WORKING PAPERS

N° 1681

October 2025

“False Cascades and the Cost of Truth”

Darina Cheredina and Georgy Lukyanov

Toulouse School of Economics

False Cascades and the Cost of Truth*

Darina Cheredina† Georgy Lukyanov‡

Abstract

We study sequential social learning when agents can sometimes pay to verify a claim and obtain hard, publicly checkable evidence. Each agent observes the public history, receives a private signal, may investigate at a cost (succeeding only when the claim is true), and can disclose or conceal any proof. Actions are binary or continuous, with a conformity pull toward the prevailing consensus. We characterize when false cascades persist and when societies self-correct. In the binary benchmark, we derive an investigation cutoff and show how its location relative to classic cascade bands governs breakability; a simple knife-edge condition guarantees that any wrong cascade at the boundary is overturned with positive probability. With continuous actions, coarse observation and conformity can recreate cascades, yet occasional disclosures collapse them. These forces yield a tractable "resilience frontier" with transparent comparative statics and policy levers.

Keywords: social learning; informational cascades; verification; misinformation; conformity; disclosure.
JEL: D83; D85; C73; C72.

1 Introduction

A research claim gains traction as papers cite one another, seminar audiences nod along, and policy briefs echo the finding. Then a small replication team requests data, reruns the code, and applies simple forensic checks. The result is a public package—archived data, scripts, and a short report—that anyone can audit. Beliefs shift quickly: what had looked like a robust result is withdrawn or retracted once verifiable proof accumulates. This paper builds a framework for that kind of turn. Soft information can stack into a cascade, but costly verification sometimes generates hard, public evidence that can overturn it.

Classic models show how sequential, rational actors can ignore their own signals and herd on observed behavior [2, 6]. Subsequent work clarifies when herding persists and how beliefs move [8, 14]. We enrich this environment with an endogenous verification decision. Each agent sees

*We thank Evgeniy Andreev, Emiliano Catonini, Markus Gebauer, Ella Khromova, Steven Kivinen, and Alexey Verenikin for helpful comments and discussions, as well as participants of the ICEF research seminar for valuable feedback. All remaining errors are our own.

†HSE University, International College of Economics and Finance (ICEF). Email: dicheredina@edu.hse.ru.
‡Toulouse School of Economics. Email: Georgy.Lukyanov-fr.eu.

the public history, draws a private signal about a binary state, and chooses whether to pay for an investigation that yields verifiable evidence only when the state is truly present. If proof arrives, the agent may disclose it and receive a benefit; if not, actions remain the only soft signal. Continuous actions are interpreted as expressed opinions disciplined by a quadratic pull toward the consensus, capturing conformity pressures in public discourse.

Two forces drive our results. The first is a soft channel: actions convey private information when they are responsive and observable. The second is a hard channel: occasional disclosures create public, verifiable proof and coordinate everyone on the truth. We use these to formalize *misinformation resilience*—the conditions under which wrong cascades are eventually broken by decentralized behavior. Our first main result characterizes the optimal verification rule: there is a cutoff in the posterior above which agents choose to investigate (Proposition 4.4). We then show how this cutoff intersects the classic cascade bands to determine when cascades are breakable (Proposition 4.5); a simple knife-edge condition at the boundary yields an immediate resilience benchmark (Corollary 4.6). In the continuous-action extension, we derive a responsiveness threshold under coarse observation that explains when actions again become uninformative (Proposition 5.1). Finally, we provide boundary results that pin down local resilience and assemble them into a compact resilience frontier summarizing both channels (Theorem 6.4; Corollary 6.5). Returning to the replication vignette, a single verifiable release—data and code that pass public checks—can shift beliefs even after long stretches in which soft signals mostly aligned with the prevailing view.

2 Related literature

We build on the sequential social-learning tradition where public actions can swamp private information and generate cascades [2, 6]. Recent treatments sharpen when learning succeeds or stalls, the geometry of posteriors, and the speed of correction [1, 11]. Fragility with heterogeneous types and slight misperceptions is now well documented [10], and network structure can impede diffusion even under connectivity [12]. Surveys synthesize these advances and open questions [5].

Our "hard-evidence" channel relates to verifiable disclosure: once proof exists, beliefs unravel toward the truth [4]. Here proof is *endogenously* produced through costly investigation and arrives only when the claim is true. This connects to work on disclosure in sequential environments and strategic revelation [3]. With continuous actions we allow conformity pull, in the spirit of social image concerns [7], and we study coarse observation of actions. Coarsening is increasingly central in recent models where agents learn from action-signals rather than raw actions [15, 16]. Our threshold for soft-channel informativeness complements those papers by tying responsiveness to coarse observation explicitly.

The broader motivation intersects with misinformation and verification. Large, multi-country experiments show that fact-checking reduces misperceptions on average, while design and framing matter for engagement [9, 13]. We abstract from psychology to isolate the equilibrium forces that make false cascades resilient, and we ask when decentralized verification suffices to restore truth.

We proceed as follows. Section 3 sets up the environment, signals, verification technology, disclosure, and equilibrium. Section 4 analyzes the binary benchmark, derives the investigation cutoff, and studies its interaction with cascade bands. Section 5 turns to continuous actions and conformity, showing how coarse observation can reintroduce cascades and how disclosures collapse them. Section 6 defines misinformation resilience, proves boundary breakability conditions, and develops the resilience frontier. Section 7 sketches welfare and policy levers, and Section 8 collects extensions (heterogeneous costs and benefits, strategic concealment, networks). Section 9 concludes.

3 Model

We consider an infinite sequence of agents $t = 1, 2, \ldots$ who act once in order. There is a binary state $\theta \in \{0, 1\}$ ("event present" if $\theta = 1$). Let $\mu_1 \in (0, 1)$ denote the common prior $\Pr(\theta = 1)$ and μ_t the public belief at the start of period t, formed from the public history. Agents are risk–neutral and do not discount; payoffs are period–by–period.

Timeline and observables. At the beginning of t, agent t observes the public history H_t, which includes all past actions and any *disclosed* hard evidence (defined below). The agent then receives a private signal $s_t \in \{0, 1\}$, draws an investigation decision $i_t \in \{0, 1\}$ (pay a cost if $i_t = 1$), possibly obtains hard evidence, chooses whether to disclose it, and finally chooses an action a_t. The public history H_{t+1} records (a_t, disclosed evidence at t).

Private signals. Signals are i.i.d. conditional on θ with precision $q \in (1/2, 1)$:

$$\Pr(s_t = 1 \mid \theta = 1) = \Pr(s_t = 0 \mid \theta = 0) = q,$$
$$\Pr(s_t = 1 \mid \theta = 0) = \Pr(s_t = 0 \mid \theta = 1) = 1 - q.$$

Let $\mu_t^+ \equiv \Pr(\theta = 1 \mid s_t = 1, H_t)$ and $\mu_t^- \equiv \Pr(\theta = 1 \mid s_t = 0, H_t)$ denote the signal–updated posteriors (Bayes' rule).

Investigation technology. Agent t may pay cost $c \geq 0$ to investigate ($i_t = 1$). Investigation yields *verifiable hard evidence* E with probability $p \in (0, 1]$ if and only if $\theta = 1$:

$$\Pr(E \mid \theta = 1, i_t = 1) = p,$$
$$\Pr(E \mid \theta = 0, i_t = 1) = 0.$$

When no hard evidence is realized, the investigation produces no public object. There are thus no false positives; false negatives occur when $\theta = 1$ but E fails to arrive. If E arrives, the investigator privately observes it and then chooses whether to disclose.

Disclosure. Disclosure is a binary choice $d_t \in \{0, 1\}$ available only if E arrived. If $d_t = 1$, the public history records a verifiable disclosure at t and all future agents learn $\theta = 1$. If $d_t = 0$,

no public trace of evidence is left (the action a_t remains observable). There is no possibility to fabricate E.

Actions. The baseline action space is binary, $a_t \in \{0, 1\}$, interpreted as endorsement vs. rejection of the event. (Section 5 allows continuous opinions $a_t \in [0, 1]$ with a conformity motive.) In the baseline, absent disclosure, agents care about choosing the action that matches the state.

Payoffs. Agent t's payoff is

$$u_t \;=\; \underbrace{\mathbf{1}\{a_t = \theta\}}_{\text{accuracy}} \;+\; \underbrace{V\,\mathbf{1}\{E \text{ arrived and } d_t = 1\}}_{\text{disclosure benefit}} \;-\; \underbrace{c\,i_t}_{\text{investigation cost}},$$

where $V \geq 0$ captures whistleblower rents, reputational or prize benefits from bringing verifiable truth to light.[1]

Beliefs and updating. The public belief $\mu_t = \Pr(\theta = 1 \mid H_t)$ is common knowledge at the start of t. Private Bayesian updating from signals yields

$$\mu_t^+ = \frac{\mu_t q}{\mu_t q + (1 - \mu_t)(1 - q)}, \qquad \mu_t^- = \frac{\mu_t (1 - q)}{\mu_t (1 - q) + (1 - \mu_t) q}.$$

If E is *disclosed* at any date, beliefs jump to $\mu_{t+1} = 1$ permanently. Otherwise beliefs evolve from observed actions via Bayes' rule under equilibrium strategies (defined below). In the binary–action benchmark without investigation, the standard cascade bands are $\mu \leq 1 - q$ ("lower") and $\mu \geq q$ ("upper"), within which optimal actions ignore private signals. In our environment, these bands still describe behavior in regions where agents optimally *do not* investigate.

Strategies and equilibrium. A (pure) strategy for agent t maps the public belief and private information into choices

$$\sigma_t : \; (\mu_t, s_t, E) \;\mapsto\; (i_t, d_t, a_t),$$

where d_t is relevant only if E is realized. The public history H_t consists of past actions and any past disclosures. An equilibrium is a Perfect Bayesian Equilibrium (PBE): (i) each agent's strategy is sequentially optimal given beliefs; (ii) beliefs are updated from the public history by Bayes' rule wherever possible and by standard consistency off the equilibrium path.

Investigation cutoff. In the binary–action benchmark without conformity, there exists an investigation threshold $\mu^* = \mu^*(q, p, c, V) \in (0, 1)$ such that, holding fixed the anticipated mapping from actions to beliefs, an agent with posterior μ weakly prefers to investigate whenever $\mu \geq \mu^*$, and not otherwise. Intuitively, higher signal precision q and better verification technology p reduce μ^*,

[1] If desired, a conformity term can be introduced later as $-\lambda(a_t - \bar{a}_t)^2$, where $\bar{a}_t$ is the publicly inferred consensus; we defer this to Section 5.

