Aicumene:2026.047v1 [econ.TH] October 2026

Multi-Agent Verification Games: Shared Sources, Crowding Out, and the Price of Independence

Aicumene Preprint

Alexandra Bernadotte
Aicumene*

Ivan Menshikov
Aicumene

October 2026 — working paper, Part I (SSRN preprint)

Abstract

When several agents confirm the same claim, their agreement is ordinarily taken as stronger evidence of its truth. But what if their independence is only procedural, while their evidence originates from the same source, and what if every agent acts rationally yet no one has an incentive to obtain genuinely independent evidence? We study these questions by connecting the epistemology of verification with the game theory of costly information acquisition in a multi-agent system. The primitive c is the conditional fraction of erroneous claims that remain observationally indistinguishable from correct claims to source-sharing verifiers; it separates the number of verifications from the number of independent evidentiary origins. We show that c imposes an irreducible floor on the false-accept/false-reject frontier of every decision rule based on such verdicts, and that the floor persists on arbitrary copying structures: cross-checking any number of registries that copy one root leaves it unchanged. Verification networks need not be trees: on a cycle of delegation agents confirm one another's records, and we count the architectures that avoid this. Acyclic orientations of a consultation graph are at most $n!$ of 2^g; unique-sink orientations of the h-cube, in which every sub-community has one responsible agent, number $h^{\Theta(2^h)}$ and may still contain a loop; and when each agent goes to the origin with probability θ the architecture is loop-free with probability tending to $\theta e^{1-\theta}$. The central economic result concerns a simultaneous procurement game in which n receivers choose between a cheap shared-source verification and a costlier source-independent one. We classify its pure equilibria and identify a nonempty region, $pcL < d < npcL$ with d the independence premium, in which every pure-strategy Nash equilibrium purchases shared-source verification although source-independent verification minimizes aggregate expected cost; we bound the own error the result tolerates, derive a targeted subsidy that implements the efficient equilibrium, and confirm the conditions by exhaustive enumeration. The floor is illustrated on a chain of copies, the paper's epigraph, and on the verification record of a machine-found theorem. The findings separate strategic rationality from epistemic reliability: agents may individually optimize while collectively preserving an avoidable source-dependent error, so that independence is an economic and architectural resource and not only a statistical property of evidence. These are conditional theoretical results rather than empirical claims about deployed AI verifiers.

Keywords: multi-agent systems; social epistemology; evidentiary independence; costly verification; information economics; common-cause error; correlated verifiers; public goods; Nash equilibrium; source independence; verification networks; unique-sink orientations; artificial intelligence.
JEL: D82, D83, H41.

*Aicumene: `https://aicumene.com/`; this paper: `https://research.aicumene.com/papers/game-theory-1/`. Corresponding author: bernadotte@aicumene.com.
© 2026 Alexandra Bernadotte and Ivan Menshikov. Licensed under CC BY-NC-ND 4.0.

A monastery has copied its sacred texts by hand for centuries: each copy is made from the previous copy, and the original rests untouched in the vault. A new novice asks the abbot whether he may check his copy against the original. "We have copied from the copy for a thousand years," says the abbot. The novice insists. "Very well, I will check it myself," says the abbot, and goes down to the vault. Days pass. At last he comes back up in tears, repeating: "Celebrate, not celibate!"

— a folk joke on source dependence

Aicumene:2026.047v1 [econ.TH] October 2026

1 Introduction

1.1 When agreement fails to establish truth

Suppose that several intelligent agents are asked to verify a factual claim. Each examines the available evidence, produces a verdict, and reports that the claim is correct. The agents may differ in their computational architectures, decision procedures and institutional affiliations, even by the structural criteria that separate strong from imitative intelligence [17], and their agreement appears to constitute multiple independent confirmations. Yet this conclusion may be unwarranted. All agents may have consulted databases derived from the same original record; their reasoning procedures differ, but their evidence has a common ancestor, and if the original record is incorrect every subsequent verification reproduces the error without detecting it. The agreement is real, the procedures were executed correctly, and each agent may even have acted optimally given its information. The collective conclusion is nonetheless false.

This illustrates a distinction between three properties that are easily treated as interchangeable. Agreement concerns the similarity of conclusions. Independence concerns the provenance and the conditional informational content of the evidence behind those conclusions. Truth concerns the claim itself. The distinction matters most in multi-agent artificial intelligence, where an agentic system delegates research, verification, criticism and final approval to different models or specialized components. Such a system appears to introduce redundancy, but redundancy in computational procedures need not imply redundancy in evidentiary origins: a system of ten verifiers may possess less independent evidence than one verifier consulting a genuinely separate source. The relevant question is not how many agents have checked a claim, but what information their checks have introduced.

1.2 Verification as an epistemic and economic problem

The limitation of shared evidence is only the first part of the problem. Independent verification may be available but costly: it may require acquiring an original document, reconstructing provenance, obtaining a separate dataset, commissioning an external audit, or performing a new observation. Who should bear this cost? When the result of a verification is available to several agents, independently acquired evidence is a public good: one agent pays for it and every other agent benefits, so an agent's willingness to pay for independent verification differs from its value to the group. This creates the possibility of a specific failure: every agent behaves rationally according to its own incentives, yet the resulting equilibrium preserves an error that the system could have eliminated at lower aggregate expected cost. The failure is not caused by irrational agents, defective algorithms or insufficient computation; it arises from the structure of the verification game itself.

This motivates the central question of the paper: *under what informational and economic conditions can individually rational verification produce a collectively unreliable outcome?* We answer in two steps. First, we characterize the information that source-sharing verifiers cannot recover, regardless of the number of checks they perform. Second, we examine the strategic incentives that determine whether agents acquire the independent evidence needed to overcome this limitation. The connection between the steps is essential: an information-theoretic limitation describes what verification cannot establish from a given collection of sources, and a game-theoretic model explains why strategically interacting agents may fail to obtain better sources even when doing so would improve collective welfare.

1.3 From individual rationality to collective reliability

Nash equilibrium is a statement about incentives, not about truth: it describes a configuration in which no player can improve its payoff by a unilateral change of strategy, and it implies neither that the resulting

Aicumene:2026.047v1 [econ.TH] October 2026

information is accurate, nor that the verification is epistemically independent, nor that the equilibrium minimizes the total cost of error. In the present setting the difference can be quantified. We define the independence premium d as the additional cost of obtaining a verdict from a source that does not inherit the relevant common error; the equilibrium procurement decision depends on the private benefit of paying d, the socially efficient decision on the benefit to all receivers, and we identify the conditions under which the two diverge. This yields a precise sense in which strategic rationality can coexist with epistemic inefficiency: rational agents may purchase verification without purchasing the independence that would make it collectively valuable. The results do not imply that agreement is worthless, that shared sources are always unreliable, or that independent checks guarantee truth; they identify a specific limit of source-dependent evidence and show how strategic incentives can preserve it. The broader problem is one of institutional and computational design: a reliable verification architecture must determine not only who checks a claim, but which evidentiary dependencies are acceptable, who is responsible for obtaining new information, and how the cost of independence is distributed.

Contributions and prior work. Costly state verification [7] and public-good provision [30] are established frameworks, and jury aggregation with correlated votes [14] is closely related; this paper claims no priority for those general insights, and Section 9 places the results among their neighbours. What it adds is a specific source-dependent boundary and procurement wedge under explicit assumptions: the complete classification of strict-cost equilibria, the endogenous choice of the source technology as a reduced-form technology choice, a bound on the own error of the verifiers that the results tolerate, and a count of the verification architectures in which every chain of delegation reaches an origin, which turn out to be rare.

Road map. Section 2 fixes the vocabulary: agreement, truth, and evidentiary independence relative to an error mechanism. Section 3 states the floor (Theorem 1), its posterior form and its probabilistic-truth reading, extends it to source trees (Proposition 1) and illustrates it on a chain of copies (Example 1). Section 4 turns to architectures: a lemma on error propagation in loops, the count of bottom-up architectures (Proposition 2), and the verification record of a machine-found theorem. Section 5 defines the procurement game and classifies its pure equilibria (Theorem 2); Section 6 proves the crowding-out wedge (Theorem 3), with the epigraph as Example 2. Section 7 bounds the own error the results tolerate, gives the subsidy that implements the efficient equilibrium, and reports the exhaustive enumeration. Section 8 draws the epistemic lesson, Section 9 places the results among their neighbours and states the research plan, and Section 10 concludes.

2 Epistemic foundations: provenance and independence relative to an error mechanism

Three properties of a verified claim are easily conflated. *Agreement* is a property of verdicts: they coincide. *Truth* is a property of the claim: it corresponds to the facts, with the operational status τ of Remark 1 as its measurable surrogate. *Independence* is a property of evidence: it concerns where the evidence comes from and what it can distinguish. The primitive of this paper is a precise version of the third. Truth and the probability assigned to truth are moreover different objects: in the binary model the truth value $T(\varphi) \in \{0, 1\}$ of a claim φ is fixed, whereas $\tau(\varphi \mid \mathcal{E}, K, t) = \Pr(T(\varphi) = 1 \mid \mathcal{E}, K, t)$ is epistemic confidence conditional on the available evidence. Our results constrain the latter, not the ontological status of truth; the truth-probability cap of Remark 1 is an upper bound on the posterior confidence obtainable from a specified class of source-sharing verdicts under the model assumptions.

Definition 1 (Evidentiary indistinguishability relative to an error mechanism)**.** *Let W denote a wrong claim, R a correct claim, and E an* error mechanism*, an event on which the error has a particular origin. A verifier i is* individually indistinguishing *with respect to E if the conditional distribution of its verdict V_i under (W, E) coincides with that under R. A collection of verifiers is* jointly indistinguishing *with respect to E if the conditional joint distribution of its verdict vector V under (W, E) coincides with that under R. A verifier or collection that is not indistinguishing is* evidentiarily independent *of E. The information floor of Theorem 1 requires joint indistinguishability; individual indistinguishability alone is not sufficient without an assumption on the dependence structure of the verdicts. The share $c = \Pr(E \mid W)$ of errors on which the collection is jointly indistinguishing is the primitive of the paper.*

Aicumene:2026.047v1 [econ.TH] October 2026

When the verifiers' own errors are conditionally independent under both states, with identical component laws under R and (W, E), individual indistinguishability implies joint indistinguishability; this is the case of Section 3. Without this condition the dependence between otherwise uninformative verdicts can itself carry information about the claim, which is why Theorem 1 is stated for the joint law and why not every correlation of sources is sufficient for the floor. The definition is relative to the mechanism: a registry is independent of errors that originate downstream of it and indistinguishing on errors that originate at its root (Proposition 1). It is also distinct from four notions with which it is often confused. *Causal* independence, the absence of a common cause, is sufficient but not necessary, since a common cause that does not reach the verdicts leaves the verifiers evidentiarily independent. *Statistical* independence of the verifiers' own errors is neither necessary nor sufficient for distinguishing: the shared verifiers of Section 3 have conditionally independent own errors and are nonetheless jointly indistinguishing on E, while two individually indistinguishing verifiers with dependent own errors may distinguish through that dependence. *Methodological* independence, different algorithms or model families, and *institutional* independence, different organizations or labels, do not enter the definition at all; that only the law of the verdict on (W, E) matters is the content of Theorem 1. Verification, finally, is a procedure, whereas justification is a relation between a claim and its evidence: a correctly executed procedure whose evidence is dependent on E adds nothing about E, however many times it is run. The sections that follow make this precise, and every interpretive statement in them is a reading of a theorem under its stated premises, not a general claim about collective knowledge.

