

Aggregation Is Not Awareness: Recursive Epistemic Mediation and Individual Authorship under Generative AI

A. Sepúlveda-Jiménez¹
National University, QDR Labs
San Diego, California, USA

April 2026

Abstract

Popular and semi-technical commentary increasingly describes generative artificial intelligence (AI) as producing, or disclosing, a form of “collective mind” shared across its users. This paper argues that such framings conflate three distinct phenomena: functional collective intelligence, distributed cognition, and phenomenal collective consciousness, and substitute rhetorical momentum for conceptual precision. We first separate these notions. We then construct a formal framework that isolates what is actually occurring when large populations are mediated by a shared generative model, identifying three disjoint recursion regimes (*training recursion*, *stylistic recursion*, and *cognitive recursion*) whose evidence bases and intervention points differ sharply. Using epistemic logic in the Halpern–Moses tradition, together with Aumann’s agreement theorem, we show that AI-mediated populations exhibit an ersatz form of common knowledge that we term *pseudo-common belief*: the common-knowledge operator C is approximated by bounded-depth mutual knowledge E^k via a shared channel, producing the coordination consequences of common knowledge without its epistemic foundations. We derive a distribution-shift upper bound on the homogenization of public expression under mixed human–model corpora, empirically calibrated via simulation with sub-1% identifiability of the imitation-rate parameter. We cast individual authorship as a signaling game and derive an explicit pooling threshold $\beta^*(\gamma, \kappa) = (\Delta c - \kappa) / (\Delta c - \kappa + \gamma^2 \Delta U_R)$ above which all D1-stable perfect Bayesian equilibria pool, establishing a “paradox of fluency”: better mediators force pooling at lower mediation intensities. We close with an impossibility theorem showing that no intervention preserving the receiver’s information set can restore separating equilibria in the pooling regime; epistemic provenance infrastructure is therefore the structurally unique class of admissible solutions.

Keywords: generative AI; collective intelligence; epistemic logic; common knowledge; model collapse; signaling games; epistemic sovereignty; distribution shift.

1 Introduction

A recurring claim in popular and semi-technical writing on generative AI holds that the technology is producing — or revealing — a kind of collective mind, a coordinated cognitive stratum in which individual thought participates (Floridi, 2014; Ho et al., 2024). Sympathetic readings of this claim invoke the long lineage of distributed cognition (Hutchins, 1995; Clark and Chalmers, 1998); skeptical readings warn of homogenization, erosion of authorship, and cognitive outsourcing (Postman, 1992; Carr, 2010). Both sides, we argue, trade on an ambiguity between three notions that should be kept distinct.

Let CI_F denote *functional collective intelligence*: the demonstrated capacity of groups to outperform individuals on task batteries, operationalized as a latent factor in the Woolley–Malone program (Woolley et al., 2010). Let CI_D denote *distributed cognition*: the thesis that cognitive processes can be constitutively realized across agents and artifacts (Hutchins, 1995; Clark and Chalmers, 1998). Let CI_P denote *phenomenal collective consciousness*: the much stronger claim that a group (with or without artifactual mediators) is itself the subject of unified experience (Dennett, 1991). The last is controversial under any condition; the first two are well-established but do not entail it. The philosophical structure of this three-way distinction has been developed in detail by List and Pettit (2011), who argue that group agency — a weaker

notion than CI_P but stronger than CI_F — admits coherent formalization through aggregation-function theory without requiring phenomenal consciousness. Our framework below inherits their distinction between functional, agential, and phenomenal notions and treats the diagnosis of collective mind as a category error specifically about which level is at issue.

Tsapara (2026) has recently argued that what looks like emergent collective intelligence under AI mediation is better described as “synthetic recursion”, aggregation without awareness. We believe this diagnosis is essentially legitimate but underdeveloped. The term “synthetic recursion” is not defined operationally, the argument slides among CI_F , CI_D , and CI_P , and the concluding normative claim — that human “authorship of interior life” is at stake — is asserted rather than derived.

This paper tightens the argument. Section 2 distinguishes three recursion regimes and provides a formal definition of each. Section 3 develops the epistemic-logic core, showing why AI mediation produces an ersatz common knowledge that has the coordination effects of \mathbf{C} without its grounds, and uses Aumann’s theorem and the Halpern–Moses results to handle the infinite regress cleanly. Section 4 gives a distribution-shift bound on expressive diversity under mixed corpora, empirically calibrated in Appendix C. Section 5 casts authorship as a signaling game and derives an explicit pooling threshold. Section 6 addresses the philosophical objection that all prior media “rewired” cognition without catastrophe, and isolates what is categorically novel. Section 7 lists falsifiable predictions and a constructive proposal. Full proofs appear in Appendices A and B, and an impossibility theorem on provenance-free restoration appears in Appendix D.