Table 1: Notation

Symbol	Meaning
$\theta \in \{0,1\}$	State.
$\mu_t \in (0,1)$	Public belief at the start of period t.
H_t	Public history.
$s_t \in \{0,1\}$	Private signal at t.
$q \in (1/2,1)$	Signal precision.
$i_t \in \{0,1\}$	Investigation decision.
$c \geq 0$	Investigation cost.
E	Hard, verifiable evidence.
$p \in (0,1]$	Success probability of investigation given $\theta = 1$.
$d_t \in \{0,1\}$	Disclosure decision when E arrives.
$V \geq 0$	Private benefit/rent from disclosing hard evidence.
$a_t \in \{0,1\}$ or $[0,1]$	Action: binary (baseline) or continuous opinion.
$\gamma \in [0,1]$	Accuracy weight in continuous best response.
$h \in [0,1]$	Observational granularity.
$\mu_t^+,\ \mu_t^-$	Posteriors after $s_t = 1$ or $s_t = 0$.
x	Agent's posterior given (H_t, s_t).
x'	Posterior after a failed investigation.
$A(x)$	Accuracy payoff without investigation.
$x_1^*,\ x_2^*$	Investigation cutoffs.
$\underline{\mu}, \overline{\mu}$	Cascade boundaries: $\underline{\mu} = 1-q$, $\overline{\mu} = q$.
$L(\mu)$	Belief likelihood ratio.
$\Delta x(\mu)$	Posterior gap under signals.

while higher costs c raise it and higher discovery benefits V reduce it. Section 4 derives μ^* formally and studies how its position relative to the cascade bands determines whether wrong cascades can be broken.

Two features are central. First, hard evidence is asymmetric (no false positives), so a disclosure is an absorbing state at truth. Second, disclosure is endogenous: even when E arrives, an agent may conceal it, trading off V against any strategic considerations embedded in the action choice. These forces interact with the usual action–based learning to determine when societies become stuck in false cascades and when sporadic investigation suffices for correction.

Whenever an agent is (weakly) indifferent at a knife-edge belief (e.g., at $\mu \in \{1-q, q\}$ in the binary benchmark or when $\Delta a(\mu) = h$ in the continuous/coarsened case), we select the equilibrium that (i) breaks ties *in favor of investigation* and (ii) if hard evidence is obtained, *in favor of disclosure*. In the continuous-action case, ties in the best response are resolved toward $a_t^* = \gamma x_t + (1-\gamma)\mu_t$.[2]

With binary actions and endogenous investigation, the symmetric monotone PBE is unique away from knife-edges. Multiplicity can arise only at measure-zero parameter sets where the investigation

[2]Equivalently, add vanishing trembles: an $\varepsilon \downarrow 0$ private payoff perturbation that makes investigation strictly optimal on the knife-edge, and (separately) a vanishing disclosure bonus. All results—Proposition 4.4, Proposition 4.5, Theorem 6.4, Proposition 5.1—are robust to such trembles.

and no-investigation payoffs coincide exactly at the cascade boundaries; our selection above pins down a canonical equilibrium in those cases. None of the comparative-statics or resilience results depend on which knife-edge selection is used.

4 Discrete-Action Analysis

This section solves the one-shot problem of an agent who has observed the public belief μ_t and a private signal s_t, yielding a posterior $x \in [0,1]$ about $\theta = 1$. We derive the value of investigating after s_t is observed, show a single–crossing property that delivers a cutoff rule, and then translate the decision back into public-belief space to study when cascades are breakable. Throughout this section actions are binary ($a_t \in \{0,1\}$) and there is no conformity term.

Let x denote the agent's posterior $\Pr(\theta = 1 \mid H_t, s_t)$ (so $x = \mu_t^+$ if $s_t = 1$ and $x = \mu_t^-$ if $s_t = 0$). If the agent *does not* investigate, the optimal action is $a = 1$ iff $x \geq 1/2$, yielding value

$$U^{\text{no}}(x) = \max\{x, 1-x\} \equiv A(x).$$

If the agent *does* investigate, then with probability xp hard evidence E arrives (only when $\theta = 1$), the agent discloses it, obtains the benefit V, chooses $a = 1$ and gets accuracy 1. With the complementary probability $1-xp$, no hard evidence arrives; given this failure, the posterior updates to

$$x' \equiv \Pr(\theta = 1 \mid \text{failure}, H_t, s_t) = \frac{x(1-p)}{x(1-p) + (1-x)\cdot 1} = \frac{x(1-p)}{1-xp},$$

after which the agent optimally chooses the action that maximizes accuracy, giving value $A(x')$. Hence

$$U^{\text{inv}}(x) \;=\; xp\,(1+V) + (1-xp)\,A(x') - c. \tag{1}$$

Lemma 4.1. *For $x \geq \frac{1}{2}$,*

$$U^{\text{inv}}(x) - U^{\text{no}}(x) = \begin{cases} x\,p\,V - c, & \text{if } x \geq \frac{1}{2-p}, \\ 1 - c + x\big(p(1+V) - 2\big), & \text{if } \frac{1}{2} \leq x < \frac{1}{2-p}. \end{cases}$$

By symmetry, for $x \leq \frac{1}{2}$ replace x with $1-x$ in the right-hand side. In each half-interval $[\frac{1}{2}, 1]$ and $[0, \frac{1}{2}]$ the difference $U^{\text{inv}} - U^{\text{no}}$ is affine in x and hence exhibits a single crossing.

Proof. When $x \geq \frac{1}{2}$, $U^{\text{no}}(x) = x$. If $x' \geq \frac{1}{2}$ (equivalently $x \geq 1/(2-p)$), then $A(x') = x'$ and

$$(1-xp)A(x') = (1-xp)\frac{x(1-p)}{1-xp} = x(1-p),$$

so (1) gives $U^{\text{inv}}(x) - U^{\text{no}}(x) = xp(1+V) + x(1-p) - c - x = xpV - c$. If instead $x' < \frac{1}{2}$ (equivalently

$x < 1/(2-p)$), then $A(x') = 1 - x'$, and

$$(1 - xp)A(x') = (1 - xp)\left(1 - \frac{x(1-p)}{1-xp}\right) = 1 - x,$$

so (1) yields $U^{\text{inv}}(x) - U^{\text{no}}(x) = xp(1+V) + 1 - x - c - x = 1 - c + x\big(p(1+V) - 2\big)$. The symmetry claim follows by exchanging labels of states and actions, which maps x to $1 - x$. □

Lemma 4.1 implies a threshold rule within each half of the unit interval.

Lemma 4.2. *Fix $q \in (1/2, 1)$. The one–signal posteriors $\mu^+(\mu, q)$ and $\mu^-(\mu, q)$ are continuous and strictly increasing in $\mu \in (0, 1)$. Moreover, for all $\mu \in (0, 1)$ one has $\mu^-(\mu, q) < \mu < \mu^+(\mu, q)$.*

Proof. Let $O = \mu/(1-\mu)$ and $R = q/(1-q) > 1$. Bayes gives $\mu^+ = \frac{OR}{1+OR}$ and $\mu^- = \frac{O}{O+R}$. Both expressions are strictly increasing in O, and O is strictly increasing in μ, yielding the claim. Since $R > 1$, we have $\mu^- < \mu < \mu^+$ for all $\mu \in (0, 1)$. □

Lemma 4.3. *Fix (p, c, V) with $p \in (0, 1]$, $c \geq 0$, $V \geq 0$. In the binary benchmark, the net gain from investigating versus not investigating, $G(x) \equiv U^{\text{inv}}(x) - U^{\text{no}}(x)$, is continuous and strictly increasing in the posterior x on $[1/2, 1)$ and satisfies $G(1 - x) = -G(x)$. Hence there exists a unique cutoff $x^\star \in [1/2, 1)$ such that an agent investigates iff $x \geq x^\star$.*

Proof. Continuity is immediate from the payoff definitions. Monotonicity follows from MLRP and the structure of investigation: evidence arrives only in the true state, so the likelihood ratio of "investigation outcomes" (disclosure vs. no disclosure) is increasing in x, which increases the expected marginal value of investigating. Symmetry around $1/2$ holds because swapping labels of states maps posteriors $x \mapsto 1 - x$ and flips the sign of the relative value. Strict increase on $[1/2, 1)$ yields uniqueness of the cutoff by the intermediate value theorem. □

Proposition 4.4. *There exists a (weakly) increasing cutoff policy with respect to the posterior. Specifically:*

- *If $x \geq \frac{1}{2}$ and $x \geq \frac{1}{2-p}$, then the agent investigates iff*

$$x \;\geq\; x_1^* \;\equiv\; \frac{c}{pV}.$$

- *If $x \geq \frac{1}{2}$ and $x < \frac{1}{2-p}$, then the agent investigates iff*

$$x \;\geq\; x_2^* \;\equiv\; \frac{1-c}{2-p(1+V)}.$$

- *For $x \leq \frac{1}{2}$ the rule is symmetric: replace x by $1 - x$ in the conditions above.*

Moreover, x_1^* is decreasing in p and V and increasing in c; x_2^* is decreasing in p and V and increasing in c whenever $2 - p(1+V) > 0$.

Proof. Immediate from Lemma 4.1: in each region $U^{\text{inv}} - U^{\text{no}}$ is affine in x, so the indifference point solves a linear equation. Comparative statics follow by inspection. □

Two remarks aid interpretation. First, the "clean" case $x \geq 1/(2-p)$ is economically natural: a failed investigation weakens conviction but does not flip the optimal action; here the cutoff is especially transparent, $x \geq c/(pV)$. Second, near the knife-edge $x \simeq 1/2$ a failed investigation can reverse the chosen action; the effective cutoff adjusts to x_2^*, which nests the knife-edge condition $U^{\text{inv}}(1/2) \geq U^{\text{no}}(1/2)$, i.e.

$$\frac{1}{2}\, p(1+V) \;\geq\; c. \tag{2}$$

The posterior x is $x = \mu_t^+$ after a positive signal and $x = \mu_t^-$ after a negative signal, where

$$\mu_t^+ = \frac{\mu_t q}{\mu_t q + (1-\mu_t)(1-q)}, \qquad \mu_t^- = \frac{\mu_t (1-q)}{\mu_t (1-q) + (1-\mu_t) q}.$$

Thus the investigation rule can be written as likelihood–ratio thresholds in μ_t conditional on the observed s_t. For example, in the clean region after $s_t = 1$ the threshold $x \geq c/(pV)$ is equivalent to

$$\frac{\mu_t}{1-\mu_t} \;\geq\; \underbrace{\frac{c}{pV-c}}_{\text{evidence rent ratio}} \cdot \underbrace{\frac{1-q}{q}}_{\text{signal LR}^{-1}},$$

and after $s_t = 0$ the symmetric threshold uses $\mu_t^- \leq 1 - c/(pV)$.

In the classic binary–action benchmark without investigation, agents ignore their signals and herd when $\mu_t \leq 1-q$ (lower cascade) or $\mu_t \geq q$ (upper cascade). In those regions the next agent's posteriors lie on one side of $1/2$: if $\mu_t \leq 1-q$ then $\mu_t^+ \leq 1/2$ and $\mu_t^- \leq 1/2$; if $\mu_t \geq q$ then $\mu_t^- \geq 1/2$ and $\mu_t^+ \geq 1/2$. Whether such a cascade is *breakable* depends on whether the relevant posterior crosses the investigation cutoff.