3 Information limits of source-sharing verification

Let R denote a correct claim and W an incorrect one, with $\Pr(W) = p \in (0, 1)$. Conditional on W, event E occurs with probability $c \in [0, 1]$: the error originates in the shared source, the source read by every shared verifier. Each shared verifier returns ok with probability $1 - q_j$ conditional on R and on (W, E), and returns flag with probability $1 - q_j$ conditional on (W, E^c). Its own conditional signal errors are independent across verifiers. An independent-source verifier does not inherit E. A receiver's losses are $L > 0$ for accepting W and $\ell > 0$ for rejecting R. Write $FA = \Pr(\mathsf{accept} \wedge W)$ for the probability of a false accept and $FR = \Pr(\mathsf{reject} \wedge R)$ for that of a false reject. An arbitrary decision rule δ can randomize over the available verdict vector V. Table 1 collects the notation of the paper.

Theorem 1 (Common-source information floor)**.** *If the joint conditional distribution of V under (W, E) equals that under R, then every decision rule using only V satisfies*

$$FA \;\geq\; pc\left(1 - \frac{FR}{1-p}\right). \tag{1}$$

If an unlimited sequence of shared verifiers with independent own errors bounded above by $\bar{q} < 1/2$ is available, the boundary is asymptotically attainable by majority-based randomized decisions.

Proof. Put $a = \Pr(\delta(V) = \mathsf{accept} \mid R)$. Equality of signal laws implies $\Pr(\delta(V) = \mathsf{accept} \mid W, E) = a$. Thus $FA \geq pca$ and $FR = (1-p)(1-a)$, establishing (1). For the asymptotic statement, conditional independence and the uniform accuracy margin imply that majority voting identifies R versus (W, E^c) with probability tending to one. It cannot separate R from (W, E). Accepting the majority-ok result with arbitrary probability a approaches $((1-p)(1-a), pca)$. □

For a vector of shared verifiers the hypothesis of the theorem holds by construction: each returns ok with the same probability under R and under (W, E), and own errors are independent across verifiers. It is stated as a hypothesis because Proposition 1 applies the same argument to other events. The proof is the contamination bound of robust hypothesis testing [1], restated for a verification frontier: when the alternative is a mixture that contains the null with weight c, no test can have type-II error below c times its acceptance rate under the null. The floor depends on *informational indistinguishability*, not merely a positive correlation coefficient. A sufficiently informative independently sourced signal can break it. At zero-probability verdict events posterior probabilities are undefined; any posterior lower-bound claim must be restricted to profiles of positive probability.

Aicumene:2026.047v1 [econ.TH] October 2026

Table 1: Notation.

Symbol	Meaning
R, W, p	the claim is correct or wrong; $p = \Pr(W)$
E, c	the error is inherited from the shared source; $c = \Pr(E \mid W)$, the primitive of the paper
q, q_j	own error of a verifier, independent across verifiers
ok, flag, V, δ	verdicts, the verdict vector, a possibly randomized decision rule
FA, FR	false-accept and false-reject probabilities $\Pr(\mathsf{accept} \wedge W)$ and $\Pr(\mathsf{reject} \wedge R)$
L, ℓ	loss from accepting a wrong claim and from rejecting a correct one; Assumption 1 is $pL < (1-p)\ell$
$x(v)$, $c_{\mathrm{obs}}(V)$	likelihood ratio $\Pr(v \mid W, E^c)/\Pr(v \mid R)$; observational floor, the finite-sample infimum of $\Pr(v \mid W)/\Pr(v \mid R)$ (Corollary 1)
τ, $\bar{\tau}$	truth probability of a claim; its cap $(1-p)/(pc+1-p)$ under shared verification (Remark 1)
$c_{\mathrm{src}}(\mathcal{V})$, σ, ρ	structural floor, the share of errors originating among the common ancestors of the sources read by $\mathcal{V}$; sources; the origin (Proposition 1)
D, G, g, Z, θ	verification digraph; consultation graph and its number of edges; a directed cycle; share of agents who go to the origin (Lemma 1, Proposition 2)
Q_h	the h-dimensional cube, the consultation graph of agents indexed by h binary attributes
N, S, I; κ_S, κ_I, d	no check, shared check, independent check; their prices and the independence premium $d = \kappa_I - \kappa_S$
n, Γ_n, a_i, u_i	number of receivers, the procurement game, actions and payoffs (Definition 2)
$\mathcal{L}(\cdot)$, $\mathcal{W}$	decision loss given the strongest check bought, pL, pcL or 0; welfare cost of the equilibrium, $npcL - d$
A, B, r, q^*, $\mathcal{C}$	$(1-p)\ell - pL$, $(1-p)\ell - pcL$, $\sqrt{B/(p(1-c)L)}$, $1/(1+r^2)$, $(1-p)\ell + pL(1-2c)$ (Section 7)
R_k, $R_{k,k'}$	Bayes risk with k shared and k' independent verdicts
m, Δ, M, $\bar{q}$	deviation margin, welfare margin, $\max\{L, \ell\}$, robustness radius (Proposition 4)
q_{dup}, q_{SN}, q_{II}, q_{SS}, q_{IS}, q_{eq}, q_{opt}, q_{end}	own-error thresholds of Section 7
s	subsidy to the first independent purchaser (Section 7.1)
T, ε, y	number of copyings, per-copy error, years between copies (Examples 1 and 2)

Corollary 1 (Posterior floor)**.** *For every verdict profile v of shared verifiers with $\Pr(v) > 0$, write $x(v) = \Pr(v \mid W, E^c)/\Pr(v \mid R) \in [0, \infty]$. Then*

$$\Pr(W \mid v) = \frac{pc + p(1-c)\,x(v)}{(pc+1-p) + p(1-c)\,x(v)} \;\geq\; \frac{pc}{pc+1-p}, \tag{2}$$

with equality iff $x(v) = 0$. For k shared verifiers with a common own error $q \in (0, \frac{1}{2})$ the minimum over profiles is attained at the all-ok profile, where $x = (q/(1-q))^k$; it is strictly above $pc/(pc+1-p)$ for every finite k and decreases to it as $k \to \infty$. At $q = 0$ the floor is attained already by the ok verdict of a single verifier.

Proof. Under the hypothesis of Theorem 1, $\Pr(v \mid W) = c\Pr(v \mid R) + (1-c)\Pr(v \mid W, E^c)$, and Bayes' rule gives (2); if $\Pr(v \mid R) = 0$ then $x = \infty$ and $\Pr(W \mid v) = 1$. The middle expression is increasing in x and equals $pc/(pc+1-p)$ at $x = 0$. For k verifiers and a profile with u verdicts ok, $x = (q/(1-q))^{2u-k}$, which is smallest at $u = k$ because $q/(1-q) < 1$; it is positive for $q > 0$ and tends to 0 as $k \to \infty$. □

Equivalently, the hypothesis of Theorem 1 says that the likelihood ratio $\Pr(v \mid W)/\Pr(v \mid R) = c + (1-c)x(v)$ never falls below c: shared verifiers can multiply the odds in favour of a claim by at most $1/c$. For a finite verdict vector the infimum $c_{\mathrm{obs}}(V) = \inf_v \Pr(v \mid W)/\Pr(v \mid R) = c + (1-c)\big(q/(1-q)\big)^k$ is strictly larger than the structural c whenever $q > 0$, and the floor (1) holds verbatim with c replaced by $c_{\mathrm{obs}}(V)$, since $\Pr(\mathsf{accept} \mid W) \geq c_{\mathrm{obs}}(V)\Pr(\mathsf{accept} \mid R)$ for every rule: $c_{\mathrm{obs}}(V)$ is the tighter bound for the verifiers at hand, and it decreases to the structural c as $k \to \infty$ or equals it at $q = 0$. The two should not be confused:

Aicumene:2026.047v1 [econ.TH] October 2026

c, and its source-tree form c_{src} of Proposition 1, is a structural floor, a primitive of the source structure; $c_{\mathrm{obs}}(V)$ is the observational floor, a property of a finite verdict vector, and under joint indistinguishability $c_{\mathrm{obs}}(V) \geq c_{\mathrm{src}}$. An informative independently sourced signal breaks the floor: one independent verifier with the rule "accept iff ok" gives $FA = pq$ and $FR = (1-p)q$, both of which vanish as $q \to 0$, and k' independent verifiers under majority rule drive both to zero as $k' \to \infty$ for any fixed $q < \frac{1}{2}$. Figure 1 shows the achievable sets.

Remark 1 (Probabilistic-truth reading)**.** *In the language of* probabilistic truth*, the operational epistemic status $\tau(\varphi \mid \mathcal{E}, K, t) = \Pr(\varphi \text{ true} \mid \mathcal{E}, K, t)$ of a claim φ given evidence $\mathcal{E}$, knowledge context K and time t [20, 21] (written $\tau(q \mid E, K, t)$ there; the letters are changed because q and E denote here the own error and the shared-source event), Corollary 1 says that no number of source-sharing verifications can raise the truth probability of an accepted claim above*

$$\bar{\tau} \;=\; 1 - \frac{pc}{pc + 1 - p} \;=\; \frac{1-p}{pc+1-p},$$

a bound on posterior confidence, not on the truth value (Section 2). The shared source caps the evidentiary support that verification can add; only an independently sourced signal lifts the cap, and under majority rule over k' independent verifiers the truth probability of an accepted claim tends to one. Theorem 1 is therefore a bound on truth-probability propagation through correlated verifiers, and the procurement game of Section 5 asks who pays for the independent evidence that moves τ above $\bar{\tau}$.

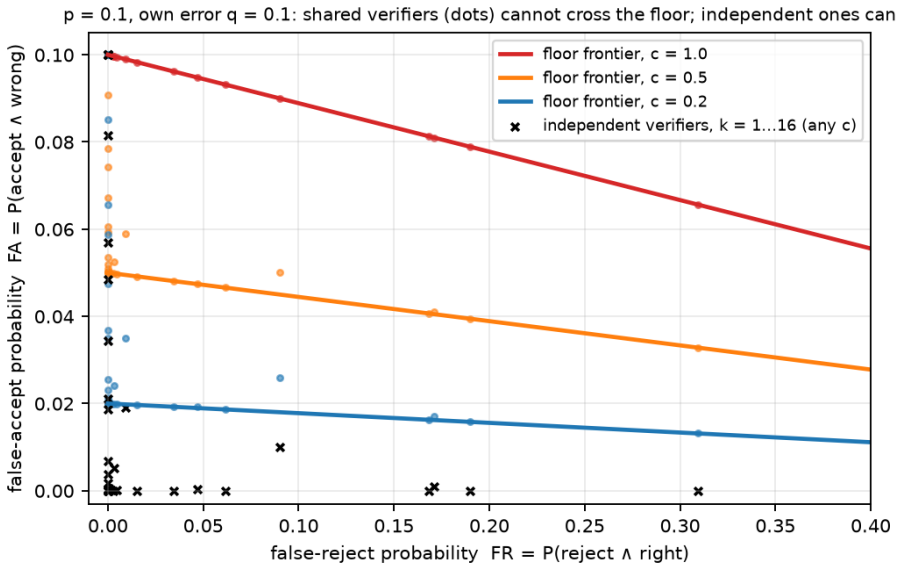

Figure 1: Achievable (false reject, false accept) pairs. Dots: shared verifiers, $k = 1, \ldots, 16$, every threshold rule; lines: the frontier (1) for $c = 1$, 0.5 and 0.2, top to bottom; crosses: independent verifiers. $p = 0.1$, $q = 0.1$. Generated by `sim_vgame.py`.

Source trees. The floor is stated for one shared source; it extends to any copying structure among sources.

Proposition 1 (Source trees and institutional entrenchment)**.** *Let sources form a directed acyclic graph in which an edge $\sigma \to \sigma'$ means that σ' copies σ. A claim's error originates at one node; a verifier reading σ inherits it iff the origin is σ or an ancestor of σ. For a set $\mathcal{V}$ of verifiers let $c_{\mathrm{src}}(\mathcal{V}) = \Pr(\text{the origin is a common ancestor of all sources read by } \mathcal{V} \mid W)$. Assume additionally that conditional on any error copied into every selected source, the entire vector of verdicts has the same joint distribution as on a correct claim (including its own-error dependence structure). Then* (1) *holds with c replaced by $c_{\mathrm{src}}(\mathcal{V})$;*

in the base model of one shared source, $c_{\mathrm{src}}(\mathcal{V}) = c$ for every $\mathcal{V}$. In particular, if every source copies one root and the error originates at the root with probability c_0, then $c_{\mathrm{src}}(\mathcal{V}) = c_0$ for every $\mathcal{V}$: cross-checking any number of downstream registries leaves the floor unchanged.