2 Three Recursion Regimes

Definition 1 (Training recursion). Let \mathcal{C}_t be the corpus used to train a generative model M_t at epoch t . Let p_H denote the data-generating distribution of authentically human-authored content, and let $p_{M_{t-1}}$ denote the output distribution of the previous model. *Training recursion* obtains when

$$\mathcal{C}_t \sim \alpha_t p_H + (1 - \alpha_t) p_{M_{t-1}}, \quad \alpha_t \in [0, 1], \quad (1)$$

and α_t decreases over t .

Training recursion is well-documented and produces *model collapse*: the progressive loss of distributional tails (Shumailov et al., 2024; Alemohammad et al., 2024). It is a property of training pipelines, not of users.

Definition 2 (Stylistic recursion). Let $\sigma : \mathcal{X} \rightarrow \mathcal{S}$ map expressions to stylistic embeddings. Let π_t be the empirical distribution of human public expression at time t , and let μ be the output style distribution of the dominant generative mediator. *Stylistic recursion* obtains when $\sigma_\# \pi_t \Rightarrow \sigma_\# \mu$ in total variation as $t \rightarrow \infty$.

Stylistic recursion is empirically emerging: evidence includes the homogenization of creative outputs (Doshi and Hauser, 2024), the pervasiveness of large language model (LLM) use in ostensibly human-produced crowd-sourced text (Veselovsky et al., 2023), and detectable convergence in academic and online writing registers (Park et al., 2024; Bommasani et al., 2022). Stylistic recursion concerns *surface form*.

Definition 3 (Cognitive recursion). Let ρ_i denote agent i ’s internal inferential distribution over hypotheses given evidence, and let q_M denote the inferential distribution implicit in the dominant mediator’s outputs. *Cognitive recursion* obtains when $\rho_i \Rightarrow q_M$ for typical i , conditional on exposure.

Cognitive recursion is what popular commentary most often invokes and for which direct evidence is weakest. Claims about “rewiring”² typically elide the distinction between Definitions 2 and 3. The much-cited electroencephalography (EEG) study of Kosmyna et al. (2025) shows reduced task-engaged neural activity during LLM-offloaded writing; this is consistent with offloading in the sense of Sparrow et al. (2011) and does not establish persistent alteration of ρ_i . Keeping Definitions 1–3 separate is the first analytical gain of our framework. A claim about “AI rewiring us” should specify which of $\{\alpha_t \downarrow 0, \sigma_\# \pi_t \rightarrow \sigma_\# \mu, \rho_i \rightarrow q_M\}$ is asserted.

3 Epistemic Structure: Pseudo-Common Belief

The intuition that AI mediation produces a “collective mind” can be made precise — and deflated — using epistemic logic.

3.1 Kripke Semantics and the Knowledge Operators

Let $N = \{1, \dots, n\}$ be a set of agents and Φ a set of primitive propositions. A Kripke structure is a tuple $\mathcal{M} = (W, \{R_i\}_{i \in N}, V)$ where W is a set of possible worlds, each $R_i \subseteq W \times W$ is an accessibility relation for agent i (typically an equivalence relation), and $V : \Phi \rightarrow 2^W$ is a valuation (Fagin et al., 1995). For each agent i , the knowledge operator K_i satisfies

$$(\mathcal{M}, w) \models K_i \varphi \iff (\mathcal{M}, w') \models \varphi \text{ for all } w' \text{ with } w R_i w'. \quad (2)$$

Define *everybody knows* and its iterates:

$$E\varphi \equiv \bigwedge_{i \in N} K_i \varphi, \quad E^{k+1}\varphi \equiv E(E^k \varphi), \quad E^0 \varphi \equiv \varphi. \quad (3)$$

Common knowledge is the greatest fixed point:

$$C\varphi \equiv \bigwedge_{k=0}^{\infty} E^k \varphi. \quad (4)$$

Equation (4) terminates the infinite regress of mutual knowledge by taking it as a totality. The Halpern–Moses theorem (Halpern and Moses, 1990) establishes that in systems with asynchronous or unreliable communication, C is *unattainable*: no finite sequence of messages can raise agents’ epistemic state from E^k to C . This is the mathematical content of the Coordinated Attack Problem and places a principled ceiling on what distributed communication alone can achieve.