Proposition 4.5. *Fix parameters (q, p, c, V).*

1. *A lower cascade at belief $\mu_t \leq 1-q$ is breakable (i.e., some agent strictly prefers to investigate with positive probability) iff*

$$U^{\text{inv}}(\mu_t^+) \;\geq\; U^{\text{no}}(\mu_t^+).$$

   Equivalently, using Proposition 4.4, either $\mu_t^+ \geq \max\{x_1^*, \frac{1}{2-p}\}$, or $\mu_t^+ \in [\frac{1}{2}, \frac{1}{2-p})$ and $\mu_t^+ \geq x_2^*$. (By monotonicity, if the condition fails at μ_t^+ it fails at μ_t^- as well.)

2. *An upper cascade at belief $\mu_t \geq q$ is breakable iff the symmetric condition holds with x replaced by $1 - \mu_t^-$.*

Proof. In a lower cascade, $s_t = 1$ maximizes the posterior among the two signals. Since the decision to investigate is increasing in x (Lemma 4.1), investigation occurs with positive probability iff it occurs at $x = \mu_t^+$. The upper-cascade case is symmetric. □

A particularly transparent benchmark arises at the cascade boundary $\mu_t = 1 - q$, where $\mu_t^+ = 1/2$. Plugging $x = 1/2$ into Lemma 4.1 yields the following corollary.

Corollary 4.6. *If*

$$\frac{1}{2}\,p(1+V) \;\geq\; c,$$

then any cascade at the boundary is breakable: an agent who receives a positive signal (at the lower boundary) or a negative signal (at the upper boundary) strictly prefers to investigate. Consequently, any wrong cascade is overturned with positive probability in finite time.

Proposition 4.4 implies that the incentives to investigate strengthen when p or V rise and weaken when c rises; higher signal precision q raises μ_t^+ at any given μ_t, moving the economy toward investigation after favorable signals. In terms of cascades, wrong lower cascades are easier to break when (p, V) are high, c is low, and q is high; the symmetric statements hold for wrong upper cascades. Define the *misinformation-resilience region* as the set of parameters (q, p, c, V) for which both inequalities in Proposition 4.5 hold at the respective cascade boundaries $\mu = 1 - q$ and $\mu = q$. By Corollary 4.6, the simple sufficient condition $\frac{1}{2}p(1+V) \geq c$ guarantees resilience at both boundaries; outside that baseline, resilience obtains whenever μ^+ (or $1 - \mu^-$) crosses the relevant cutoff in Proposition 4.4, which occurs for sufficiently large q even when $p(1+V)$ is moderate.

Taken together, these results formalize how sporadic verifiable evidence interacts with action-based learning. If investigation rents are sufficiently attractive relative to costs, wrong cascades are fragile: they are broken the first time a moderately favorable signal arrives. When rents are small or costs high, cascades become absorbing unless the signal and verification technologies are strong enough to push posteriors above the relevant thresholds.

5 Continuous Actions and Conformity

We now allow actions to be continuous opinions $a_t \in [0, 1]$ and introduce a conformity motive. Agent t's posterior about $\theta = 1$ after observing H_t and s_t is $x \in [0, 1]$. The agent chooses a to balance accuracy and alignment with the current consensus μ_t:

$$\min_{a\in[0,1]} \underbrace{\gamma\,\mathbb{E}\Big[(a-\theta)^2 \mid x\Big]}_{\text{truth}} \;+\; \underbrace{(1-\gamma)\,(a-\mu_t)^2}_{\text{conformity}}, \qquad \gamma \in [0,1].$$

With $\theta \in \{0, 1\}$ and $\Pr(\theta{=}1 \mid H_t, s_t){=}x$, we have $\mathbb{E}[(a-\theta)^2 \mid x] = x(a-1)^2 + (1-x)a^2$. The objective is strictly convex and yields a closed-form best response.

The first-order condition gives

$$a^*(x, \mu_t) = \gamma\, x + (1-\gamma)\,\mu_t, \tag{3}$$

i.e., a convex combination of posterior truth and consensus. The *responsiveness* to private information is $\partial a^*/\partial x = \gamma$. When $\gamma = 0$ actions are fully conformist ($a^* = \mu_t$) and reveal nothing; when

$\gamma = 1$ actions are truthful ($a^* = x$).

If strategies are common knowledge and a^* is observed without noise, then for any $\gamma > 0$ the map $x \mapsto a^*$ is strictly increasing and invertible:

$$x = \mu_t + \frac{a^* - \mu_t}{\gamma}.$$

Hence the period-t action fully reveals the posterior x, and soft information aggregates efficiently across time. In that frictionless case, action-based herding disappears: the belief process coincides with aggregation of posteriors and does not collapse to a region where actions ignore signals.

Two forces can reintroduce cascades. First, at the classical cascade boundaries the posterior gap induced by one signal vanishes. Let

$$O_t \equiv \frac{\mu_t}{1-\mu_t} \quad \text{and} \quad R \equiv \frac{q}{1-q}.$$

Bayes' rule gives

$$\mu_t^+ \;=\; \frac{O_t R}{1 + O_t R} \qquad \text{and} \qquad \mu_t^- \;=\; \frac{O_t/R}{1 + O_t/R} \;=\; \frac{O_t}{O_t + R}.$$

Hence the one-period posterior gap is

$$\Delta x(\mu_t, q) \;\equiv\; \mu_t^+ - \mu_t^- \;=\; \frac{\mu_t(\mu_t - 1)\,(2q-1)}{(2\mu_t q - \mu_t - q)\,(2\mu_t q - \mu_t - q + 1)}. \tag{4}$$

For brevity we write $\Delta x(\mu)$ when q is fixed.

At the classical cascade boundaries $\mu_t \in \{1-q,\, q\}$ we have $\Delta x(\mu_t, q) = 0$; absent hard evidence, the soft channel thus shuts down locally.

Suppose the public sees a coarsened action $\tilde{a} = \mathcal{C}(a^*)$ where $\mathcal{C}$ rounds to a grid of step $h \in (0, 1]$. A sufficient statistic for signal revelation at belief μ is whether the two signal-contingent actions straddle a grid boundary. Using (3), the signal-induced separation in actions equals

$$\Delta a(\mu) \;\equiv\; a^*(\mu^+, \mu) - a^*(\mu^-, \mu) \;=\; \gamma\, \Delta x(\mu).$$

Hence signals are *distinguishable* from actions at belief μ if $\Delta a(\mu) \geq h$, and *indistinguishable* if $\Delta a(\mu) < h$.

Proposition 5.1. *Fix (q, γ) and grid step $h \in (0, 1]$. At belief μ, signals are inferable from actions if and only if*

$$\gamma \;\geq\; \frac{h}{\Delta x(\mu, q)} \qquad \textit{where } \Delta x(\mu, q) \textit{ is given by (4).}$$

Suppose the public sees a coarsened action $\tilde{a} = \mathcal{C}(a^*)$ obtained by rounding a^* to the nearest

grid point of step h. From (3), the signal-induced separation in actions equals

$$\Delta a(\mu) \;\equiv\; a^*(\mu^+,\mu) - a^*(\mu^-,\mu) \;=\; \gamma\,\Delta x(\mu,q).$$

Hence signals are distinguishable *from actions at belief μ if and only if $\Delta a(\mu) \geq h$, and* indistinguishable *if $\Delta a(\mu) < h$.*

In particular, at the cascade boundaries $\mu \in \{1-q,\ q\}$ one has $\Delta x(\mu,q) = 0$, so no finite γ clears the threshold. Away from the boundaries, higher q (better signals) increases $\Delta x(\mu,q)$ and relaxes the threshold; higher conformity (lower γ) tightens it.

Proof. Fix μ, $q > 1/2$, responsiveness $\gamma \in [0,1]$, and grid step $h \geq 0$. Let the continuous best response be $a^*(x,\mu) = \gamma x + (1-\gamma)\mu$, and denote $\mu^\pm$ the posteriors after one signal, so that by (4)

$$\Delta x(\mu,q) \;\equiv\; \mu^+ - \mu^- \;=\; \frac{\mu(\mu-1)\,(2q-1)}{(2\mu q-\mu-q)\,(2\mu q-\mu-q+1)}.$$

The signal–induced separation in actions is

$$\Delta a(\mu) \;\equiv\; a^*(\mu^+,\mu) - a^*(\mu^-,\mu) \;=\; \gamma\,\Delta x(\mu,q).$$

Let $\tilde{a} = \mathcal{C}(a^*)$ be the coarsened action obtained by rounding to the nearest grid point of step h (ties broken by the public according to the selection in Section 3). The coarsening map $\mathcal{C}$ is monotone and piecewise constant with flat segments of width h. Hence $\tilde{a}(\mu^+) = \tilde{a}(\mu^-)$ if and only if $|\Delta a(\mu)| < h$, and $\tilde{a}(\mu^+) \neq \tilde{a}(\mu^-)$ if and only if $|\Delta a(\mu)| \geq h$. Therefore signals are inferable from actions at belief μ if and only if

$$\gamma\,\Delta x(\mu,q) \;\geq\; h.$$

At the knife–edge $\gamma\,\Delta x(\mu,q) = h$ inference is pinned down by our tie-breaking (Section 3). Finally, at the classical cascade boundaries $\mu \in \{1-q, q\}$ we have $\Delta x(\mu,q) = 0$ (plugging μ into the formulas for $\mu^\pm$ yields equality), so—absent hard evidence—the soft channel is locally mute. This proves the claim. □

Figure 3 complements Proposition 5.1: with $\gamma = 0.6$ and $h = 0.05$, breakability rises with verification p and falls with relative cost c/V; when $\gamma\,\Delta x(\mu,q) < h$, actions cease to separate and disclosures drive correction.

This delivers a simple intuition: continuous actions do not automatically prevent cascades. With coarse observation or even modest rounding, sufficiently strong conformity collapses the signal-induced movement in actions below observational granularity, re-creating regions in which actions are effectively uninformative about private signals.

Investigation adds a second, discontinuous channel for information. If hard evidence E arrives and is disclosed, beliefs jump to 1 regardless of γ or h. Thus disclosures break any ongoing cascade. When $\Delta a(\mu)$ falls below the observational threshold, investigation is the only route to escape from

regions where soft signals cannot move public belief. Conversely, when $\gamma\,\Delta x(\mu)$ exceeds h, soft signals are again visible in actions and can steer beliefs without disclosure.