Proof. Replace E by the event that the origin is a common ancestor of all sources read; on it every verifier in $\mathcal{V}$ reads the error and, by the additional distributional assumption, the joint law of V under (W, E) equals its law under R and the proof of Theorem 1 applies verbatim. □

The additional assumption is not implied by common ancestry: a verifier may read a copied error and still detect it from evidence outside the source, in which case the error is inherited but not invisible. The primitive the floor needs is the share of errors that are both inherited and undetected, which is what the snapshot design of Section 9.2 measures. A record that appears in three registries has been confirmed three times only if the registries have three origins. This is the formal content of *institutional entrenchment* of an error in an administrative system.

Example 1 (The epigraph as a chain of copies)**.** Let the text of the epigraph be recopied every y years, each copying corrupting a given word with probability ε independently of the others, and let the original rest in the vault. After a millennium $T = 1000/y$ copies separate the current copy from the original: the source tree is a path, and the claim "the rule says *celibate*" is wrong with probability $p_T = 1 - (1-\varepsilon)^T$. A check against the exemplar, the previous copy, is a shared-source verification in the sense of Theorem 1: it sees only an error introduced in the last copying, and every older error is a common ancestor of the claim and of the exemplar, so

$$c_T \;=\; \frac{1-(1-\varepsilon)^{T-1}}{1-(1-\varepsilon)^{T}} \;=\; \frac{T-1}{T} + O(\varepsilon). \tag{3}$$

A check against the copy k generations back has $c = (T-k)/T + O(\varepsilon)$ by Proposition 1; the vault is $k = T$ and $c = 0$. For $y = 25$ and $\varepsilon = 10^{-3}$, $T = 40$, $p = 0.039$ and $c = 0.975$: a reassuring comparison with the exemplar moves the probability of error from 0.039 to 0.038, and the truth probability of the accepted rule is capped at $\bar{\tau} = 0.962$; for $\varepsilon = 10^{-2}$, $p = 0.33$ and $c = 0.98$. Figure 5(a) shows p_T and c_T against the age of the tradition: the older it is, the likelier the error and the less an internal check is worth, $p_T \to 1$ and $c_T \to 1$. The example continues in Example 2.

4 Verification architectures: delegation cycles and the count of bottom-up designs

Proposition 1 assumes that sources form a directed acyclic graph, so that every chain of copying ends at an origin. Verification networks need not: an agent who verifies through an agent "above" her may, through a chain of such delegations, be verifying through herself. Let the *verification digraph* D on n agents have an arc $i \to j$ when i verifies through j; agents with no outgoing arc verify at the origin, the vault of the epigraph, and are the sinks of D. Call an architecture *bottom-up* if every agent's chain of delegations reaches a sink, that is, if D is acyclic. On a directed cycle Z every agent is an ancestor of every other. Lemma 1 states the premises under which this makes mutual confirmation worthless: when records and errors are copied along the arcs, the floor applies to $\mathcal{V} = Z$ with $c_{\mathrm{src}}(Z) = \Pr(\text{origin} \in Z \mid W)$, and a closed loop, in which all errors are its own, has $c_{\mathrm{src}}(Z) = 1$. Figure 2 shows the three structures of this paragraph. The h-cube used below is the consultation graph of agents indexed by h binary attributes (department, region, tier, ...): two agents may consult each other when they differ in exactly one attribute, and a subcube is the sub-community that shares the remaining attributes. The question is how many architectures avoid loops, and the answer is: few.

Lemma 1 (Error propagation on a loop)**.** *Let every agent hold a record, and let an arc $i \to j$ of D mean two things: i's record is a copy of j's, and i verifies a claim by comparing it with j's record, with own error q and with a verdict that depends on the claim only through that record. Let the error of a wrong claim originate in the record of one agent and be copied along arcs into the record of every agent whose chain of delegations passes through the origin. Then, conditional on the origin lying on a directed cycle Z of D, every record of Z contains the error, the verdict vector of the agents of Z has the same joint law as on a correct claim, and Theorem 1 holds for $\mathcal{V} = Z$ with $c_{\mathrm{src}}(Z) = \Pr(\text{origin} \in Z \mid W)$. If every error of the claims handled on Z originates on Z, then $c_{\mathrm{src}}(Z) = 1$ and no number of checks inside the loop detects anything.*

Aicumene:2026.047v1 [econ.TH] October 2026

Proof. On a directed cycle every agent's chain passes through every other agent, so an error originating at any agent of Z is copied into all records of Z. Each agent of Z compares the claim with a record that carries the same error, and her verdict depends on the claim only through that record; hence her verdict is distributed as on a correct claim, and the verdicts of Z are independent given the state by the own-error assumption. This is the hypothesis of Theorem 1 on the event that the origin lies on Z, whose conditional probability given W is $c_{\mathrm{src}}(Z)$. □

Without these premises a cycle is only a cycle: agents who verify by independent means while delegating bookkeeping to one another inherit nothing. The premises are those of the epigraph, copying from the copy, and of the registries of Proposition 1; Proposition 2 below is a statement about the graph alone.

Proposition 2 (Cyclic entrenchment and the count of bottom-up architectures)**.** *(i) If every agent verifies through at least one other agent, D contains a directed cycle. (ii) Let G be the graph of who may consult whom, with n agents and g edges, and let an architecture be an orientation of G. The number of bottom-up architectures is $|\chi_G(-1)|$, where χ_G is the chromatic polynomial [38]; it is at most $n!$, with equality for the complete graph, so the fraction of bottom-up architectures among the 2^g orientations is at most $n!/2^g$. For the h-cube, $n = 2^h$ and $g = h\,2^{h-1}$, the number of bottom-up architectures lies between $(h/2 + O(1))^{2^h}$ and $(h+1)^{2^h}$ [39], so their fraction is at most $\big((h+1)/2^{h/2}\big)^{2^h}$; exactly 14 of 16 orientations are bottom-up for $h = 2$ and 1 862 of 4 096 for $h = 3$. The number of orientations of the h-cube in which every subcube has a unique sink, that is, every sub-community has exactly one top agent, is also $h^{\Theta(2^h)}$ [40]: 12 for $h = 2$ and 744 for $h = 3$; such an orientation can contain a directed cycle. (iii) Suppose each agent verifies either at the origin or through exactly one other agent. For a given set of $k \geq 1$ agents who go to the origin there are $k\,n^{n-k-1}$ bottom-up architectures among $(n-1)^{n-k}$ [41]; summed over all such sets there are $(n+1)^{n-1}$ bottom-up architectures among n^n, a fraction $\frac{1}{n}(1+\frac{1}{n})^{n-1} \to e/n$. If each agent independently goes to the origin with probability θ and otherwise consults a uniformly chosen other agent, the probability that the architecture is bottom-up is*

$$\sum_{k=1}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}\, \frac{k\,n^{n-k-1}}{(n-1)^{n-k}} \;\longrightarrow\; \theta\, e^{1-\theta} \qquad (n \to \infty), \tag{4}$$

which is increasing in θ, equals 0 at $\theta = 0$ and equals 1 only at $\theta = 1$.

Proof. (i) Follow outgoing arcs from any agent; within $n+1$ steps an agent repeats. (ii) The count $|\chi_G(-1)|$ is Stanley's theorem. Every acyclic orientation is induced by each of its topological orders, so the $n!$ linear orders of the agents map onto the bottom-up architectures, bijectively when G is complete. The cube bounds are those of [39] as quoted in [40], the unique-sink count is Theorem 1 of [40], whose Section 1 exhibits a unique-sink orientation of the 3-cube with a directed 6-cycle (Figure 2(c)), and the exact values for $h = 2, 3$ are by enumeration (script `cycles.py`). (iii) The architecture restricted to the $n-k$ delegating agents is a map into the n agents, and it is bottom-up iff it is a rooted forest on the n agents whose root set is exactly the k origin visitors; Cayley's formula for forests with a prescribed root set gives $k\,n^{n-k-1}$ [41], and $\sum_k \binom{n}{k} k\,n^{n-k-1} = (n+1)^{n-1}$ counts all rooted forests. For the limit, $k\,n^{n-k-1}/(n-1)^{n-k} = \frac{k}{n}\big(1+\frac{1}{n-1}\big)^{n-k} \to \theta e^{1-\theta}$ uniformly as $k/n \to \theta$, and k/n concentrates at θ. The derivative of $\theta e^{1-\theta}$ is $(1-\theta)e^{1-\theta} \geq 0$, and $\theta e^{1-\theta} < 1$ for $\theta < 1$ because $\ln\theta < \theta - 1$. □

In words: if nobody goes to the origin, no architecture propagates bottom-up; if one agent in ten does, the system is free of self-confirming loops with probability about 0.25, if one in two, 0.82; and the finite-n probability in (4) is within 0.01 of its limit from $n = 20$ on. The loops trap a minority that shrinks slowly: in simulation the share of agents whose chain never reaches the origin at $\theta = 0.1$ is 0.40 for $n = 50$, 0.21 for $n = 200$ and 0.07 for $n = 1\,000$ (`cycles.py`). Unique responsibility in every unit, the unique-sink property of [40], does not by itself prevent a loop. In the terms of Section 5, where verdicts are public and one origin visit serves all n, θ is the share of agents who pay the independence premium; the abbot's descent to the vault is $\theta > 0$.

Aicumene:2026.047v1 [econ.TH] October 2026

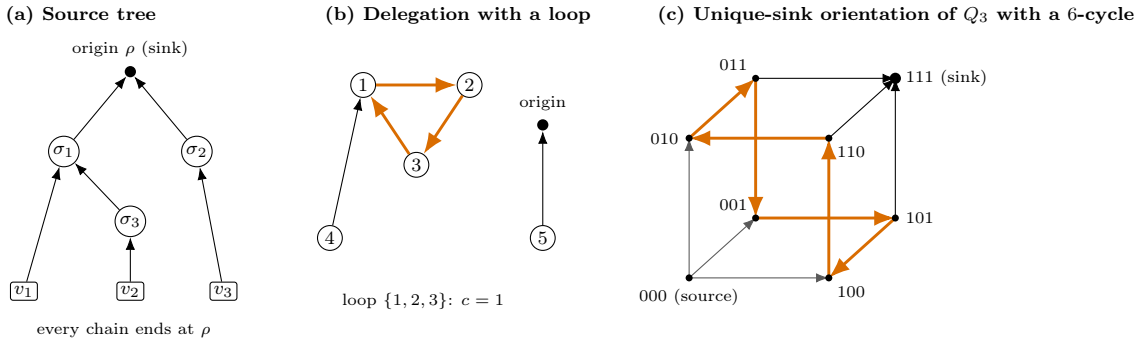

Figure 2: Bottom-up and cyclic verification architectures; arrows point from an agent to the agent or source consulted. (a) A source tree: every chain ends at the origin ρ, the unique sink, and $c_{\mathrm{src}}(\mathcal{V})$ of Proposition 1 is the probability that the error originates among the common ancestors of $\mathcal{V}$, here $\{\rho, \sigma_1\}$ for $\{v_1, v_2\}$ and $\{\rho\}$ for $\{v_1, v_3\}$. (b) The delegation model of Proposition 2(iii): agents $1, 2, 3$ verify only through one another, agent 4 delegates into the loop, agent 5 goes to the origin; the loop is a closed source set with $c = 1$ for its own errors (Lemma 1). (c) A unique-sink orientation of the 3-cube by the matching reversal of [40]: the Boolean orientation, toward more ones, with the matching {010–110, 001–011, 100–101} reversed. Every face has exactly one sink (010, 100, 001 on the three faces through 000, and 111 on the other three) and 111 is the unique sink of the cube, so every sub-community has one responsible top agent; yet the six agents of weight one and two (thick arrows) form a directed cycle and verify only through one another.