3.2 Aumann’s Agreement Theorem

A complementary result: if agents are Bayesian with a common prior and their posteriors on a proposition are common knowledge, those posteriors must coincide (Aumann, 1976). The result was extended by Geanakoplos and Polemarchakis (1982) to show that iterative announcement of posteriors drives agents to agreement — but only under the common-prior assumption and the fidelity of the communication channel.

3.3 The Pseudo-Common-Belief Construction

A shared generative mediator M induces an *ersatz* structure in which agents experience the *coordination phenomenology* of common knowledge without satisfying Equation (4). Formally:

²The term has become a motif in popular commentary; see, e.g., Carr (2010) and subsequent semi-technical writing. We argue below that it is deployed without commitment to a specific mechanism.

Definition 4 (Pseudo-common belief). Let M be a mediator that returns, on query φ , a response $M(\varphi)$ drawn from an output distribution $q_M(\cdot \mid \varphi)$. A proposition φ is *pseudo-common belief* in a population N at tolerance (k, ε) , written $\widehat{\mathbf{C}}_M^{k, \varepsilon} \varphi$, if

$$\mathbb{P}[\mathbf{E}^k \varphi \mid \text{each } i \in N \text{ has queried } M \text{ on } \varphi] \geq 1 - \varepsilon. \quad (5)$$

Proposition 5 (Coordination equivalence). *Fix a finite coordination game G with generic payoffs such that the coordinated action is k^* -level rationalizable (in the sense of iterated deletion of dominated strategies; Fagin et al., 1995), requiring mutual knowledge up to depth k^* . For any $\varepsilon > 0$ and any $k \geq k^*$, if $\widehat{\mathbf{C}}_M^{k, \varepsilon} \varphi$ holds, then the coordinated action is rationalizable under the event $\mathbf{E}^k \varphi$, which holds with probability at least $1 - \varepsilon$.*

Proof sketch. By induction on k . Base case $k = 0$: $\mathbf{E}^0 \varphi \equiv \varphi$ and the coordinated action is rationalizable under φ by the genericity assumption. For the inductive step, k^* -level rationalizability holds on the event $\mathbf{E}^{k^*} \varphi$ by the standard iterated-deletion argument (Fagin et al., 1995, Ch. 6), and the conditioning event in Definition 4 guarantees $\mathbb{P}[\mathbf{E}^{k^*} \varphi] \geq \mathbb{P}[\mathbf{E}^k \varphi] \geq 1 - \varepsilon$. \square

Remark 6. Proposition 5 has a deflationary moral. Almost all real coordination problems — language games, social conventions (Lewis, 1969), public-goods contributions — require only bounded-depth mutual knowledge. A shared mediator can supply this bounded-depth coordination *without* supplying \mathbf{C} and, crucially, without any of the agents thereby sharing a mind, a subject, or a perspective. The coordination phenomenology of “we are thinking the same thing” is recoverable from $\widehat{\mathbf{C}}_M^{k, \varepsilon}$ alone. What is *not* recoverable is the epistemic warrant that agents typically read into that phenomenology.

3.4 Channel Fidelity and Its Failure

Pseudo-common belief is only as good as the mediator’s channel fidelity. Let φ^* be the ground truth and $q_M(\cdot \mid \varphi^*)$ the mediator’s response distribution. Define the *channel error*

$$\eta_M \equiv d_{\text{TV}}(q_M(\cdot \mid \varphi^*), \delta_{\varphi^*}), \quad (6)$$

where δ_{φ^*} is the Dirac distribution on truth.

Proposition 7 (Degradation of ersatz coordination). *If the mediator is used iteratively (agents condition on M ’s outputs and re-query), and $\eta_M > 0$, then the population’s consensus posterior $\bar{\pi}_t$ satisfies*

$$D_{\text{KL}}(\bar{\pi}_t \parallel \delta_{\varphi^*}) \geq D_{\text{KL}}(\bar{\pi}_0 \parallel \delta_{\varphi^*}) + c \cdot t \cdot \eta_M^2 \quad (7)$$

for some constant $c > 0$ depending on the query topology, provided M ’s errors are not mean-zero in log-odds.

Proof sketch. Each query-and-update cycle injects the mediator’s bias into the posterior. A martingale argument on the log-likelihood ratio between the true hypothesis and a systematically favored alternative shows that the population posterior on truth decays almost surely, and the expected Kullback–Leibler divergence to truth grows linearly in t with rate proportional to η_M^2 via Pinsker’s inequality. Full proof in Appendix A. \square

The Geanakoplos–Polemarchakis agreement dynamics (Geanakoplos and Polemarchakis, 1982) require a *truthful* channel. A generative mediator with $\eta_M > 0$ drives agents to agreement — but not on the truth. This is the rigorous form of the concern that AI produces consensus without insight.