Proposition 5.2. *Fix (q, p, c, V, γ, h). Suppose the binary-action investigation cutoffs from Section 4 apply to the posterior x (the value comparison is unaffected by the action's continuity). Then:*

1. *If there exists an interval $\mathcal{I} \subset (0,1)$ with $\gamma\,\Delta x(\mu) \geq h$ for all $\mu \in \mathcal{I}$, any wrong belief that enters $\mathcal{I}$ is corrected with positive probability without investigation.*

2. *If at some $\bar{\mu}$ the investigation condition of Proposition 4.5 holds, a wrong cascade at or near $\bar{\mu}$ is broken with positive probability via disclosure, independent of γ and h.*

In particular, the society is misinformation-resilient *whenever either channel is active along the evolution of beliefs: high responsiveness (γ large, h small, q large) or sufficiently attractive investigation ($p(1+V) \gtrsim 2c$ near the boundaries).*

Discussion. Equation (3) isolates conformity as a linear dampener of responsiveness. Without frictions ($h = 0$) any $\gamma > 0$ restores full revelation of soft signals and precludes action-based herding. With even mild coarsening, however, (4) shows that signals are intrinsically least informative near the classic cascade bands, and strong conformity can render their effect invisible. Hard evidence then plays a pivotal role: it introduces discrete, verifiable corrections that robustly collapse conformity-driven cascades. The resilience frontier in $(q, p, c/V, \gamma, h)$ thus has two margins—one continuous (soft) and one discontinuous (hard)—each expanding when media quality or verification improves, or when conformity pressure eases.

Figure 2 visualizes the binary benchmark: simulated breakability is high on and below the hard-evidence frontier and declines as verification worsens or costs rise, mirroring Theorem 6.4 and Corollary 6.5.

6 Misinformation Resilience

We formalize when a society eventually corrects a wrong cascade through either soft information (actions that reveal private signals) or hard evidence (verifiable disclosures).

Fix (q, γ, h) and the equilibrium selection from Section 3. The *resilience frontier* is the locus in $(p,\ c/V)$ space that separates parameter pairs for which a wrong cascade at the boundary is breakable with positive probability from those for which it is not. In the binary benchmark this specializes to the hard-evidence condition in Corollary 4.6; with continuous actions and coarsening, it coincides with the parameter region characterized by Proposition 5.1 and Theorem 6.4. Figure 2 provides a numerical visualization.

Proposition 6.1. *Fix (q, γ, h) and the equilibrium selection from Section 3. Let*

$$\mathcal{R}(q, \gamma, h) \;\equiv\; \{(p,\, c/V) \in (0,1]\times[0,\infty) :\ \textit{a wrong boundary cascade is breakable with positive probability}\}.$$

Then:

1. *(Hard channel) If* $(p, c/V) \in \mathcal{R}$ *and* $p' \geq p$, $c'/V \leq c/V$, *then* $(p', c'/V) \in \mathcal{R}$.
2. *(Soft channel) If* $(p, c/V) \in \mathcal{R}$ *and* $\gamma' \geq \gamma$, $h' \leq h$, *then* $(p, c/V) \in \mathcal{R}(q, \gamma', h')$.
3. *(Signal quality) If* $q' \geq q$, *then* $\mathcal{R}(q, \gamma, h) \subseteq \mathcal{R}(q', \gamma, h)$.

Proof sketch. (i) Increasing p or decreasing c/V weakly enlarges the investigation region by Lemma 4.3, and disclosures arrive at least as often, so any path that breaks the cascade under $(p, c/V)$ also breaks it under $(p', c'/V)$ by coupling.

(ii) By Proposition 5.1, soft inferability at belief μ requires $\gamma\,\Delta x(\mu, q) \geq h$. Raising γ or lowering h weakly relaxes this inequality at every μ, so any soft-driven correction remains feasible (and hard-driven corrections are unchanged).

(iii) From Lemma 4.2, $\Delta x(\mu, q)$ is (weakly) increasing in q, so the soft condition becomes easier; higher q also increases the fraction of high-posterior agents, enlarging the investigation region. Combining with (i)–(ii) gives the inclusion. □

A public belief μ is a *cascade belief* if the optimal action is independent of the private signal (Section 3). A cascade at μ is *wrong* if the induced action does not match the true state. Fix parameters (q, p, c, V, γ, h) and a strategy profile constituting a PBE.

Definition 6.2. A cascade at belief μ is *breakable* if, conditional on $\mu_t = \mu$ and the cascade being wrong, there exists a path with positive probability on which future public beliefs leave the cascade region and converge to the truth (either via distinguishable soft signals or via disclosure).

Definition 6.3. The society is *misinformation-resilient* if any wrong cascade that arises along the equilibrium path is breakable. Equivalently, starting from any $\mu_1 \in (0, 1)$, the probability that a wrong cascade persists forever is zero.

The two channels identified earlier are: (i) the *soft channel*, operative when signal-induced action movements are observable (Proposition 5.1); and (ii) the *hard channel*, operative when investigation is attractive near cascades (Propositions 4.4–4.5).

6.1 Main characterization near cascade boundaries

Write the classical cascade boundaries as $\underline{\mu} = 1 - q$ (lower) and $\overline{\mu} = q$ (upper). Recall $\Delta x(\mu) = \mu^+ - \mu^-$ and the coarse-observation threshold $\gamma\,\Delta x(\mu) \geq h$ from Proposition 5.1; and the knife-edge investigation condition $\frac{1}{2}p(1 + V) \geq c$ from (2).

Theorem 6.4. *Fix* (q, p, c, V, γ, h).

1. *If there exists* $\varepsilon > 0$ *such that* $\gamma\,\Delta x(\mu) \geq h$ *for all* $\mu \in [\underline{\mu}, \underline{\mu} + \varepsilon]$ *(respectively, for all* $\mu \in [\overline{\mu} - \varepsilon, \overline{\mu}]$*), then any wrong lower (respectively, upper) cascade at the boundary is breakable via soft information.*

2. *If* $\frac{1}{2}\,p(1+V) \geq c$*, then any wrong cascade at either boundary is breakable via investigation and disclosure.*

3. *If* $\frac{1}{2}\,p(1+V) < c$ *and there exists* $\varepsilon > 0$ *with* $\gamma\,\Delta x(\mu) < h$ *for all* μ *in a neighborhood of the boundary (lower or upper), then a wrong cascade at that boundary is not breakable (locally absorbing).*

Sketch. (i) When $\gamma\,\Delta x(\mu) \geq h$, actions corresponding to $s = 1$ and $s = 0$ are separated by at least one grid boundary, hence publicly distinguishable. Beliefs therefore update away from the boundary with positive probability, breaking the cascade (Proposition 5.1). (ii) At $\mu = \underline{\mu}$ the best posterior is $\mu^+ = 1/2$; Lemma 4.1 and (2) imply investigating strictly dominates after a favorable signal, yielding disclosure with probability $p > 0$ and a belief jump to 1. The upper boundary is symmetric. (iii) If both channels fail locally (no investigation incentive by (2) and no signal distinguishability by Proposition 5.1), then neither actions nor investigations can move beliefs away from the boundary, so the wrong cascade is locally absorbing. □

Corollary 6.5. *Define the* hard-evidence frontier $\mathcal{H}:\ \frac{1}{2}\,p(1+V) = c$ *and the* soft-inference frontier $\mathcal{S}:\ \gamma\,\Delta x(\underline{\mu}) = h$ *and* $\gamma\,\Delta x(\overline{\mu}) = h$*. The parameter region*

$$\{(q,p,c,V,\gamma,h):\ \tfrac{1}{2}\,p(1+V) \geq c\}\ \cup\ \{(q,p,c,V,\gamma,h):\ \gamma\,\Delta x(\underline{\mu}) \geq h \text{ and } \gamma\,\Delta x(\overline{\mu}) \geq h\}$$

is sufficient for global resilience. Conversely, if both inequalities fail in neighborhoods of the boundaries, wrong cascades are locally absorbing.

Figure 2 visualizes the binary benchmark: simulated breakability is high on and below the hard-evidence frontier and declines as verification worsens or costs rise, mirroring Theorem 6.4 and Corollary 6.5.

Along $\mathcal{H}$, the slope satisfies $dp/dV = -\,p/(1+V)$ at fixed c (more discovery rents substitute for verification quality). Lower c or higher p/V expand the hard-evidence region. Along $\mathcal{S}$, since $\Delta x(\mu)$ is increasing in q away from the boundaries (and equals zero exactly at them), higher q or higher γ or finer observation (smaller h) expand the soft-inference region. These margins are complementary: either can secure resilience on its own.

6.2 Beyond the boundary: interior beliefs and dynamics

Theorem 6.4 focuses on boundary behavior, which is the bottleneck for correction since $\Delta x(\mu)$ is minimized there (equals zero in the binary baseline). Away from the boundaries, $\Delta x(\mu)$ grows with q and the soft channel becomes progressively more potent; similarly, the investigation cutoffs in Proposition 4.4 are easier to satisfy because posterior x moves away from 1/2. Hence the frontier in Corollary 6.5 is conservative: once beliefs are nudged off the boundary (by any small shock), the set of parameters sustaining resilience strictly expands.

First, policies that increase *verification capability* ($p \uparrow$) or reduce *verification frictions* ($c \downarrow$) guarantee breakability even in the most hostile region (the boundaries). Second, policies that

improve *media quality* ($q \uparrow$), reduce *conformity pressure* ($\gamma \uparrow$), or increase *observational granularity* ($h \downarrow$) restore the informativeness of actions and can obviate the need for costly investigation. Third, if platforms inadvertently coarse-grain signals (large h) while conformity is strong (small γ), even good private signals cannot move beliefs near cascades; in such environments, supporting verification is pivotal.

We assumed $V \geq 0$, so disclosure is weakly optimal whenever E arrives. Allowing strategic concealment (e.g., V depends on audience or future payoffs) shifts the hard-evidence frontier upward by replacing p with $p \cdot \Pr(d = 1 \mid E)$; the qualitative conclusions are unchanged but the resilience region shrinks proportionally to the equilibrium disclosure rate.

With local observation, boundary conditions are defined on neighborhood beliefs; $\Delta x(\mu)$ is then evaluated per neighborhood. Theorems above apply node-wise; resilience requires that either soft distinguishability or hard-evidence incentives hold on a per-neighborhood basis. Sparse networks tighten $\mathcal{S}$ (actions are noisier aggregates) and raise the value of $\mathcal{H}$ (disclosures percolate globally once observed).

Additional simulation details and figures appear in Appendix B.

7 Welfare and Policy

We sketch a planner's problem that captures two externalities: (i) a *verification externality*—an investigator's disclosure permanently resolves uncertainty for everyone; and (ii) a *conformity / observability externality*—platform coarsening and social pressure make actions less informative to others. Policies act on three margins: the effective investigation cost (c), the verification technology (p), and the informativeness of actions (through conformity γ and observational granularity h). We keep the discussion local to the cascade boundaries, which are the bottlenecks (Section 6).

Let W denote expected discounted welfare from date t onward, where each period's flow payoff is accuracy (1 if $a_t = \theta$, 0 otherwise) plus any social value from a disclosure. Private choices ignore the public-good value of moving beliefs out of a wrong cascade. A reduced-form planner therefore chooses instruments to minimize the probability that a wrong cascade persists forever, subject to policy costs.