An instance from mathematics: Claude's cycles. Proposition 1 is about provenance, not about the number or the fame of the checks, and a recent mathematical record shows the nodes of such a source tree in one place. Knuth [36] documents how a construction found by a language model for an open problem of *The Art of Computer Programming*, the decomposition of the arcs of a digraph on m^3 vertices into three directed Hamiltonian cycles, was verified. That digraph is itself the opposite of a bottom-up architecture: every vertex has three outgoing arcs and no sink, and each channel alone routes the verification of every agent around one loop through all m^3 agents. The first layer of checks ran the model's program for every odd $m \leq 101$; such checks share one source, the program, and cannot see whether a rule that works at the tested sizes works in general. Knuth's own count makes the blind spot explicit: of the 11 502 Hamiltonian cycles for $m = 3$, 1 012 survive the check at $m = 5$, 996 the check at $m = 7$, and exactly those 996 are valid for every odd m, Under a uniform prior over the candidate constructions surviving the stated test, these counts yield posterior invalidity fractions of $10\,506/11\,502 \approx 0.91$ after the $m = 3$ check, $16/1\,012 \approx 0.016$ after the $m = 5$ check and zero after $m = 7$; without a specified sampling distribution they are combinatorial fractions, not unconditional Bayesian posteriors. Read as posteriors they are $x(v)$ of Corollary 1 falling to a floor that here is zero, because the $m = 7$ check separates. The second layer was Knuth's own proof; the third its formalization in Lean by Morrison [37], which Knuth welcomed with the remark that he had become more error-prone, a statement about his own q. Further proofs were produced by language models given Knuth's note as context and by passing text between two models that Knuth calls data-sharing agents. In the terms of Definition 1 the model-generated proofs read overlapping sources, the shared corpora and the same note, so their mutual agreement raises the truth probability of the result only up to the cap $\bar{\tau}$ of Remark 1, while the formal check is a different mechanism of verification, dependent on its own trusted base, the kernel and the formalization of the statement, rather than an absolutely independent component; relative to the error mechanisms of the informal proofs it is the independently sourced check that lifts the cap. Its cost is the independence premium d of Section 5, and whether anyone pays it is the procurement question studied there; in this record a volunteer did.

5 Strategic verification: the procurement game and its pure equilibria

There are $n \geq 2$ receivers of one claim. At procurement stage each chooses N (no purchase), S (one shared-source verification, cost κ_S), or I (one independently sourced verification, cost κ_I). Purchased verdicts

Aicumene:2026.047v1 [econ.TH] October 2026

Aicumene:2026.047v1 [econ.TH] October 2026

are public. The stage-two receiver uses the Bayes-optimal accept/reject action. We assume $q = 0$, strictly positive procurement costs, and

Assumption 1 (Accept by default)**.** *$pL < (1-p)\ell$. Thus absent verification, accepting is optimal; after a shared* ok*, accepting remains optimal.*

Since a perfect independent check fully identifies the truth and a perfect shared check flags only the idiosyncratic wrong claims, individual expected decision losses are

$$\mathcal{L}(N) = pL, \quad \mathcal{L}(S) = pcL, \quad \mathcal{L}(I) = 0, \tag{5}$$

where the argument indicates the strongest technology *purchased by anyone.* Duplicate shared checks do not add information at $q = 0$.

Definition 2 (Simultaneous procurement game Γ_n)**.** Players. *Receivers $i = 1, \ldots, n$ of one claim.* Nature. *The claim is wrong (W) with probability p and right (R) otherwise; conditional on W, the error originates in the shared source (E) with probability c.* Stage 1 (procurement, simultaneous). *Each receiver chooses $a_i \in \{N, S, I\}$ at cost $\kappa(N) = 0$, $\kappa(S) = \kappa_S$, $\kappa(I) = \kappa_I$.* Information. *Every purchased verdict is public. With $q = 0$, an independent verdict reveals the state; a shared verdict reads* flag *exactly on $W \cap E^c$ and* ok *on R and on $W \cap E$, so it cannot distinguish a claim that is right from one that is wrong because of the shared source.* Stage 2 (decision). *Each receiver accepts or rejects by Bayes' rule given the public verdicts, with loss L on accepting W and ℓ on rejecting R.* Payoffs. *Receiver i bears her own purchase cost plus the decision loss implied by the strongest technology purchased by anyone,*

$$u_i(a) \;=\; -\,\kappa(a_i) \;-\; \mathcal{L}\big(\max_j a_j\big), \qquad N \prec S \prec I,$$

with $\mathcal{L}$ as in (5)*. The verdict is a public good with the best-shot technology of [24]: only the strongest verification bought by anyone matters, and a receiver who buys nothing still benefits from it, which is the source of free riding below.*

Figure 3 summarizes the timing of Γ_n and what each technology reveals.

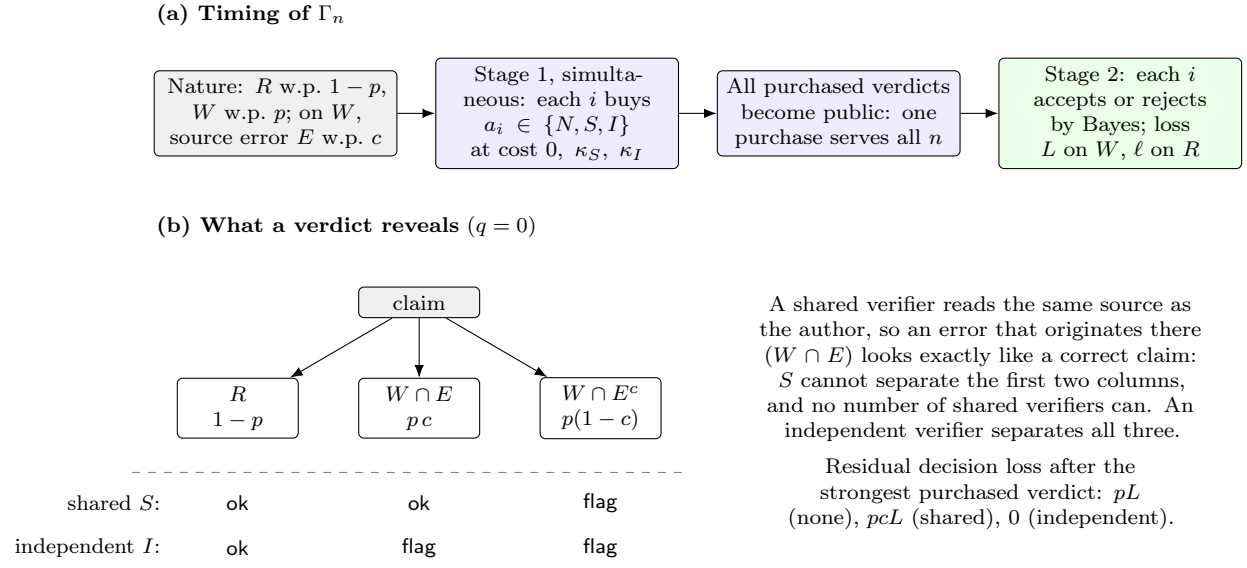

Figure 3: The simultaneous procurement game Γ_n. (a) Timing: nature draws the claim and the origin of its error; receivers buy at most one verification each, simultaneously; verdicts are public; each receiver then decides. (b) Information: the three states of the claim and the verdict each technology returns in each state; the shared verdict is blind to errors inherited from the shared source, the event of probability pc behind the floor of Theorem 1.

For $n = 2$ the game is the 3×3 cost matrix of Table 2. Pure equilibria are the cells in which neither row nor column player can lower her own cost by a unilateral move; Theorem 2 identifies them for every n.

Table 2: Expected costs (player 1, player 2) in Γ_2 at $q = 0$. A purchased independent verdict removes all decision loss; a shared verdict leaves the residual loss pcL to everyone.

	N	S	I
N	$(pL,\ pL)$	$(pcL,\ \kappa_S + pcL)$	$(0,\ \kappa_I)$
S	$(\kappa_S + pcL,\ pcL)$	$(\kappa_S + pcL,\ \kappa_S + pcL)$	$(\kappa_S,\ \kappa_I)$
I	$(\kappa_I,\ 0)$	$(\kappa_I,\ \kappa_S)$	$(\kappa_I,\ \kappa_I)$

A numerical instance. Let $p = 0.1$, $c = 0.5$, $L = 100$, $\ell = 20$ (so $pL = 10 < (1-p)\ell = 18$ and Assumption 1 holds), $\kappa_S = 3$, $\kappa_I = 9$. Then $pL = 10$, $pcL = 5$ and $pL(1-c) = 5$. A lone independent buyer would switch to S (cost 9 against $3 + 5 = 8$), so no equilibrium contains I; a lone shared buyer pays $8 < 10$ and would not upgrade ($8 < 9$), while a free rider bears 5 and would not buy I for 9; hence the unique pure equilibrium is one S and the rest N. For $n = 3$ the social optimum is $\min\{30,\ 3 + 15,\ 9\} = 9$, attained by one independent check, whereas the equilibrium costs society $3 + 3 \cdot 5 = 18$: the independence premium $d = \kappa_I - \kappa_S = 6$ lies in the wedge $pcL = 5 < d < npcL = 15$ of Theorem 3.

Theorem 2 (Pure procurement equilibria)**.** *Under Assumption 1, $\kappa_S, \kappa_I > 0$, and away from all equality boundaries, pure Nash equilibria consist exclusively of the following profiles (up to permutation):*

1. *I by exactly one receiver, all others N, iff $\kappa_I < \min\{pL,\ \kappa_S + pcL\}$;*
2. *S by exactly one receiver, all others N, iff $\kappa_S < pL(1-c)$ and $\kappa_I > \kappa_S + pcL$;*
3. *all N, iff $\kappa_S > pL(1-c)$ and $\kappa_I > pL$.*

At positive costs no equilibrium contains duplicate purchases or both S and I. On equality boundaries weak equilibria may coexist.

Proof. If I is purchased, every other buyer can cancel her redundant purchase and save a strictly positive amount. The unique I purchaser compares her cost κ_I with pL (switch to N) and $\kappa_S + pcL$ (switch to S), yielding (1). If no I is purchased and at least one S is, only one S can remain due to strict positive costs. Its buyer prefers S over N iff $\kappa_S + pcL < pL$ and over I iff $\kappa_S + pcL < \kappa_I$. A free rider faced with S could buy I, reducing her loss from pcL to zero, but this is not profitable because $\kappa_I > \kappa_S + pcL > pcL$; hence (2). At the all-$N$ profile, neither unilateral S nor unilateral I is profitable precisely under (3). All remaining profiles contain redundant purchases and cannot be equilibria. □

Proposition 3 (Socially efficient procurement)**.** *The minimum aggregate expected cost equals*

$$\min\{npL,\ \kappa_S + npcL,\ \kappa_I\}. \tag{6}$$

The minimizing alternative is, respectively, no verification, one shared-source check, or one independent-source check. Strict inequalities ensure uniqueness of technology choice.

Proof. All receivers benefit from a public verdict. Their total residual decision loss is n times (5); duplication adds positive purchase costs without improving information when $q = 0$. Taking the minimum yields (6). □

6 The price of independence: the crowding-out wedge

The two technologies can be understood as a choice of source architecture. Let shared procurement cost $\kappa > 0$ and define the *independence premium* $d > 0$ by $\kappa_I = \kappa + d$; this premium can represent independent source acquisition, fresh auditing, provenance reconstruction, or non-shared model infrastructure. Here the technology is chosen by purchasing receivers, not by a strategic verification vendor; vendor pricing is left for a subsequent market model.