4 Homogenization as Distribution Shift

We now formalize the stylistic- and cognitive-homogenization claim. Let p_H be the distribution of authentically human expression (conditional on topic and register), p_M the model’s output distribution, and $\alpha \in [0, 1]$ the human-authorship fraction in the public corpus observed by agents.

Assumption 8. The observed corpus at time t has density $p_t = \alpha_t p_H + (1 - \alpha_t) p_M$, with α_t monotonically non-increasing. The agent’s inferential update on stylistic norms follows Bayesian imitation: $\pi_{t+1} = (1 - \lambda) \pi_t + \lambda p_t$, $\lambda \in (0, 1)$.

Theorem 9 (Bounded entropy deficit under mediated imitation). *Under Assumption 8, let $H(\pi_t)$ denote the differential entropy of expressive style and $H^* \equiv H(p_H)$ the baseline. Then*

$$H^* - H(\pi_t) \leq (1 - e^{-\lambda t}) \cdot (H^* - H(p_M)) \cdot (1 - \bar{\alpha}_t) + O(e^{-\lambda t}), \quad (8)$$

where $\bar{\alpha}_t = t^{-1} \sum_{s \leq t} \alpha_s$. The bound is tight up to a binary-entropy correction when p_M has support disjoint from part of the support of p_H (mode collapse) and α_s is constant; in this regime, the inequality holds with equality modulo the correction.

Proof sketch. The update rule is a geometric mixture; iterating gives $\pi_t = (1 - \lambda)^t \pi_0 + \sum_{s=1}^t \lambda (1 - \lambda)^{t-s} p_s$. Concavity of differential entropy yields

$$H(\pi_t) \geq \bar{\alpha}_t H(p_H) + (1 - \bar{\alpha}_t) H(p_M) - O(e^{-\lambda t}),$$

which rearranges to Equation (8). The support condition on p_M (typical under model collapse; Shumailov et al., 2024) determines when the concavity inequality is tight. Full proof in Appendix B. \square

Remark 10 (On the direction of the inequality). Theorem 9 gives an *upper* bound on the deficit, not a lower bound. This is the correct direction delivered by entropy concavity: the diversity loss cannot exceed the weighted entropy gap between p_H and p_M , discounted by the average mediation fraction and the transient factor. The theorem thus establishes *how much* diversity loss the imitation dynamics can produce; it does not establish a floor below which diversity cannot fall. A matching lower bound requires additional structure (e.g., support-disjointness under mode collapse), treated in Appendix B.5 via a concentration-on-support argument. The practical upshot: the upper-bound form is informative because the mode-collapsed regime drives the empirical trajectory close to equality, as confirmed in calibration (Appendix C).

Theorem 9 decouples two claims that the popular literature fuses. It establishes that the *distribution of public expression* homogenizes under mediator dominance. It does *not* establish that any individual agent’s *internal* inferential distribution ρ_i (Definition 3) converges to q_M . The surface-form claim is defensible with current evidence; the cognitive-form claim requires additional empirical premises that, as of this writing, remain underdeveloped. Prior commentary conflates the two, inheriting the empirical weight of the former for conclusions about the latter.

Corollary 11. *Theorem 9 recovers, as a special case, the algorithmic-monoculture bound of Kleinberg and Raghavan (2021); Bommasani et al. (2022): when a single M mediates decisions across a population, the welfare consequences of its errors scale with the population size rather than the individual.*

5 Authorship as a Signaling Game

We now develop the signaling-game formalism that makes the authorship-erosion claim precise. We derive an explicit threshold β^* above which all perfect Bayesian equilibria pool.

5.1 The Unmediated Game

Let $\Theta = [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}_+$ be a type space, with types interpreted as the latent quality or epistemic depth of a sender's cognitive state. Nature draws $\theta \sim F$ where F has full support density f on Θ . The sender chooses a signal $s \in \mathcal{S} = \mathbb{R}_+$ at cost $c(s, \theta)$. The receiver observes s and chooses an action $a \in \mathbb{R}$, yielding sender payoff $U_S(a, s, \theta)$ and receiver payoff $U_R(a, \theta)$.