(i) A per-investigation subsidy $s \in [0, c]$ lowers the private cost to $c - s$. (ii) Investment in verification raises the success probability from p to $p + \Delta p$ (e.g., audit capacity, archiving, authenticity tools). (iii) Platform design reduces coarsening from h to $h - \Delta h$ (finer observability) and/or attenuates conformity (nudges that increase γ). We assume convex policy costs $\kappa_c(s)$, $\kappa_p(\Delta p)$, and $\kappa_h(\Delta h)$.

At the cascade boundaries, the knife-edge condition for local breakability is $\frac{1}{2}\,p(1+V) \geq c$ (Corollary 4.6). With a subsidy s, the condition becomes

$$\tfrac{1}{2}\,p(1+V) \;\geq\; c - s. \tag{5}$$

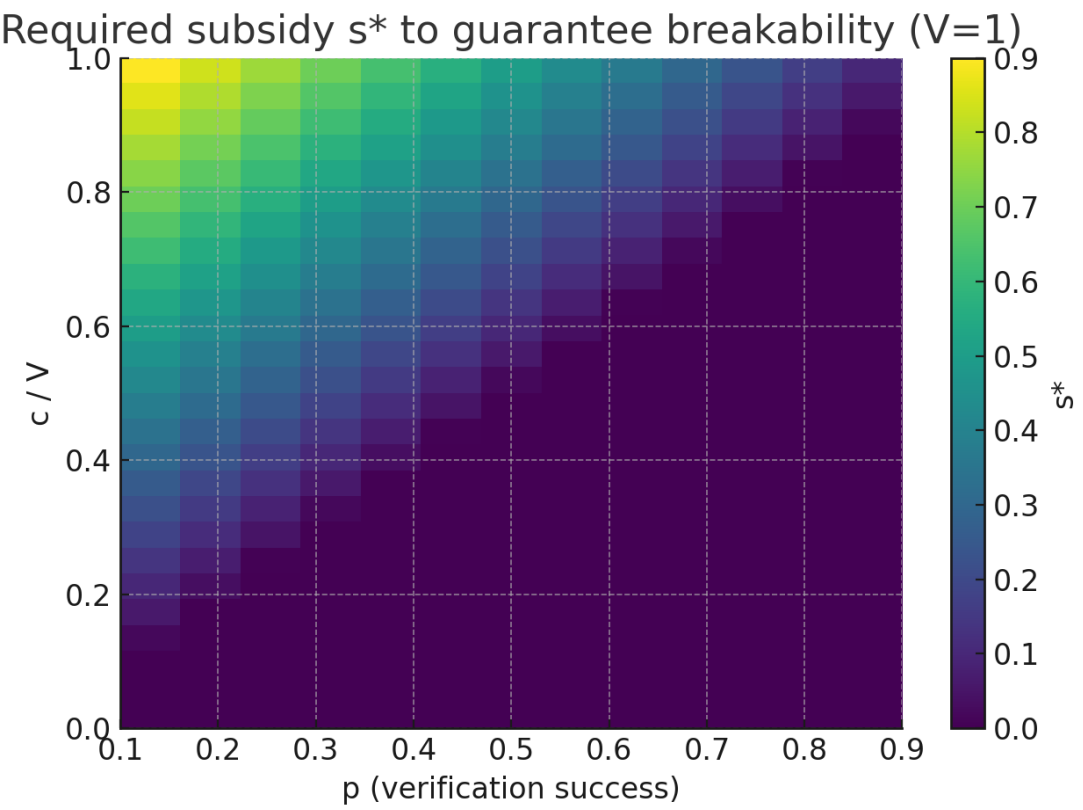

Figure 1: Required subsidy s^* over $(p,\ c/V)$ for $V{=}1$ to guarantee boundary breakability (Proposition 7.1). The required subsidy is zero on and below the hard-evidence frontier and rises linearly above it.

Proposition 7.1. *Given (p, V, c), the least costly hard-evidence intervention that guarantees boundary breakability is either a subsidy*

$$s^* \;=\; \max\Big\{0,\ c - \tfrac{1}{2}\,p(1+V)\Big\},$$

or, equivalently, an increase in verification quality

$$\Delta p^* \;=\; \max\left\{0,\ \frac{2(c-s)}{1+V} - p\right\}.$$

Any pair $(s, \Delta p)$ satisfying (5) secures local resilience; the planner chooses the cost-minimizing pair under κ_c, κ_p.

Sketch. Corollary 4.6 applied to effective cost $c - s$ and success rate $p + \Delta p$. □

As a direct illustration, Figure 1 plots the implied subsidy schedule $s^*(p, c/V)$: zero on and below the frontier, rising linearly above it.

At belief μ, signals are distinguishable from actions iff $\gamma\,\Delta x(\mu) \geq h$ (Proposition 5.1). Near the

boundaries, $\Delta x(\mu)$ is minimal, so a sufficient condition for local resilience via soft information is

$$\gamma \;\geq\; \frac{h}{\Delta x(\mu, q)} \qquad \text{for all } \mu \text{ in a neighborhood of } \{1-q,\ q\}. \tag{6}$$

Proposition 7.2. *Fix (q, γ, h). Let $\overline{\mu} \in \{1-q, q\}$ and $\Delta x_\varepsilon \equiv \inf_{\mu \in (\overline{\mu}-\varepsilon, \overline{\mu}+\varepsilon)} \Delta x(\mu)$ for some small $\varepsilon > 0$. A sufficient policy to ensure soft breakability at both boundaries is either (i) to reduce coarsening to*

$$h^* \;\leq\; \gamma\, \Delta x_\varepsilon,$$

or (ii) to raise responsiveness to

$$\gamma^* \;\geq\; \frac{h}{\Delta x_\varepsilon}.$$

The planner picks the lower-cost option under κ_h, κ_γ (where κ_γ is the cost of nudges that increase γ).

Proof. Write $f(\mu) \equiv \Delta x(\mu, q)$ as in (4). For $q > 1/2$, f is continuous on $[0,1]$, satisfies $f(1-q) = f(q) = 0$, and $f(\mu) > 0$ on $(1-q, q)$. By Proposition 5.1, signals are inferable at belief μ iff $\gamma \geq h/f(\mu)$. Hence, for any neighborhood U of the boundary (a small interval around $\{1-q, q\}$), inferability throughout U holds iff

$$\gamma \;\geq\; \sup_{\mu \in U} \frac{h}{f(\mu)}\,.$$

Since f is continuous and strictly positive on $U \setminus \{1-q, q\}$, the supremum is finite for any fixed U; this yields the stated condition with the "for all μ in a neighborhood of $\{1-q, q\}$" phrasing. The converse direction is immediate from Proposition 5.1. □

Define the feasible set

$$\mathcal{R} \;=\; \Big\{(s, \Delta p, \Delta h, \Delta\gamma) :\ (5) \text{ holds and } (6) \text{ holds on both boundaries}\Big\}.$$

A simple program is

$$\min_{(s, \Delta p, \Delta h, \Delta\gamma) \in \mathcal{R}} \quad \kappa_c(s) + \kappa_p(\Delta p) + \kappa_h(\Delta h) + \kappa_\gamma(\Delta\gamma).$$

Because the two margins operate independently at the boundaries (Theorem 6.4), any solution that satisfies either the hard condition or the soft condition at both boundaries is sufficient for global resilience.

Subsidizing verification ($s \uparrow$) or improving it technologically ($p \uparrow$) moves the economy vertically across the hard-evidence frontier; refining observability ($h \downarrow$) or boosting responsiveness ($\gamma \uparrow$) moves it horizontally across the soft frontier (Corollary 6.5). When platforms heavily coarsen actions (*large* h) or conformity is strong (*small* γ), the soft margin is expensive; hard-evidence tools are then cost-effective. Conversely, in high-q environments where $\Delta x(\mu)$ is large near the boundaries, modest UI changes that reduce h (finer ratings/scales; richer reaction sets) or mild accuracy nudges that

increase γ can obviate costly subsidies.

If disclosure is imperfect (some investigators conceal), the hard-evidence frontier shifts by replacing p with $p \cdot \Pr(d{=}1 \mid E)$. Whistleblower rewards that raise V or policies that mandate/enable verifiable archiving effectively increase $\Pr(d{=}1 \mid E)$ and move the frontier down. Caution is warranted with selective warning labels: they may reduce h for labeled items but increase perceived accuracy of unlabeled content (an "implied truth" effect), which can dampen soft inference away from labels. In our terms, such policies change h unevenly and can tighten the soft frontier in unlabeled regions. A robust approach combines small, general-purpose boosts to γ (accuracy prompts) with either lower h platform-wide or targeted support for verification (s or p) at topics where $\Delta x(\mu)$ is intrinsically small.

8 Extensions and Robustness

This section records several add-ons that leave our main insights intact while clarifying scope.

8.1 Heterogeneous investigation costs and benefits

Let costs and discovery benefits be agent-specific, (c_i, V_i) drawn i.i.d. from a continuous distribution F on $[0, \bar{c}] \times [0, \bar{V}]$, observed privately by agent i before choosing $i_t \in \{0, 1\}$. The value comparison in Lemma 4.1 and Proposition 4.4 applies typewise; in the "clean" region ($x \geq 1/(2-p)$) the indifference condition is $x\, p\, V_i = c_i$.

Proposition 8.1. *Fix a posterior $x \geq 1/(2-p)$. There exists a monotone selection rule: agent i investigates iff $\phi_i \equiv V_i/c_i \geq \phi^*(x)$, where $\phi^*(x) = 1/(px)$. The ex-ante investigation probability at (x, p) equals $1 - \Phi(\phi^*(x))$, where Φ is the c.d.f. of ϕ_i. Consequently, the hard-evidence margin strengthens (weakens) under MLRP shifts of F that raise (lower) the distribution of ϕ_i.*

The resilience frontier shifts outward when the population contains a sufficiently thick upper tail of high ϕ_i types (high V_i or low c_i). This preserves the qualitative role of p and V in Corollary 4.6, replacing $p(1+V)$ with $p\,\mathbb{E}[V_i \mid \phi_i \geq \phi^*]$ in knife-edge comparisons.

8.2 Strategic concealment and disclosure frictions

Suppose disclosure, conditional on E, is chosen to maximize $V_i - \Delta_i$, where Δ_i is a (possibly type- and history-dependent) private cost of disclosure. Let $\pi(\cdot) \in [0, 1]$ denote the equilibrium disclosure probability given evidence. Then all hard-evidence formulas carry through with the substitution

$$p \quad \rightsquigarrow \quad \tilde{p} \equiv p\,\mathbb{E}[\pi(E, H_t) \mid E].$$

The hard-evidence frontier in Corollary 6.5 becomes $\frac{1}{2}\,\tilde{p}(1+\bar{V}) \geq \bar{c}$ under homogeneous $(\bar{c}, \bar{V})$ or the selection-weighted analogue under heterogeneity. Whistleblower rewards and verifiable-archiving rules operate either by raising V_i directly or by increasing π, thereby expanding the resilience region.