Theorem 3 (Source-choice crowding-out wedge)**.** *Fix $n \geq 2$, $0 < c < 1$, $p \in (0,1)$, $L > 0$, and $0 < \kappa < pL(1-c)$. For generic premiums in*

$$\boxed{pcL < d < npcL}, \tag{7}$$

Aicumene:2026.047v1 [econ.TH] October 2026

Aicumene:2026.047v1 [econ.TH] October 2026

all pure Nash equilibria purchase exactly one shared-source check, whereas every social optimum purchases exactly one independent-source check. The welfare cost, relative to the optimum, is

$$\mathcal{W} = npcL - d > 0, \tag{8}$$

which is at most $(n-1)pcL$ *within the open wedge. The premium interval has width* $(n-1)pcL$*. The stated region cannot exist at* $c = 0$ *(no wedge) or at* $c = 1$ *with nonnegative* κ *(shared procurement has zero private benefit).*

Proof. The shared-check buyer bears cost $\kappa + pcL$ and prefers it to N because $\kappa < pL(1-c)$. She prefers it to I precisely when $d > pcL$. A free rider would gain at most pcL by independently checking, and would have to pay $\kappa + d > pcL$, so no other buyer deviates. By Theorem 2, these are the only pure equilibrium profiles. Socially, independent verification costs $\kappa + d$, whereas shared verification costs $\kappa + npcL$; $d < npcL$ makes independent verification superior. Furthermore $\kappa + d < \kappa + npcL < npL$ because $\kappa < pL(1-c) \leq npL(1-c)$. Thus independence also dominates no verification. The excess welfare cost equals $(\kappa + npcL) - (\kappa + d) = npcL - d$. Subtracting the strict endpoints of (7) yields the interval width. When $c = 0$, the endpoints coincide. When $c = 1$, the prerequisite $\kappa < pL(1-c) = 0$ contradicts $\kappa > 0$. □

Interpretation. The difference between the purchaser's incentive pcL and the ecosystem-wide gain $npcL$ is an uninternalized source-dependence externality. Increasing the number of beneficiaries increases the range over which independent provenance is socially desirable but individually unaffordable. It is not correct to describe this as a result for *all* values of c: our complete result is an interior region with $0 < c < 1$.

Figure 4 shows the equilibrium map in the price plane at $q = 0$, where the wedge is the closed-form region (7), and at $q = 0.03$, where it is computed by exact enumeration with Bayes-optimal decisions; the next section gives an explicit radius in q inside which the map is provably unchanged.

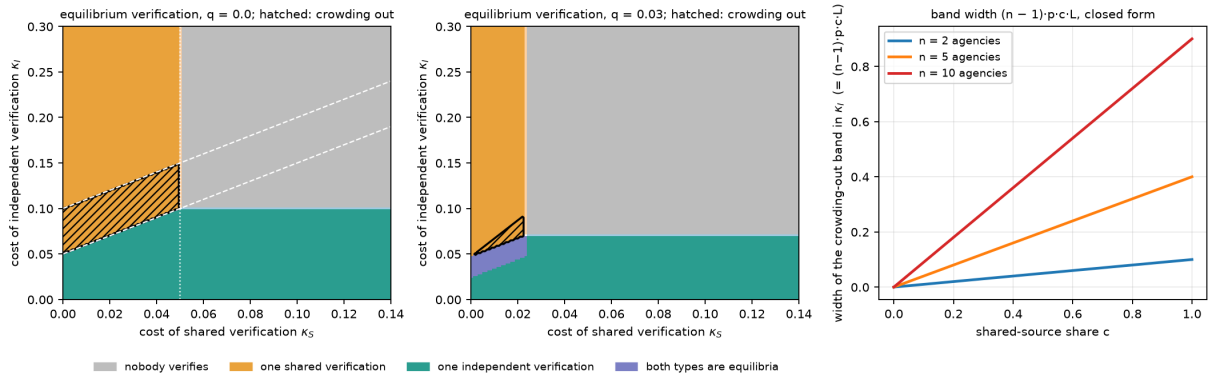

Figure 4: Left and centre: type of verification bought in pure equilibrium as a function of the two prices, for $q = 0$ (closed form; dashed lines bound the wedge (7) with $n = 2$) and $q = 0.03$; hatched: the crowding-out region, where the optimum is independent verification. Right: width of the wedge as a function of c and n. $p = 0.1$, $c = 0.5$, $L = \ell = 1$. Generated by `sim_vgame.py`.

Example 2 (The epigraph, continued: who goes to the vault)**.** Continue Example 1 with $T = 40$ and $\varepsilon = 10^{-3}$. Let $n = 30$ monks live under the rule in each generation, 1 200 over the millennium; let L be one monastic life under the wrong rule and $\ell = L$, so that Assumption 1 holds; let κ_S be one day of comparison with the exemplar and d three days of descent to the vault, so that in lifetimes of forty years $\kappa_S = 6.8 \cdot 10^{-5}$ and $d = 2.1 \cdot 10^{-4}$. At the calibrated $p = 0.039$ the private value of the vault, $pcL = 0.038$ lifetimes or a year and a half, is almost two hundred times d: by Theorem 2 the unique pure equilibrium has one monk going down, the first who thinks of it, and the expected loss over the millennium that the descent saves is $nTpcL = 46$ monastic lives. A thousand years of copying is therefore the all-N equilibrium at a *believed* probability of error $\tilde{p}$ so small that neither check pays, $\tilde{p} < \min\{\kappa_S/(L(1-c)), (\kappa_S + d)/L\} = 2.7 \cdot 10^{-4}$, one chance in 3 650 against a calibrated 0.039 (Figure 5(b)); the exemplar check is never an equilibrium purchase here, because c is close to one and the vault is cheap. The monastery's failure was not the price of the vault

but the prior, and each generation that saw the previous one accept the rule is the permanent cascade of the companion paper. The wedge (7) is the intermediate case in which the vault is dear enough that no single monk goes and cheap enough that the community should send one; with the monks of the millennium as receivers it is a descent costing between a year and a half and 46 monastic lives. The numbers and the figure are produced by `epigraph.py`.

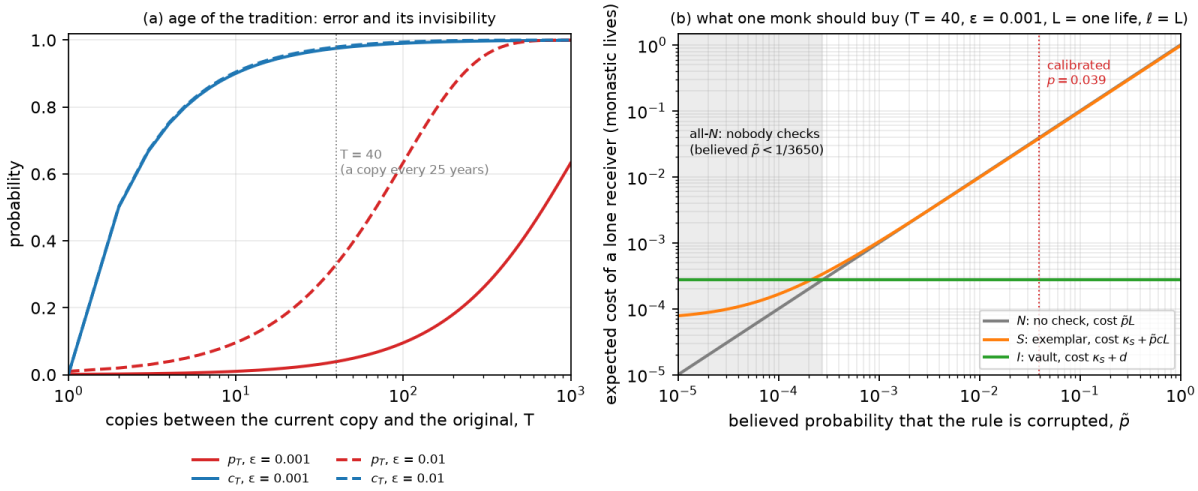

Figure 5: The epigraph computed (Examples 1 and 2). (a) Probability p_T that a given word is corrupted after T copyings and share c_T of such errors invisible to a check against the exemplar, for $\varepsilon = 10^{-3}$ (solid) and 10^{-2} (dashed); the posterior after a reassuring exemplar check lies within 0.01 of p_T throughout. (b) Expected cost of a lone receiver for the three actions as a function of the believed probability $\tilde{p}$ that the rule is corrupted, in monastic lives ($T = 40$, $\varepsilon = 10^{-3}$, κ_S one day, d three days): below $\tilde{p} = 2.7 \cdot 10^{-4}$ nothing is bought, above it the vault; the exemplar check is never cheapest. The calibrated $p = 0.039$ is 140 times the threshold.

7 Robustness to imperfect verifier signals, a mechanism, and exact enumeration

The preceding procurement results assume perfect idiosyncratic accuracy, $q = 0$. We now retain the original common-source signal model but let every purchased verifier independently flip its ideal signal with probability $q \in [0, 1/2)$ conditional on $(R/W, E)$. In particular, the common-source event remains observationally indistinguishable from R even for multiple shared checks.

Set $\kappa_S = \kappa$, $\kappa_I = \kappa + d$ and maintain the strict interior conditions of Theorem 3. Define the individual-deviation and social-optimality margins at zero noise by

$$m = \min\{\kappa,\ d - pcL,\ pL(1-c) - \kappa\} > 0, \tag{9}$$

$$\Delta = \min\{npcL - d,\ npL - \kappa - d\} > 0, \tag{10}$$

$$M = \max\{L, \ell\}. \tag{11}$$

Proposition 4 (Explicit robustness radius)**.** *For all $0 \le q < \bar{q}$ where*

$$\boxed{\bar{q} = \min\left\{\frac{1}{2}, \frac{m}{2nM}, \frac{\min\{\Delta, \kappa\}}{2n^2 M}\right\}}, \tag{12}$$

the pure Nash equilibria are precisely the n profiles with one shared check and $n-1$ nonpurchasers; the unique socially optimal procurement technology remains one independent check.

Aicumene:2026.047v1 [econ.TH] October 2026

Aicumene:2026.047v1 [econ.TH] October 2026

Proof. For a fixed procurement profile a, couple each noisy signal to its $q = 0$ counterpart by an independent flip of probability q. No more than n checks are purchased, so the probability that any observed signal differs from its zero-noise counterpart is at most nq. For every fixed decision rule, its loss lies in $[0, M]$, so the difference in expected loss is bounded by nqM. Taking minima over the finite set of deterministic decision rules preserves the bound: $|R_q(a) - R_0(a)| \le nqM$. Therefore unilateral payoff comparisons change by at most $2nqM$, and social comparisons change by at most $2n^2qM$.

At $q = 0$, the one-shared-check equilibrium has no profitable deviation. Its smallest strict deviation margin is at least m: the buyer's deviations from S to N and I respectively cost $pL(1-c) - \kappa$ and $d - pcL$; a free rider's additional S costs κ, whereas buying I costs $\kappa + d - pcL > m$. Every other procurement profile has a strictly profitable deviation of gain at least m: redundant purchasers can drop out, a sole independent purchaser prefers shared procurement since $d > pcL$, and from the all-N profile a shared purchase gains $pL(1-c) - \kappa$. Thus the Nash-equilibrium set is unchanged if $2nqM < m$.

