 + b)$ in $\theta = 1$ decreases with λ. Therefore $G(\mu, 0; \cdot)$ is weakly increasing and $G(\mu, 1; \cdot)$ weakly decreasing in λ. If $F' \succeq_{\text{FOSD}} F$, then $\lambda(\mu; F') \geq \lambda(\mu; F)$ pointwise, and the ex–ante region where $K < G$ weakly expands, yielding Proposition 5.3.

A.7. Worked–out special case

Assume $\pi = \frac{1}{2}$, $\varepsilon \to 0$, and $F(x) = x$ on $[0, 1]$. Then $\lambda(\mu; F) = \mu(1 - \mu)$ and the EPIC condition (A.2) reduces to

$$(1 - \mu_s(1 - \mu_s))\mu_s = 2b \quad\Longleftrightarrow\quad \mu_s^3 - \mu_s^2 + \mu_s - 2b = 0,$$

with a unique solution $\mu_s(b) \in [\frac{1}{2}, 1]$. The left–hand side is maximized at $\frac{1}{2}$ where it equals $\frac{3}{8}$, so for $b \geq \frac{3}{16}$ the solution saturates at $\mu_s = \frac{1}{2}$. The verifying mass at silence is $\lambda(\mu_s) = \mu_s(1-\mu_s)$, which equals $\frac{1}{4}$ at the corner. This reproduces Proposition 3.1 in the text.

A.8. Continuous states

The binary state simplifies exposition. Suppose now $\theta \in [0, 1]$ with prior G and Receiver loss $(a - \theta)^2$. Verification reveals θ at cost c_i. For any public posterior distribution over θ with mean m, a non–verifier chooses $a = m$ and incurs Bayes risk $\mathbb{E}\Big[(\theta - m)^2 \mid m\Big] = \mathrm{Var}(\theta \mid m)$. The benefit of verification equals this posterior variance, yielding a cutoff $c_i \leq \mathrm{Var}(\theta \mid m)$ and a verifying mass $\lambda(m; F) = F(\mathrm{Var}(\theta \mid m))$. The aggregate action is $(1 - \lambda)m + \lambda\,\theta$, and the Sender's continuation value takes the form

$$v(m; F) = -\Big(b^2 + (1 - \lambda(m; F))^2\,\mathrm{Var}(\theta \mid m)\Big),$$

which is single–peaked and becomes more concave in m when F improves. Concavification and the EPIC construction (now with hard evidence about θ on *intervals* and a minimal soft layer) go through with obvious modifications; details are omitted for brevity.

Summary. The appendix established: (i) the cutoff verification rule and expression for $\lambda(\mu; F)$; (ii) the shape and comparative statics of the indirect value; (iii) concavification–based coarsening in the benchmark; (iv) EPIC implementability under hard evidence, including the closed–form indifference $(1 - \lambda(\mu_s; F))\mu_s = 2b$ and monotonicity of μ_s; (v) coarsening under EPIC with a minimal soft layer; (vi) a constructive implementation of the optimal binary experiment; and (vii) substitution patterns and thresholds for falsification and violence.

Acknowledgments

We thank Olivier Gossner, Johannes Hörner, Yukio Koriyama, Yves Le Yaoanq, Alessandro Riboni, François Salanié, Anna Sanktjohanser, Ludvig Sinander, Nikita Zakharov and all the participants of the PEDD 2023 Conference, the TSE Applied Theory Workshop and the seminars at CREST for useful comments and discussions. Mira Hajar provided excellent research assistance. Any errors are our own.

Statements and Declarations

Funding This research is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047).

Competing interests The authors declare that they have no competing interests.

Data availability This study is theoretical and does not analyze empirical datasets. All results are derived analytically. No data or code are required to reproduce the findings beyond the statements and proofs contained in the paper.

References

Arieli, I. and Babichenko, Y. (2019). Private Bayesian persuasion. *Journal of Economic Theory*, 182:185–217.

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2019). Mechanisms with evidence: Commitment and robustness. *Econometrica*, 87(2):529–566.

Bergemann, D. and Morris, S. (2016). Information design, bayesian persuasion, and Bayes correlated equilibrium. *American Economic Review: Papers & Proceedings*, 106(5):586–591.

Caillaud, B. and Tirole, J. (2007). Consensus building: How to persuade a group. *American Economic Review*, 97(5):1877–1900.

Doval, L. and Skreta, V. (2022). Mechanism design with limited commitment. *Econometrica*, 90(4):1463–1500.

Gehlbach, S., Luo, Z., Shirikov, A., and Vorobyev, D. (2021). A model of censorship, propaganda, and repression. Working paper.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483.

Guriev, S. and Treisman, D. (2019). Informational autocrats. *Journal of Economic Perspectives*, 33(4):100–127.

Guriev, S. and Treisman, D. (2020). A theory of informational autocracy. *Journal of Public Economics*, 186:104–114.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., and Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964.

Matysková, L. and Montes, A. (2023). Bayesian persuasion with costly information acquisition. *Journal of Economic Theory*, 211:105678.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391.

Morris, S. and Shin, H. S. (2002). Social value of public information. *American Economic Review*, 92(5):1521–1534.

Seidmann, D. J. and Winter, E. (1997). Strategic information transmission with verifiable messages. *Econometrica*, 65(1):163–170.

Shin, H. S. (1994). News management and the value of firms. *RAND Journal of Economics*, 25(1):58–71.

Shin, H. S. (2003). Disclosures and asset returns. *Econometrica*, 71(1):105–133.

Skreta, V. and Pérez-Richet, E. (2022). Test design under falsification. *Econometrica*, 90(3):1109–1142.

Titova, M. (2022). Persuasion with verifiable information. Working paper.

Yang, L. L. (2024). Information design with costly state verification. Discussion Paper 502, CRC TR 224, University of Bonn and University of Cologne.