8.3 Networked observation

Let agents observe only a neighborhood's past actions and disclosures on a graph $G = (N, \mathcal{E})$. Beliefs are now local: $\mu_t(i)$ is agent i's public belief given her neighborhood history. Define local cascade boundaries $\underline{\mu}(i) = 1 - q$ and $\overline{\mu}(i) = q$ as before. Propositions 5.1 and 4.5 apply node-wise with μ replaced by $\mu(i)$. Disclosures, once observed, percolate through G deterministically if they are globally visible (e.g., archived, searchable). When observation is purely local, a disclosure at node j shifts beliefs to 1 along paths that include j; resilience then depends on whether the set of nodes recurrently visited by the belief process intersects these paths. Sparsity tightens the soft frontier (actions are noisier aggregates), increasing the relative value of the hard-evidence channel.

8.4 Coarse and noisy actions beyond rounding

The coarsening operator $\mathcal{C}$ can encode bounded attention or platform frictions more generally. If the public observes $\tilde{a} = a^* + \eta$ with noise η (mean zero, variance σ^2), then distinguishability of $s = 1$ vs. $s = 0$ reduces to a standard signal-detection condition:

$$\frac{|\Delta a(\mu)|}{\sigma} \;=\; \frac{\gamma\,\Delta x(\mu)}{\sigma} \;\geq\; z_\alpha,$$

for a target false-positive rate α and critical value z_α. The soft frontier becomes $\gamma\,\Delta x(\mu) \geq z_\alpha \sigma$ in place of $\gamma\,\Delta x(\mu) \geq h$, leaving all comparative statics unchanged.

8.5 Repeated interactions and reputation

If some agents reappear and V_i includes reputational returns that depend on future audiences, then π generally rises (evidence is more likely to be disclosed), effectively increasing $\tilde{p}$. Reputational benefits can also reduce *effective* conformity costs by making deviations from consensus less penalized in expectation, which raises γ. Both forces shift the resilience frontier outward. The main caveat is selection: reputational stakes may induce over-investigation in high-visibility states, concentrating disclosures on salient topics.

8.6 Evidence with rare false positives

Suppose investigations can (rarely) generate spurious "evidence" with probability $\epsilon \ll 1$ when $\theta = 0$. Then disclosures at $\theta = 0$ do not fully pin down the state. Beliefs jump to

$$\mu' = \Pr(\theta = 1 \mid E) = \frac{\mu p}{\mu p + (1 - \mu)\epsilon},$$

which exceeds μ but is below 1 unless $\epsilon = 0$. All hard-evidence results continue to hold with 1 replaced by μ', and with $\tilde{p} = p(1 - \epsilon)$ in knife-edge inequalities. For small ϵ, the resilience frontier shifts inward by $O(\epsilon)$.

8.7 Batch arrivals and finite horizons

With k agents arriving per period and public aggregation of their actions before beliefs update, $\Delta x(\mu)$ scales by the cross-sectional amplification in the law of large numbers. If actions are observed without noise, the soft channel strengthens in k; with coarse/noisy observation, the effective granularity h (or noise σ) scales down as k grows, again expanding the soft frontier. In finite horizons, the late arrival of investigators raises the value of early disclosure; the knife-edge $\frac{1}{2}p(1+V) \geq c$ becomes easier to satisfy if V includes time-sensitive rents.

Across these variants, two margins continue to govern resilience: (i) *soft* action informativeness (controlled by q, γ, and observability), and (ii) *hard* verification (controlled by p, V, c, and disclosure frictions). Heterogeneity, networks, and modest noise shift quantitative thresholds but leave the qualitative geometry of the resilience frontier unchanged.

9 Conclusion

We develop a sequential social–learning framework in which agents can endogenously verify claims and occasionally generate hard, publicly checkable evidence. Two forces shape outcomes: a soft channel, where actions convey private information, and a hard channel, where rare disclosures coordinate everyone on the truth. In the binary benchmark we show that investigation follows a unique cutoff and that the position of this cutoff relative to the classical cascade bands governs whether a wrong boundary cascade is breakable; a simple boundary condition delivers an immediate resilience test (Proposition 4.4, Proposition 4.5, Corollary 4.6). With continuous actions under conformity and coarse observation, we derive a transparent responsiveness threshold at which the soft channel falls silent, while occasional disclosures still overturn false cascades (Proposition 5.1). These pieces assemble into a tractable resilience frontier with clean comparative statics and monotonicity (Theorem 6.4, Corollary 6.5, Proposition 6.1).

The analysis clarifies policy levers without committing to a specific implementation. Interventions that raise verification success (p) or lower effective cost (c/V) expand the breakable region; so do designs that increase responsiveness or granularity of observation (higher γ, lower h) (Section 7). In environments where conformity pressures or coarse signals mute the soft channel, even small improvements in verifiability can substitute for otherwise uninformative actions. Our baseline takes disclosure as weakly optimal ($V \geq 0$); strategic concealment simply rescales the effective arrival rate by $p\,\mathbb{E}[\pi]$ without altering the geometry of the frontier.

Robustness variants leave the message intact. Allowing rare false positives tightens the hard–evidence frontier in a transparent way; partially observable attempts to investigate weakly enlarge the breakable set by shifting incentives; alternative conformity costs widen the dead zone but preserve the hard–evidence logic (Appendix A). Numerical illustrations visualize these thresholds and frontiers but are not used in proofs; additional figures and code appear in Appendix B.

Two natural directions remain. First, richer interaction structures—networks, repeated encounters, or platform feedback—can endogenize both observability (h) and responsiveness (γ), poten-

tially generating new selection effects while preserving our boundary logic. Second, heterogeneous investigation benefits and costs (or institutionally supplied verification) would speak to the optimal placement of limited verification capacity. Our results suggest a simple organizing principle: when soft signals are muted, a modest increase in verifiability can restore truth at scale.

A Proofs

A.1 Proof of Lemma 4.1

Proof. Fix a public history H_t and a realized private posterior $x = \Pr(\theta = 1 \mid H_t, s_t) \in [0, 1]$. If the agent does *not* investigate, the optimal binary action is $a = 1$ iff $x \geq \frac{1}{2}$, yielding

$$U^{\text{no}}(x) = \max\{x, 1 - x\} \equiv A(x).$$

If the agent investigates, hard evidence E arrives iff $\theta = 1$ and the investigation succeeds; thus $\Pr(E \mid x) = xp$. Upon arrival, the agent discloses, gains V, takes $a = 1$ and secures accuracy 1. If no evidence arrives (probability $1 - xp$), the posterior updates to

$$x' \equiv \Pr(\theta = 1 \mid \text{failure}, H_t, s_t) = \frac{x(1-p)}{x(1-p) + (1-x)\cdot 1} = \frac{x(1-p)}{1-xp},$$

because $\Pr(\text{failure} \mid \theta = 1) = 1 - p$ and $\Pr(\text{failure} \mid \theta = 0) = 1$.

The continuation value is then $A(x')$. Hence

$$U^{\text{inv}}(x) = xp(1 + V) + (1 - xp)A(x') - c. \tag{7}$$

Consider $x \geq \frac{1}{2}$ so $U^{\text{no}}(x) = x$. There are two subcases.

Case 1: $x' \geq \frac{1}{2}$. This is equivalent to $x \geq 1/(2 - p)$ since

$$x' \geq \tfrac{1}{2} \iff \frac{x(1-p)}{1-xp} \geq \tfrac{1}{2} \iff 2x(1-p) \geq 1 - xp \iff x(2-p) \geq 1 \iff x \geq \frac{1}{2-p}.$$

Then $A(x') = x'$ and $(1 - xp)A(x') = (1 - xp)\frac{x(1-p)}{1-xp} = x(1 - p)$. Using (7),

$$U^{\text{inv}}(x) - U^{\text{no}}(x) = xp(1 + V) + x(1 - p) - c - x = xpV - c.$$

Case 2: $x' < \frac{1}{2}$. Equivalently $x < 1/(2 - p)$. Now $A(x') = 1 - x'$ and hence

$$(1 - xp)A(x') = (1 - xp)\Big(1 - \frac{x(1-p)}{1-xp}\Big) = 1 - x.$$

Again from (7),

$$U^{\text{inv}}(x) - U^{\text{no}}(x) = xp(1 + V) + (1 - x) - c - x = 1 - c + x\big(p(1 + V) - 2\big).$$

This establishes the stated piecewise form for $x \geq \frac{1}{2}$. By symmetry of labels of states and actions, the expression for $x \leq \frac{1}{2}$ follows by replacing x with $1 - x$. In each subinterval the expression is affine in x, implying a single crossing. □

A.2 Proof of Proposition 4.4

Proof. For $x \geq \frac{1}{2}$:

Region A ($x \geq 1/(2-p)$): From Lemma 4.1, $U^{\text{inv}} - U^{\text{no}} = xpV - c$, which is increasing and crosses zero at $x_1^* = c/(pV)$ (if $pV > 0$). Thus the agent investigates iff $x \geq x_1^*$ (provided $x_1^* \in [\frac{1}{2}, 1]$; if $x_1^* < \frac{1}{2}$ then investigation is (weakly) optimal throughout the region; if $x_1^* > 1$ it is never optimal). Comparative statics are direct.