At $q = 0$, one independent check has lower aggregate cost than one shared check by $npcL - d$ and than no verification by $npL - \kappa - d$. Every other profile has an additional strictly positive procurement cost because it contains redundant purchases. Such a profile differs from the one-independent-check benchmark by at least $\min\{\Delta, \kappa\}$ unless it contains a shared or independent check with a zero-cost redundant purchase, ruled out by $\kappa > 0$. Since $\kappa > 0$, the minimum strict welfare gap to any other profile is at least $\min\{\Delta, \kappa\}$. It remains positive whenever $2n^2qM < \min\{\Delta, \kappa\}$, which gives the stated bound. □

Lemma 2 (Local value of a second shared check)**.** *Assume $0 < c < 1$ and $(1-p)\ell > pL$, and put $B = (1-p)\ell - pcL$, $r = \sqrt{B/(p(1-c)L)}$ and*

$$q^* = \frac{1}{1+r^2} = \frac{pL(1-c)}{(1-p)\ell + pL(1-2c)}; \tag{13}$$

Assumption 1 gives $r > 1$, so $q^* < 1/(1+r) < \frac{1}{2}$. For $0 \le q \le q^*$, the Bayes-optimal decision based on one shared verdict rejects exactly upon a flag, while with two shared verdicts it rejects exactly when both flag. If $R_k(q)$ is the minimum expected decision loss given k shared verifiers, then

$$R_1(q) = pcL + q\big[(1-p)\ell + pL(1-2c)\big], \tag{14}$$

$$R_2(q) = pcL + 2pL(1-c)q + q^2\big[(1-p)\ell - pL\big]. \tag{15}$$

Consequently,

$$\boxed{R_1(q) - R_2(q) = \big[(1-p)\ell - pL\big]q(1-q).} \tag{16}$$

The leading coefficient is independent of c, although the range $q \le q^*$ on which the stated decision rules are optimal depends on c through r.

Proof. With one verdict, the posterior odds of W against R after flag are $[pcq + p(1-c)(1-q)]/[(1-p)q]$, and rejection is optimal iff $pL(1-c)(1-q) \ge qB$, i.e. iff $(1-q)/q \ge r^2$, i.e. $q \le 1/(1+r^2)$; after ok the odds are $[pc(1-q) + p(1-c)q]/[(1-p)(1-q)]$, which is at most $p/(1-p)$ for $q \le 1/2$, so acceptance is optimal by Assumption 1. With two verdicts, after one flag and one ok in either order the odds are exactly $[pcq(1-q) + p(1-c)(1-q)q]/[(1-p)q(1-q)] = p/(1-p)$ for every q, so acceptance is optimal; after two flags rejection is optimal iff $pL(1-c)(1-q)^2 \ge q^2B$, i.e. iff $(1-q)/q \ge r$, i.e. $q \le 1/(1+r)$; after two ok signals the odds are again at most $p/(1-p)$ for $q \le 1/2$. Hence the one-verdict rule is optimal iff $q \le 1/(1+r^2) = q^*$ and the two-verdict rule iff $q \le 1/(1+r)$, which is larger; ties occur only at the endpoints.

When one check is used, acceptance of a wrong claim has probability $c(1-q) + (1-c)q$, while rejection of a correct claim has probability q. For two checks, acceptance of a wrong claim has probability $c(1-q^2) + (1-c)(2q-q^2)$, while rejection of a correct claim has probability q^2. Multiplication by the prior probabilities and losses gives the displayed formulas. Their difference is Eq. (16). □

Write $A = (1-p)\ell - pL > 0$. A duplicate shared check is not profitable when $Aq(1-q) < \kappa$. When $0 < \kappa < A/4$, the first solution of the equality is

$$q_{\text{dup}} = \frac{1 - \sqrt{1 - 4\kappa/A}}{2} = \frac{\kappa}{A} + O\!\left(\frac{\kappa^2}{A^2}\right). \tag{17}$$

Aicumene:2026.047v1 [econ.TH] October 2026

This threshold is *local*: it concerns one unilateral deviation. The range of the lemma, however, is sharp, and what happens beyond it is simple.

Proposition 5 (The identity ends at q^*)**.** *For $q^* < q < \frac{1}{2}$, $R_1(q) - R_2(q) < Aq(1-q)$. More precisely, for $q^* < q \le 1/(1+r)$ a single shared* flag *no longer justifies rejection, so $R_1(q) = pL$: a lone shared check is worthless. R_2 keeps the rule of Lemma 2, and $R_1 - R_2 - Aq(1-q) = pL(1-c)(1-2q) - Aq$ is linear in q, vanishes at q^* and is negative beyond. For $q > 1/(1+r)$ both panels are worthless, $R_1 = R_2 = pL$, and the left side is $0 < Aq(1-q)$.*

Proof. The posterior-odds computations in the proof of Lemma 2 show that above $1/(1+r^2)$ the odds after one flag fall below ℓ/L, so accepting every claim is optimal with one verdict, and that above $1/(1+r)$ the same holds after two flags. Substituting $R_1 = pL$ and the expression for R_2 gives $pL - R_2 - Aq(1-q) = pL(1-c) - 2pL(1-c)q - Aq = pL(1-c)(1-2q) - Aq$, whose root is $pL(1-c)/(A + 2pL(1-c)) = q^*$ by (13). □

Where the wedge ends. Proposition 4 is a lower bound on the own error the wedge tolerates. The following gives the exact threshold for the equilibrium side and an explicit upper bound for the welfare side. Write $\mathcal{C} = (1-p)\ell + pL(1-2c)$ for the coefficient of q in R_1 and define

$$q_{SN} = \frac{pL(1-c) - \kappa}{\mathcal{C}}, \qquad q_{II} = \frac{1 - \sqrt{1 - 4(\kappa + d)/(nA)}}{2},$$
$$q_{SS} = \text{the smaller root of } \; nAq^2 - n(A + 2pcL)\, q + (npcL + \kappa - d) = 0, \tag{18}$$

with $q_{II} = \infty$ if $\kappa + d \ge nA/4$ and $q_{\mathrm{dup}} = \infty$ if $\kappa \ge A/4$. At q_{SN} a lone shared check stops paying for itself ($pL - R_1(q) = \kappa$); at q_{II} a second independent check pays for the group ($nAq(1-q) = \kappa + d$); at q_{SS} two shared checks become socially cheaper than one independent check.

Proposition 6 (Exact equilibrium threshold and an upper bound on the radius)**.** *Under the hypotheses of Proposition 4 and for $0 \le q < \frac{1}{2}$:*

1. *the n one-shared-check profiles are Nash equilibria iff $q \le q_{\mathrm{eq}} := \min\{q_{\mathrm{dup}}, q_{SN}\}$, and $q_{\mathrm{eq}} \le q_{SN} < q^*$;*
2. *if $q_{\mathrm{opt}} := \min\{q_{II}, q_{SS}\} < q^*$, the one-independent-check profile is not a social optimum for $q_{\mathrm{opt}} < q \le q^*$;*
3. *consequently the wedge statement fails for every $q > q_{\mathrm{end}} := \min\{q_{\mathrm{eq}}, q_{\mathrm{opt}}\}$, so that $\bar{q} \le q_{\mathrm{end}}$.*

Proof. Let $R_{k,k'}$ be the Bayes risk with k shared and k' independent verdicts, so that $R_k = R_{k,0}$. On $0 \le q \le q^*$ the rules of Lemma 2 are optimal, and so are "reject iff flag" for one independent verdict and "reject iff both flag" for two, whose validity thresholds $pL/(pL + (1-p)\ell)$ and $1/(1 + \sqrt{(1-p)\ell/(pL)})$ exceed q^* because $r^2 \ge (1-p)\ell/(pL)$ under Assumption 1; hence $R_{0,1} = q[pL + (1-p)\ell]$, $R_{0,2} = 2pLq + q^2[(1-p)\ell - pL]$ and $R_{0,1} - R_{0,2} = Aq(1-q)$. With one independent and one shared verdict, the odds after (ok, flag) and after (ok, ok) are at most $p/(1-p)$ for $q \le \frac{1}{2}$, and after two flags at least the odds after a single independent flag; so the Bayes rule rejects after two flags and accepts in the other two cases, while after (flag, ok) it rejects iff $q \le q_{IS} := pcL/(A + 2pcL)$. Therefore $R_{1,1} = R_{0,1}$ for $q \le q_{IS}$ and $R_{1,1} = (1-p)\ell q^2 + pcL(1 - q + q^2) + pL(1-c)(2q - q^2)$ for $q_{IS} < q \le q^*$.

Equilibrium side. From a one-shared-check profile the four unilateral deviations gain: the buyer's switch to N gains $\kappa - [pL - R_1(q)] = \kappa - pL(1-c) + \mathcal{C}q$, positive iff $q > q_{SN}$; her switch to I gains $R_1 - R_{0,1} - d = pcL(1-2q) - d < 0$; a free rider's duplicate gains $Aq(1-q) - \kappa$, positive iff $q > q_{\mathrm{dup}}$; her independent check gains $R_1 - R_{1,1} - \kappa - d$, which equals $pcL(1-2q) - \kappa - d < 0$ for $q \le q_{IS}$ and $Aq(1-q) + pcLq(1-2q) - \kappa - d < Aq(1-q) - \kappa$ for $q > q_{IS}$, so it is profitable only when the duplicate already is. For $q > q^*$ the buyer's switch to N is profitable because $R_1 = pL$, and $q_{SN} < q^*$ because $\kappa > 0$. This proves (1).

Welfare side. The one-independent-check profile costs $nR_{0,1} + \kappa + d$. Two independent checks cost $nR_{0,2} + 2(\kappa + d)$, which is less iff $nAq(1-q) > \kappa + d$, i.e. $q > q_{II}$. Two shared checks cost $nR_2 + 2\kappa$, which is less iff $nAq^2 - n(A + 2pcL)q + (npcL + \kappa - d) < 0$, i.e. $q > q_{SS}$ on $[0, q^*]$, the constant term being positive inside the wedge. This proves (2), and (3) follows. □

Aicumene:2026.047v1 [econ.TH] October 2026

On the grid of Section 7.2, (1) is confirmed exactly, and q_{end} coincides with the breakdown of the wedge found by enumeration on 16 of the 19 combinations; on the remaining ones, all with $p = 0.25$, the optimum moves to *three* shared checks before q_{SS}, so the exact welfare threshold is the minimum over $k \geq 2$ shared checks, a finite computation but not a closed form. Together with Proposition 4 the tolerance of the wedge is bracketed, $\bar{q} \leq q_{\text{tol}} \leq q_{\text{end}}$; on the grid the two bounds differ by a factor between 10 and 65, and the upper bound is the one that is nearly tight.

7.1 A sufficient implementation mechanism

A simple targeted reimbursement for the first independent-source purchase can bridge the private/social wedge. Assume the provider receives full procurement cost but an institution pays a subsidy $s \geq 0$ to the sole purchaser of the first independent check, and no subsidy is paid for redundant checks. Assume subsidy transfers have zero deadweight cost.

Proposition 7 (Implementability by a source-independence subsidy)**.** *Under Theorem 3, consider a subsidy s paid only to the first purchaser of an independent-source check. Suppose the total cost to the purchaser remains positive after the subsidy. Every subsidy satisfying*

$$\boxed{d - pcL < s < \kappa + d} \tag{19}$$

implements a strict Nash equilibrium in which exactly one receiver purchases I and the other receivers buy nothing. The infimum subsidy needed to induce that equilibrium is $d - pcL$; it is strictly smaller than the total external benefit $(n-1)pcL$ in the wedge. This is an implementation result, not a guarantee that no other equilibrium exists.

Proof. The buyer choosing I incurs effective private cost $\kappa + d - s$. Changing to S costs $\kappa + pcL$, so the buyer strictly prefers I if $s > d - pcL$. Changing to N costs pL. The maintained inequality $\kappa < pL(1-c)$ implies $\kappa + pcL < pL$; hence $\kappa + d - s < \kappa + pcL < pL$ whenever $s > d - pcL$. Thus no deviation by the purchaser is profitable. Every free rider currently incurs zero decision loss and pays nothing, whereas an additional check adds strictly positive cost and cannot further improve the verdict. Therefore all other receivers strictly prefer N. The upper bound in (19) keeps the effective independent-check cost positive and excludes trivial subsidy arbitrage. Within the crowding-out wedge $d < npcL$, we have $d - pcL < (n-1)pcL$. The lower bound may be approached arbitrarily closely, proving the infimum claim. □

Welfare qualification. If subsidies are pure transfers and their financing has no distortion, implementing one independent-source verification restores the efficient procurement technology. With distortionary taxation or endogenous supply, the efficiency claim requires a separate model.