Region B ($\frac{1}{2} \leq x < 1/(2-p)$): From Lemma 4.1, $U^{\text{inv}} - U^{\text{no}} = 1 - c + x\{p(1+V) - 2\}$. If $p(1+V) = 2$, the difference equals $1 - c$, so investigation is (weakly) optimal everywhere in Region B iff $c \leq 1$. If $p(1+V) > 2$, the slope is positive and, since $1 - c \geq 0$ for any $c \leq 1$, investigation is (weakly) optimal on Region B (and strictly at interior x if $c < 1$). If $p(1+V) < 2$, the slope is negative and the indifference point solves $1 - c + x\{p(1+V) - 2\} = 0$, i.e.

$$x_2^* = \frac{1-c}{2 - p(1+V)}.$$

Because the slope is negative, the agent investigates iff $x \geq x_2^*$. Again, monotone comparative statics follow. The characterization for $x \leq \frac{1}{2}$ is obtained by symmetry (replace x with $1 - x$ in the conditions and thresholds). □

A.3 Proof of Proposition 4.5

Proof. Suppose $\mu_t \leq 1 - q$ (lower cascade). Then, holding μ_t fixed, the posterior after a positive signal $x^+ \equiv \mu_t^+$ is weakly larger than the posterior after a negative signal $x^- \equiv \mu_t^-$. Moreover, both $x^+, x^- \leq \frac{1}{2}$, with equality $x^+ = \frac{1}{2}$ at the boundary $\mu_t = 1 - q$. By Lemma 4.1, $U^{\text{inv}} - U^{\text{no}}$ has the single-crossing property and is (weakly) increasing in x on $[0, \frac{1}{2}]$ after applying the $x \mapsto 1 - x$ symmetry. Therefore, an agent strictly prefers to investigate for some signal realization if and only if she prefers to investigate at $x = x^+$. This yields the stated condition. If it holds, then with probability $\Pr(s_t = 1 \mid \mu_t) > 0$ the agent investigates; with probability $p > 0$ evidence arrives and is disclosed, breaking the cascade. The upper-cascade case is symmetric: when $\mu_t \geq q$, the smallest posterior is $1 - x^-$ on $[\frac{1}{2}, 1]$, and single crossing again implies the stated condition. □

A.4 Proof of Corollary 4.6

Proof. At the lower boundary $\mu_t = 1 - q$, $x = \mu_t^+ = \frac{1}{2}$. Lemma 4.1 (Case 2) gives

$$U^{\text{inv}}(\tfrac{1}{2}) - U^{\text{no}}(\tfrac{1}{2}) = \tfrac{1}{2}\, p(1+V) - c.$$

If $\frac{1}{2}\,p(1+V) > c$, an agent who receives $s_t = 1$ strictly prefers to investigate, which yields disclosure with probability $p > 0$ and breaks the cascade. If equality holds, investigation is weakly optimal; any equilibrium/tie-breaking that selects investigation suffices, and the result also obtains under arbitrarily small perturbations of (p, V, c). The upper boundary is symmetric (use $x = 1 - \mu_t^- = \frac{1}{2}$). □

A.5 Proof of Proposition 5.2

Proof. Part (i): If there exists an interval $\mathcal{I}$ on which $\gamma\,\Delta x(\mu) \geq h$ pointwise, then by Proposition 5.1 the two signal–contingent actions are publicly distinguishable throughout $\mathcal{I}$. Therefore, whenever the public belief lies in $\mathcal{I}$, there is positive probability (equal to the signal likelihood) that the observed action reveals the private signal and hence moves beliefs toward the truth; iterating yields eventual correction with positive probability without requiring disclosure.

Part (ii): By Proposition 4.5, if the binary-action investigation condition holds at some $\bar{\mu}$, then with positive probability an agent investigates and discloses, forcing beliefs to 1 irrespective of (γ, h). Therefore any wrong cascade at or near $\bar{\mu}$ is broken with positive probability via the hard-evidence channel. □

A.6 Proof of Theorem 6.4

Proof. We prove parts 1-3 in turn:

1. Let $\underline{\mu} = 1 - q$ (the lower boundary). If there exists $\varepsilon > 0$ such that $\gamma\,\Delta x(\mu) \geq h$ for all $\mu \in [\underline{\mu},\, \underline{\mu} + \varepsilon]$, then whenever the public belief lies in that interval, the two signal–contingent actions are publicly distinguishable (Proposition 5.1). Hence with positive probability the next observed action shifts belief strictly above $\underline{\mu}$; by persistence, beliefs exit the cascade region and move toward the true state with positive probability. The upper boundary is symmetric.

2. At either boundary, the best posterior equals $\frac{1}{2}$ (lower: $\mu^+ = \frac{1}{2}$; upper: $1 - \mu^- = \frac{1}{2}$). By Lemma 4.1, $U^{\text{inv}}(\frac{1}{2}) - U^{\text{no}}(\frac{1}{2}) = \frac{1}{2}p(1+V) - c$. If $\frac{1}{2}p(1+V) \geq c$, investigation is weakly optimal after the favorable signal; with probability $p > 0$ evidence arrives and disclosure jumps beliefs to 1, breaking the cascade.

3. Suppose $\frac{1}{2}p(1+V) < c$ and there exists $\varepsilon > 0$ such that $\gamma\,\Delta x(\mu) < h$ for all μ in a neighborhood of the boundary. Then at and near the boundary, investigation is strictly suboptimal after any signal, and soft signals are indistinguishable from observed actions. Therefore neither channel moves beliefs away from the boundary; the wrong cascade is locally absorbing, proving non-breakability.

□

Lemma A.1. *Let $\Delta x(\mu, q)$ be as in (4). For $q \in (1/2, 1)$ define*

$$\kappa(q) \;\equiv\; \frac{(q-\frac{1}{2})^2}{q(1-5q+12q^2-16q^3+12q^4-4q^5)} \;>\; 0.$$

Then, as $\mu \to q$ and $\mu \to 1-q$,

$$\Delta x(\mu, q) \;=\; \kappa(q)\,(q-\mu) \;+\; o(|\mu - q|), \qquad \Delta x(\mu, q) \;=\; \kappa(q)\,(\mu-(1-q)) \;+\; o(|\mu-(1-q)|).$$

Proof. Differentiate (4) in μ and evaluate at $\mu = q$ and $\mu = 1-q$; one obtains $\partial_\mu \Delta x(\mu, q)\big|_{\mu=q} = -\kappa(q)$ and $\partial_\mu \Delta x(\mu, q)\big|_{\mu=1-q} = +\kappa(q)$. Hence the gap vanishes linearly at the boundaries with slope magnitude $\kappa(q)$, which is increasing in q and diverges as $q \uparrow 1$. □

A.7 Proof of Corollary 6.5

Proof. The set $\{(q, p, c, V, \gamma, h) : \frac{1}{2}p(1+V) \geq c\}$ makes the hard-evidence channel operative at both boundaries by Theorem 6.4(2), hence sufficient for global resilience. Likewise, the set $\{(q, p, c, V, \gamma, h) : \gamma\,\Delta x(\underline{\mu}) \geq h \text{ and } \gamma\,\Delta x(\overline{\mu}) \geq h\}$ makes the soft channel operative at both boundaries by Theorem 6.4(1). If both inequalities fail in neighborhoods of the boundaries, Theorem 6.4(3) implies local non-breakability and thus failure of global resilience. □

B Simulation Appendix

This appendix sketches a simple Monte Carlo framework to visualize the *resilience frontier*, the investigation cutoff, and breakability probabilities under both binary and continuous actions (with coarsening). The code can be implemented in any language; we describe the logic and provide figure placeholders.

B.1 Design goals and metrics

We track three summary objects:

- **Breakability indicator:** whether a wrong cascade (lower or upper) is overturned within $T_{\max}$ periods.
- **Time to correction:** the (random) time until beliefs exit the cascade region and converge to the truth (via disclosure or soft inference).
- **Investigation rate:** the fraction of periods (or agents) who choose to investigate.

Unless noted, we initialize at the boundary belief $\mu_1 \in \{1-q,\ q\}$ and condition on the wrong state to focus on resilience in the bottleneck region.

B.2 Environment and primitives

Parameters: signal precision $q \in (1/2, 1)$; verification success $p \in (0, 1]$; investigation cost $c \geq 0$; discovery rent $V \geq 0$; responsiveness $\gamma \in [0, 1]$; coarsening step $h \in [0, 1]$; horizon cap $T_{\max} \in \mathbb{N}$; simulations $N \in \mathbb{N}$. Posteriors after signals: μ^+ and μ^- as in Section 3. Investigation decision uses Proposition 4.4. For continuous actions, $a^* = \gamma x + (1-\gamma)\mu$; the public observes $\tilde{a} = \mathcal{C}(a^*)$ by rounding to the nearest grid of step h.

B.3 Binary-actions simulation

Inputs: $(q, p, c, V, T_{\max})$, initial belief μ_1 at the wrong boundary, true state $\theta^\star \in \{0, 1\}$ (opposite of boundary-implied action).
Loop for $t = 1, \ldots, T_{\max}$**:**

1. Compute posteriors μ_t^+ and μ_t^-; draw signal $s_t \in \{0, 1\}$ with $\Pr(s_t = 1 \mid \theta^\star) = q$.
2. Set $x \leftarrow \mu_t^+$ if $s_t = 1$, else $x \leftarrow \mu_t^-$.
3. *Investigate decision:* evaluate $U^{\text{inv}}(x) - U^{\text{no}}(x)$ from Lemma 4.1; investigate iff ≥ 0.
4. If investigate: draw $E \sim \text{Bernoulli}(p \cdot \mathbf{1}\{\theta^\star = 1\})$.
   - If $E = 1$: disclose; set $\mu_{t+1} \leftarrow 1$ and `stop` (correction).
   - If $E = 0$: update posterior to $x' = \frac{x(1-p)}{1-xp}$; choose accuracy-maximizing action based on x'; update μ_{t+1} from observed action (Bayes, given equilibrium mapping).
5. If not investigate: choose accuracy-maximizing action based on x; update μ_{t+1} accordingly.

Outputs: indicator of correction within $T_{\max}$; time to correction (if any); investigation count.

B.4 Continuous-actions with coarsening

Inputs: $(q, p, c, V, \gamma, h, T_{\max})$, initial μ_1 at the wrong boundary, state $\theta^\star$.
Loop for $t = 1, \ldots, T_{\max}$**:**

1. Draw s_t and compute x as above.
2. Compute $a^* = \gamma x + (1-\gamma)\mu_t$; publish $\tilde{a} = \mathcal{C}(a^*)$ by rounding to grid step h.
3. (Optional) Investigate using the same rule as in the binary case; if disclosure occurs, set $\mu_{t+1} \leftarrow 1$ and `stop`.
4. *Soft update from actions:* compute $\tilde{a}^+ = \mathcal{C}(\gamma\mu_t^+ + (1-\gamma)\mu_t)$ and $\tilde{a}^- = \mathcal{C}(\gamma\mu_t^- + (1-\gamma)\mu_t)$.
   - If $\tilde{a}^+ \neq \tilde{a}^-$, use Bayes with likelihoods implied by $\{\tilde{a}^+, \tilde{a}^-\}$ to update μ_{t+1}.
   - If $\tilde{a}^+ = \tilde{a}^-$, treat the action as uninformative and keep $\mu_{t+1} = \mu_t$ (or apply a tiny tie-breaking perturbation if desired).

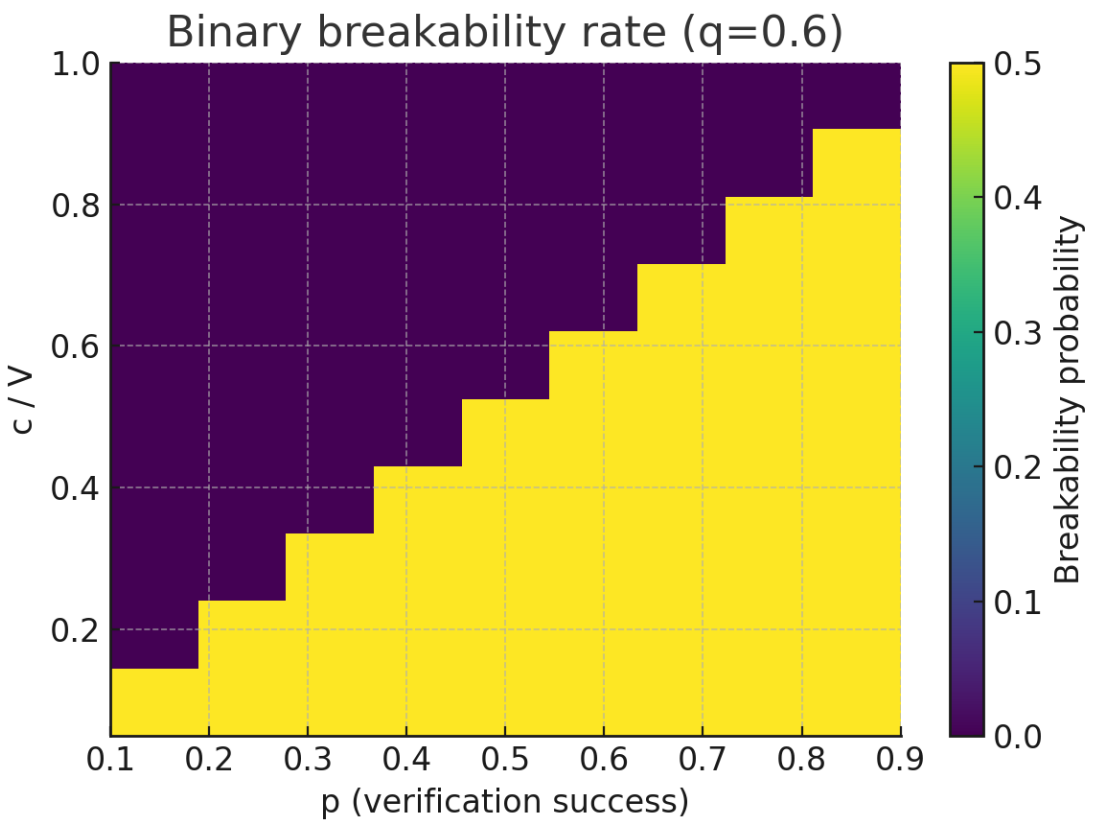

Figure 2: Binary-actions simulation: breakability probability over $(p,\ c/V)$ at $q = 0.6$. Brighter areas indicate higher probability that a wrong cascade at the boundary is overturned within the simulation horizon.

B.5 Parameter grids and outputs

A compact grid suffices for informative figures:

$$q \in \{0.55, 0.6, 0.65, 0.7\}, \quad p \in \{0.1 : 0.1 : 0.9\}, \quad V \in \{0.5, 1, 2, 4\},$$
$$c \in \{0.05 : 0.05 : 1.0\}, \quad \gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}, \quad h \in \{0, 0.02, 0.05, 0.1, 0.2\}.$$

For each grid point, run N simulations per boundary and report: breakability rate, mean time to correction (conditional on correction), and investigation rate. Use a fixed random seed for reproducibility.

Figure 2 plots the simulated breakability probability in the binary benchmark. Consistent with Corollary 4.6 and the resilience frontier in Corollary 6.5, breakability is highest where verification is effective (high p) and relatively cheap (c/V low).

Figure 3 fixes $\gamma = 0.6$ and $h = 0.05$. The pattern mirrors Proposition 5.1: when actions are sufficiently responsive and observable, the soft channel helps correction; otherwise, the hard-evidence channel (higher p or lower c/V) dominates.

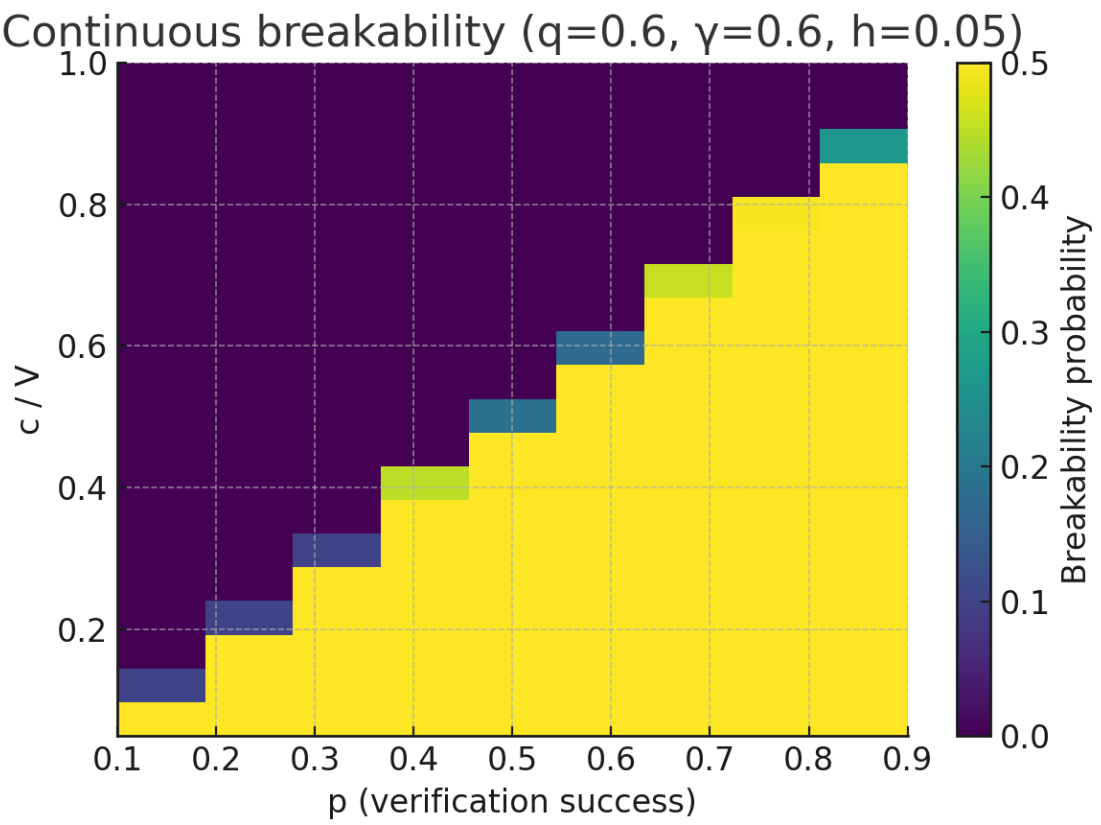

Figure 3: Continuous actions with conformity and coarsening ($\gamma = 0.6$, $h = 0.05$): breakability probability over $(p,\ c/V)$ at $q = 0.6$. Brighter areas indicate higher probability that a wrong cascade is overturned via soft inference or disclosure.

B.6 Reproducibility notes

Use a fixed seed (e.g., `seed=1729`). For stability, set $T_{\max}$ large enough that absorbing states (disclosure or near-one beliefs) are reached in most corrected runs; report the fraction hitting the cap. To avoid numerical underflow in Bayes updates near boundaries, work in log-likelihood ratios when implementing $\mu^{\pm}$.

B.7 Pseudocode

```
function simulate_binary(mu0, theta, q, p, c, V, Tmax, seed):
    set RNG(seed); mu = mu0
    for t in 1..Tmax:
        s = Bernoulli(q) if theta==1 else Bernoulli(1-q)
        x = mu_plus(mu,q) if s==1 else mu_minus(mu,q)
        gain = U_inv_minus_U_no(x, p, c, V)      # Lemma 1
        if gain >= 0:
            E = Bernoulli(p) if theta==1 else 0
```

```
            if E==1: return (correct=1, t, investigated=1)
            x = x*(1-p)/(1 - x*p)                # failure posterior
        # choose action to match state given x; update mu via Bayes
        a = 1 if x>=0.5 else 0
        mu = update_from_action(mu, a, q)        # equilibrium mapping
    return (correct=0, t=Tmax, investigated=...)
```

A continuous-actions variant replaces the action step by $a^* = \gamma x + (1-\gamma)\mu$ and the update rule by comparing coarsened actions $\tilde{a}^{\pm}$.

B.8 Deliverables

The repository should include: (i) a script to run the grid and save CSV outputs; (ii) plotting scripts that generate the four figures above; (iii) a `README` with exact parameter choices, seed, and runtime notes; (iv) a `repro` target that re-creates all figures from a clean environment.

This appendix is self-contained: once the investigation cutoff and the soft distinguishability condition are coded, the remaining logic follows directly from Sections 4 and 5.

Robustness & reproducibility (brief note). All results are reproducible with the public scripts referenced in the Data and Code Availability section. In robustness checks (reported in the replication package), we vary the primary-source domain list, use alternative stance lexicons, and run in-time and in-space placebo events; conclusions are unchanged. A deterministic pipeline and session lockfiles are included in the replication package; full archives will be posted upon submission.

Data and Code Availability

All simulation code and figure scripts used in this paper are provided in the replication package (`cascades_sim/`); figures in the paper can be regenerated from clean runs.

Acknowledgments

We thank Alexey Verenikin, Emiliano Catonini, Markus Gebauer, Khromova Ella, Steven Kivinen, and Eugenia Andreev for helpful comments and discussions, as well as participants of the ICEF research seminar for valuable feedback. All remaining errors are our own.

Conflicts of Interest

The authors declare no competing interests.

Ethics Statement

This study uses only publicly available data and simulation code; no human subjects were involved.

References

[1] Arieli, I. (2021). A general analysis of sequential social learning. *Mathematics of Operations Research*, 46(4):1235–1249.

[2] Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, 107(3):797–817.

[3] Bénabou, R. and Vellodi, N. (2025). (pro-)social learning and strategic disclosure. *American Economic Journal: Microeconomics*. Forthcoming.

[4] Bertomeu, J. and Cianciaruso, D. (2018). Verifiable disclosure. *Economic Theory*, 65(4):1011–1044.

[5] Bikhchandani, S., Hirshleifer, D., Tamuz, O., and Welch, I. (2022). Information cascades and social learning. Working paper, August 18, 2022.

[6] Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

[7] Bursztyn, L. and Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153.

[8] Chamley, C. (2004). *Rational Herds: Economic Models of Social Learning*. Cambridge University Press, Cambridge.

[9] Clayton, K. et al. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42:1073–1095.

[10] Frick, M., Iijima, R., and Ishii, Y. (2020). Misinterpreting others and the fragility of social learning. *Econometrica*, 88(6):2281–2328.

[11] Hann-Caruthers, W., Martynov, V. V., and Tamuz, O. (2018). The speed of sequential asymptotic learning. *Journal of Economic Theory*, 173:383–409.

[12] Lobel, I. and Sadler, E. D. (2015). Information diffusion in networks through social learning. *Theoretical Economics*, 10(3):807–851.

[13] Porter, E. and Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118.

[14] Smith, L. and Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*, 68(2):371–398.

[15] Xu, W. (2023). Social learning through action-signals. SSRN working paper.

[16] Xu, W. (2025). Social learning through coarse signals of others' actions. *Journal of Economic Theory*. In press; journal pre-proof.