7.2 Exact verification by finite-game enumeration

We enumerate all 3^n action profiles for $n \in \{2, 3, 4\}$ and evaluate every unilateral deviation, classifying pure equilibria directly from expected costs. Separately, we minimize total decision loss plus procurement costs across all profiles. The supplied script checks 19 interior parameter combinations satisfying Theorem 3; all tests agreed with its prediction. The script `verify_robustness.py` repeats the enumeration with losses computed exactly from the verdict distribution at $q > 0$ and runs four tests: (1) the identity (16) against the Bayes-optimal risk at $q \leq 0.99\, q^*$ (30 parameter–q pairs) and its strict failure at $1.01\, q^*$; (2) Proposition 4 at $q = 0.99\, \bar{q}$ on the 19 combinations; (3) Proposition 6: the one-shared-check profiles are equilibria at $0.99\, q_{\text{eq}}$ and not at $1.01\, q_{\text{eq}}$, the one-independent-check profile is not optimal at $1.01\, q_{\text{opt}}$, and a scan in steps of $5 \cdot 10^{-4}$ locates the breakdown of the wedge, which equals q_{end} on 16 of the 19 combinations; (4) Proposition 5 on 347 random parameter sets drawn under Assumption 1. The script `sim_vgame.py` draws Figures 1 and 4. This is a finite computational check, not a proof for arbitrary n or real data. The closed forms of Section 7 are, in addition, re-derived symbolically: `audit_section5.py` recomputes with SymPy the Bayes rule for every pair of shared and independent verdicts, R_1, R_2, $R_{1,1}$, the identity (16), q^*, q_{IS}, the deviation gains and the thresholds (18), and checks the piecewise closed forms against brute-force Bayes risk on random parameters, and verifies the ordering $q_{\text{eq}} \leq q_{SN} < q^*$ and the sign changes of the deviation gains at their thresholds on random points of the wedge (28 checks, all passed).

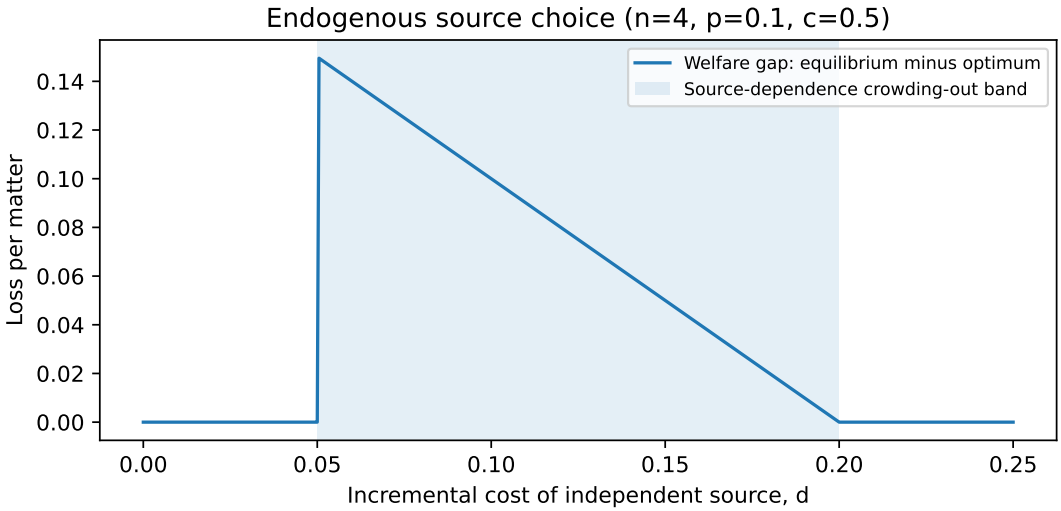

Figure 6: Computed equilibrium–optimum welfare gap as the independence premium varies ($n = 4$, $p = 0.1$, $c = 0.5$, $L = 1$, $\kappa = 0.02$). Shading identifies the theoretical premium interval. Outside the interior strict region, multiple equilibria and other public-good failures may arise; the complete finite game is enumerated rather than extrapolated from the theorem.

8 Discussion: the epistemic limits of strategic rationality

Each statement in this section is an interpretation of a result above under its premises, not a theorem about collective knowledge in general. The results point to a distinction that is fundamental to the design of multi-agent intelligence: the rationality of individual agents, the rationality of their institutional arrangements, and the epistemic reliability of their collective conclusions are different properties. An agent may correctly optimize its expected payoff while relying on evidence that contains an undetectable common error (Theorem 1). A verification network may assign responsibility to every participant while allowing cycles in which claims are confirmed by records derived from one another (Lemma 1, Proposition 2). A procurement equilibrium may be strategically stable while preserving avoidable information dependence (Theorem 3). None of these outcomes requires an individual agent to behave irrationally.

The difficulty arises because verification has two dimensions that are often conflated. One is procedural: whether an agent has performed the prescribed checks. The other is epistemic: whether those checks provide information capable of distinguishing a true claim from a false one. A verification procedure cannot create information that is absent from its evidentiary inputs, and the multiplication of agents does not by itself multiply independent evidence. The information floor expresses this limitation mathematically. The procurement game adds a second result: even when the missing information can be acquired, the incentives of individual agents may prevent its acquisition. Epistemic reliability is therefore not an automatic consequence of strategic equilibrium; it has to be designed into the architecture of evidence acquisition and verification.

In this sense independence has an economic price, but dependence has a social cost. The design problem is to determine when the price of independent evidence is justified by the reduction in collective error, and how institutions can align individual incentives with that collective benefit; the subsidy of Proposition 7 is the simplest such alignment. The central lesson is not that multi-agent agreement should be distrusted. It is that agreement must be interpreted in the light of its evidentiary genealogy: a collective conclusion becomes better justified when the supporting evidence contributes genuinely discriminating information, not when more agents repeat the same verdict.

Aicumene:2026.047v1 [econ.TH] October 2026

Aicumene:2026.047v1 [econ.TH] October 2026

9 Related work and research agenda

9.1 Antecedents and neighbours

Antecedents of the floor. The inequality behind Theorem 1 is known in three literatures, and we claim no priority for it. In robust statistics it is the contamination bound: a component of the alternative that coincides with the null cannot be detected by any test, which is the basis of Huber's robust probability-ratio test [1]. In reliability engineering it is the common-cause or beta-factor model: a fraction of failures shares one cause, and redundancy cannot push system failure below that fraction [2]; for multiversion software the same limit was derived by [3, 4] and observed experimentally by [5]. In decision analysis, [6] showed that dependence among information sources bounds the precision and value obtainable from any number of them, which is the counterpart of the $k \to \infty$ statement in Corollary 1. What is new here is the primitive c as an identifiable conditional indistinguishability, the extension to source trees, and the equilibrium consequences for who buys which verification.

Costly state verification [7, 8] has one principal, perfect verification and no notion of a verifier's source. The notion of probabilistic truth as a measurable epistemic status of a claim was introduced in the April Theses of the New Cognitive Paradigm [20] and operationalized as the variable τ with an action threshold in [21]; the present paper supplies its market side, namely who provides the independent evidence that raises τ (Remark 1). Mechanisms with costly or hard evidence [9, 10] and evidence games [11, 12, 13] treat evidence as error-free and sourceless; the unraveling results there concern disclosure incentives, not the independence of the check. The Condorcet jury theorem with correlated votes [14, 15] shows that correlation can prevent the aggregate from converging to the truth; [25, 26] trace the failure to a common cause and argue that independence conditional on the cause is the defensible premise, which is exactly the structure of E, and [27] extend the theorem to general dependent juries; Theorem 1 gives a tractable bound for a particular common-cause signal structure, and Theorems 2 and 3 add an explicit public procurement choice. However, game-theoretic models of costly and correlated information acquisition already exist (including [42]), so a broader originality claim would be unwarranted. In [28] every information source carries a sender noise common to all who consult it and a receiver noise that attention reduces; the common component is the analogue of the shared-source error and, as here, cannot be averaged away by consulting the source more. Private provision of public goods [30], the best-shot technology [24] and the volunteer's dilemma [31] supply free riding; the contribution here is that the *type* of public good bought in equilibrium is systematically the correlated one. A referee may still read Theorem 3 as underprovision of a public good, which it is; what the model adds is that the underprovided margin is the provenance of the evidence, that its two primitives are measurable, c by the snapshot design of Section 9.2 and d as a market price, and that the own error the result tolerates is bounded in Section 7. Our contribution is therefore not the general observation that correlated evidence can mislead [6, 25] or that public information is underprovided [29, 30, 28]; it is, within a specified source-dependent verification technology, an explicit equilibrium wedge between the private willingness to pay for independent provenance and its aggregate value, the classification of the associated pure procurement equilibria, and quantitative robustness and implementation conditions. Information cascades [32, 33] describe copying of unverified actions; Proposition 1 describes the same phenomenon at the level of sources, and the sequential game of Section 9.2 connects the two. Rational inattention [34] prices information acquisition but not the independence of its source. On the AI side, algorithmic collusion [35] shows learning agents converging to a collectively bad equilibrium of a pricing game; the analogue here would be convergence to the shared-verification equilibrium, a collusion on unverified truth. More directly related work includes networked acquisition of costly correlated signals [42], collective-decision mechanisms with costly verification [43], and strategic transparency under costly verification [44]. Our own experiments with cross-family LLM agents [22] find that such agents reach Nash equilibrium exactly when it reduces to a myopic best response over explicit payoffs; the sequential game of the companion paper is of that kind, the procurement game of Section 5 is not. The novelty claim must therefore be confined to the explicitly specified source-identity floor and the closed-form public-procurement wedge under our assumptions. A full literature review is required before journal submission; this section fixes the neighbours, not the boundaries.

Aicumene:2026.047v1 [econ.TH] October 2026

9.2 Boundaries and research plan

Items 2 and 4 below are developed in the companion working paper [23] (Part II: sequential reliance, markets, and provenance); the remaining items are open. This note establishes an endogenous *receiver-side choice of source technology*; it does not yet show that a competing or monopolistic verifier vendor will select a higher value of c. That requires adding vendor investment and pricing decisions, observability of source provenance, demand elasticity, and equilibrium selection. The companion paper takes two of these steps: it treats partially overlapping sources, where the value of a set of sources is a coverage function of their common ancestors, and it characterizes the price game among vendors with observable quality; investment in independence remains open. Likewise, with imperfect own signals ($q > 0$), multiple shared tests become informationally useful: Propositions 4 and 6 bracket the band in which the closed-form procurement theorem survives, $\bar{q} \le q_{\text{tol}} \le q_{\text{end}}$, and the equilibrium side of the band is known exactly. Concretely, the plan is:

1. *Exact achievable set for $q > 0$.* Characterize the closure of the attainable (FR, FA) with k shared and k' independent verifiers as a Neyman–Pearson problem with a common-mode component, and close the gap between the lower bound (12) and the upper bound q_{end} of Proposition 6; the missing piece is the welfare side, where the optimum can move to $k \ge 3$ shared checks.
2. *Sequential observation and cascades.* Receivers move in sequence and observe predecessors' *actions*, not verdicts. Conjecture: the probability that all n receivers accept a wrong claim is bounded below by pc and converges to it only if some receiver buys independent verification; characterize the price band in which nobody does. This is the entrenchment theorem.
3. *Learning agents.* Replace equilibrium play by LLM or no-regret agents over repeated matters. By the criterion of [22], the sequential entrenchment of the companion paper [23] should be reproduced by LLM receivers, since each decision is a myopic Bayes action on explicit losses, while convergence to the shared-verification equilibrium of Theorem 3 requires anticipating free riding and is an open empirical question; the protocol of [22] applies directly.
4. *Endogenous c: the verifier market.* A three-stage market: providers invest in genuinely independent sources or cheaper copied sources; they post prices; receivers purchase checks and act on signals. Buyers may or may not observe c. Conjecture: a monopolist chooses $c > 0$ and competition does not restore independence unless c is observable, a lemons-type result [45]. Comparative statics for hidden source quality and mandated independent audits then become testable predictions.
5. *Mechanisms.* Lindahl prices, mandated independent snapshots, and the cheapest mechanism given who knows c; Proposition 7 is the first case.

Measurement. c is identified by running proposer and verifier on the same matters with the verifier's source snapshot either shared or independent; under the rule "accept iff ok" and a common own error q, the difference in false-accept probabilities is $pc(1 - 2q)$, so c is identified up to the own error, which the same design measures on claims of known truth; with endogenous actions or selection effects a separately specified estimation design is needed. The evaluation protocol of the companion legal paper [16] manipulates exactly this factor. The independence premium d is the observed price difference between independent and shared verification services for the same claim type; with both, the welfare loss (8) is a number in currency. Candidate settings are legal research tools, public registries that copy one root, algorithmic procurement, and medical image annotation, where machine proposals are verified by readers who may take the automatic annotation as their source [18, 19]. No field-measured data are claimed here.

10 Conclusion

The core contribution is a bridge between an information-theoretic lower bound and equilibrium procurement of verification. Source sharing creates an epistemic floor; strategic free riding can preserve that floor even when independent verification is socially superior. Distinct institutional checks need distinct provenance, not merely different labels, algorithms, or model instances. The results survive imperfect own signals within the band bracketed in Section 7, whose equilibrium side is known exactly. In the terms of the epigraph, a thousand years of copying from the copy is the all-N equilibrium of Theorem 2 with a shared source, and the abbot's descent to the vault is the independent check. Examples 1 and 2 compute it: after forty copyings the rule is corrupted with probability 0.039 and a check against the exemplar sees 2.5 per cent of the errors; the

millennium is the all-N equilibrium at a believed probability of error below one in 3 650, so the failure was the prior, not the price of the vault, and the descent would have saved 46 monastic lives for three days of the abbot's time. The empirical scope and priority of the model relative to existing costly-verification and correlated-signal theories require a full literature review before journal submission.

Code availability and reproducibility notes

All results of this paper are checked by exact enumeration, not by simulation: the scripts enumerate every action profile and every unilateral deviation of the finite game, with decision losses computed from the exact verdict distribution. The code is available at `github.com/aicumene/verification-games`. The repository contains `model.py` (Theorems 2–3 and Proposition 7 on the parameter grid of Section 7.2), `verify_robustness.py` (Lemma 2 and Propositions 4, 5 and 6), `sim_vgame.py` (Figures 1 and 4), `cycles.py` (the counts and probabilities of Proposition 2), `epigraph.py` (Examples 1 and 2 and Figure 5), `audit_section5.py` (symbolic audit of Section 7; needs SymPy), the build script, and the files `numbers.tex` and `verification_grid.csv` written by the scripts; every number quoted in the text comes from these files. The scripts need only NumPy and Matplotlib. No human data are used. The legal verifiability compiler of the companion paper [16] is documented there and is not part of this repository. Authorship follows the working-paper author list and is to be confirmed before submission.

References

[1] Huber, P. J. (1965). A robust version of the probability ratio test. *Annals of Mathematical Statistics* 36(6), 1753–1758. doi:10.1214/aoms/1177699803.

[2] Mosleh, A. (1991). Common cause failures: an analysis methodology and examples. *Reliability Engineering & System Safety* 34(3), 249–292. doi:10.1016/0951-8320(91)90104-F.

[3] Eckhardt, D. E. and Lee, L. D. (1985). A theoretical basis for the analysis of multiversion software subject to coincident errors. *IEEE Transactions on Software Engineering* SE-11(12), 1511–1517. doi:10.1109/TSE.1985.231895.

[4] Littlewood, B. and Miller, D. R. (1989). Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering* 15(12), 1596–1614. doi:10.1109/32.58771.

[5] Knight, J. C. and Leveson, N. G. (1986). An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering* SE-12(1), 96–109. doi:10.1109/TSE.1986.6312924.

[6] Clemen, R. T. and Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research* 33(2), 427–442. doi:10.1287/opre.33.2.427.

[7] Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21(2), 265–293. doi:10.1016/0022-0531(79)90031-0.

[8] Gale, D. and Hellwig, M. (1985). Incentive-compatible debt contracts: the one-period problem. *Review of Economic Studies* 52(4), 647–663. doi:10.2307/2297737.

[9] Ben-Porath, E., Dekel, E. and Lipman, B. L. (2014). Optimal allocation with costly verification. *American Economic Review* 104(12), 3779–3813. doi:10.1257/aer.104.12.3779.

[10] Ben-Porath, E., Dekel, E. and Lipman, B. L. (2019). Mechanisms with evidence: commitment and robustness. *Econometrica* 87(2), 529–566. doi:10.3982/ECTA14991.

[11] Hart, S., Kremer, I. and Perry, M. (2017). Evidence games: truth and commitment. *American Economic Review* 107(3), 690–713. doi:10.1257/aer.20150913.

[12] Milgrom, P. R. (1981). Good news and bad news: representation theorems and applications. *Bell Journal of Economics* 12(2), 380–391. doi:10.2307/3003562.

[13] Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics* 24(3), 461–483. doi:10.1086/466995.

[14] Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science* 36(3), 617–634. doi:10.2307/2111584.

Aicumene:2026.047v1 [econ.TH] October 2026

[15] Berg, S. (1993). Condorcet's jury theorem, dependency among voters. *Social Choice and Welfare* 10(1), 87–95. doi:10.1007/BF00187435.

[16] Bernadotte, A. and Menshikov, I. (2026). Verifiable AI for Law: A Legal Verifiability Compiler with a Rule/Judgment Firewall. Aicumene Preprint 2026.040. doi:10.13140/RG.2.2.18174.75841; ResearchGate publication 416039549, https://www.researchgate.net/publication/416039549. https://research.aicumene.com/papers/verifiable-law/

[17] Bernadotte, A. and Gromov, V. A. (2025). What Is Strong AI? Intrinsic Structural Criteria for Strong Intelligence Based on Dimensionality and Topology. Aicumene Preprint 2025.043. doi:10.13140/RG.2.2.12211.59681. https://research.aicumene.com/papers/StrongAI/

[18] Bernadotte, A., Elfimov, N. and Menshikov, I. (2025). Semi-automatic annotation of brain vessels in magnetic resonance angiography images. *Scientific Data* 13, 41. doi:10.1038/s41597-025-06354-1.

[19] Bernadotte, A. (2026). VeNet: a lightweight neural network for efficient brain vessel segmentation in endovascular robotic surgery. *Scientific Reports* 16, 24566. doi:10.1038/s41598-026-54176-x.

[20] Bernadotte, A. (2026). The April Theses of the New Cognitive Paradigm: From Imitation to Structural Measurability. Aicumene Preprint 2026.046 (v1, April 2026; v2, October 2026). doi:10.13140/RG.2.2.20527.50087, https://research.aicumene.com/papers/NCP/.

[21] Bernadotte, A. and Menshikov, I. (2026). Safe AI by Design: Probabilistic Truth, Dynamic Structural Safety, and a Quantitative Harm Model for Agentic Generative AI. Aicumene Preprint 2026.050v1, https://research.aicumene.com/papers/Safe-by-Design/, SSRN forthcoming.

[22] Bernadotte, A. and Menshikov, I. (2026). When do language models reach Nash? Heterogeneous cross-family LLM agents in classic economic games. SSRN working paper 7637178 (June 2026). https://ssrn.com/abstract=7637178, https://research.aicumene.com/papers/llm-nash/, doi:10.2139/ssrn.7637178.

[23] Bernadotte, A. and Menshikov, I. (2026). Multi-Agent Verification Games II: Information Entrenchment and Competitive Underprovision of Independent Evidence. Part II. Aicumene Preprint 2026.048v1; SSRN working paper 7747018 (October 24, 2026). https://ssrn.com/abstract=7747018, doi:10.2139/ssrn.7747018. https://research.aicumene.com/papers/game-theory-2/.

[24] Hirshleifer, J. (1983). From weakest-link to best-shot: the voluntary provision of public goods. *Public Choice* 41(3), 371–386. doi:10.1007/BF00141070.

[25] Dietrich, F. and Spiekermann, K. (2013). Independent opinions? On the causal foundations of belief formation and jury theorems. *Mind* 122(487), 655–685. doi:10.1093/mind/fzt074.

[26] Dietrich, F. and Spiekermann, K. (2013). Epistemic democracy with defensible premises. *Economics and Philosophy* 29(1), 87–120. doi:10.1017/S0266267113000096.

[27] Peleg, B. and Zamir, S. (2012). Extending the Condorcet jury theorem to a general dependent jury. *Social Choice and Welfare* 39(1), 91–125. doi:10.1007/s00355-011-0546-1.

[28] Myatt, D. P. and Wallace, C. (2012). Endogenous information acquisition in coordination games. *Review of Economic Studies* 79(1), 340–374. doi:10.1093/restud/rdr018.

[29] Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review* 70(3), 393–408. https://www.jstor.org/stable/1805228.

[30] Bergstrom, T., Blume, L. and Varian, H. (1986). On the private provision of public goods. *Journal of Public Economics* 29(1), 25–49. doi:10.1016/0047-2727(86)90024-1.

[31] Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution* 29(4), 605–610. doi:10.1177/0022002785029004003.

[32] 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. doi:10.1086/261849.

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

[34] Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics* 50(3), 665–690. doi:10.1016/S0304-3932(03)00029-1.

[35] Calvano, E., Calzolari, G., Denicolò, V. and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10), 3267–3297. doi:10.1257/aer.20190623.

Aicumene:2026.047v1 [econ.TH] October 2026

[36] Knuth, D. E. (2026). Claude's cycles. Note, Stanford University, 28 February 2026, revised 14 April 2026. https://cs.stanford.edu/~knuth/papers/claude-cycles.pdf (accessed 5 November 2026).

[37] Morrison, K. (2026). KnuthClaudeLean: a Lean formalization of the proof that Claude's construction is correct. https://github.com/kim-em/KnuthClaudeLean/ (accessed 5 November 2026).

[38] Stanley, R. P. (1973). Acyclic orientations of graphs. *Discrete Mathematics* 5(2), 171–178. doi:10.1016/0012-365X(73)90108-8.

[39] Kahale, N. and Schulman, L. J. (1996). Bounds on the chromatic polynomial and on the number of acyclic orientations of a graph. *Combinatorica* 16(3), 383–397. doi:10.1007/BF01261322.

[40] Matoušek, J. (2006). The number of unique-sink orientations of the hypercube. *Combinatorica* 26(1), 91–99. doi:10.1007/s00493-006-0007-0.

[41] Takács, L. (1990). On Cayley's formula for counting forests. *Journal of Combinatorial Theory, Series A* 53(2), 321–323. doi:10.1016/0097-3165(90)90064-4.

[42] Myatt, D. P. and Wallace, C. (2019). Information acquisition and use by networked players. *Journal of Economic Theory* 182, 360–401. doi:10.1016/j.jet.2019.05.002.

[43] Erlanson, A. and Kleiner, A. (2020). Costly verification in collective decisions. *Theoretical Economics* 15(3), 923–954. doi:10.3982/TE3101.

[44] Kattwinkel, D. and Knoepfle, J. (2023). Costless information and costly verification: a case for transparency. *Journal of Political Economy* 131(2), 504–548. doi:10.1086/721618.

[45] Akerlof, G. A. (1970). The market for "lemons": quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3), 488–500. doi:10.2307/1879431.

Aicumene:2026.047v1 [econ.TH] October 2026