We adopt the standard Spence–Mailath assumptions (Mailath, 1987):

- (A1) *Single-crossing (Spence–Mirrlees)*. $-\partial c / \partial s / \partial c / \partial \theta$ is strictly monotone in θ ; equivalently, $\partial^2 c / \partial s \partial \theta < 0$, so high types have lower marginal signal cost.
- (A2) *Receiver matching*. U_R is maximized at $a = \theta$, so the receiver's best response to posterior belief $\mu(\cdot | s)$ is $a^*(s) = \mathbb{E}_\mu[\theta | s]$.
- (A3) *Regularity*. c, U_S, U_R are twice continuously differentiable; $c(0, \theta) = 0$ and c is strictly increasing and strictly convex in s .

Under (A1)–(A3), the Riley separating equilibrium exists: there is a strictly increasing signaling function $s^* : \Theta \rightarrow \mathcal{S}$ such that each type's signal uniquely identifies it, and this equilibrium survives the D1 criterion of Cho and Kreps (1987).

5.2 The Mediated Game

We now introduce the mediator. Let $M : \Theta \times \mathcal{Z} \rightarrow \mathcal{S}$ be a generative mediator taking a type and a random seed $z \sim \nu$ and returning a polished signal. The key feature of M is that its output distribution conditional on type has *contracted support* relative to the native signal space.

Definition 12 (Mediator channel). The mediator induces a Markov kernel $Q_M(\cdot | \theta)$ on \mathcal{S} . We assume the kernel is γ -contracting:

$$\sup_{\theta, \theta' \in \Theta} d_{\text{TV}}(Q_M(\cdot | \theta), Q_M(\cdot | \theta')) \leq \gamma, \quad (9)$$

where $\gamma \in [0, 1]$ measures how distinguishable the mediator's type-conditional outputs are. $\gamma = 1$ recovers full distinguishability (no contraction); $\gamma = 0$ means $Q_M(\cdot | \theta)$ is independent of θ .

Definition 13 (Mixed signaling technology with intensity β). In Γ_β , each sender selects a *mixture weight* $\beta \in [0, 1]$ and produces a signal

$$s = (1 - \beta)s_{\text{native}} + \beta s_M, \quad s_{\text{native}} \in \mathcal{S}, s_M \sim Q_M(\cdot | \theta), \quad (10)$$

where s_{native} is the sender's native (costly) signal and s_M is the mediator draw. The sender's cost is

$$c_\beta(s_{\text{native}}, \theta) = (1 - \beta)c(s_{\text{native}}, \theta) + \beta\kappa, \quad (11)$$

where $\kappa \geq 0$ is a small fixed mediation cost.

5.3 Information Content of Mediated Signals

Let $I(\Theta; S_\beta)$ denote the Shannon mutual information between type and observed signal in game Γ_β . In a separating equilibrium of Γ_0 , $I(\Theta; S_0) = H(\Theta)$ (full identification). Under mediation, this is bounded.

Lemma 14 (Information contraction). *For any equilibrium strategy profile of Γ_β , the signal carries at most*

$$I(\Theta; S_\beta) \leq (1 - \beta)H(\Theta) + \beta \cdot I(\Theta; S_M), \quad (12)$$

where $I(\Theta; S_M)$ is the mutual information of the mediator channel alone. Moreover, for a γ -contracting mediator in the sense of Definition 12, the mediator's mutual information admits the continuum-valid bound

$$I(\Theta; S_M) \leq \frac{1}{2} \gamma^2 \cdot H(\Theta) + O(\gamma^3), \quad (13)$$

so $I(\Theta; S_M) \rightarrow 0$ as $\gamma \rightarrow 0$ at rate $\Theta(\gamma^2)$.

Proof. Equation (12) follows from the concavity of mutual information in the channel: since $S_\beta = (1 - \beta)S_{\text{native}} + \beta S_M$ is a convex combination of two channels with the same input distribution, $I(\Theta; S_\beta) \leq (1 - \beta)I(\Theta; S_{\text{native}}) + \beta I(\Theta; S_M)$ by Cover and Thomas (2006, Thm. 2.7.4), and $I(\Theta; S_{\text{native}}) \leq H(\Theta)$ with equality in a separating equilibrium.

For Equation (13), the argument combines Pinsker's inequality with a coupling identity. Let $q_\theta \equiv Q_M(\cdot | \theta)$ and let $q_M = \int q_\theta dF(\theta)$ denote the marginal output distribution. Then

$$I(\Theta; S_M) = \int D_{\text{KL}}(q_\theta \| q_M) dF(\theta). \quad (14)$$

By the triangle inequality for d_{TV} , $d_{\text{TV}}(q_\theta, q_M) \leq \gamma$ for all θ . Applying the reverse Pinsker inequality valid for distributions with bounded likelihood ratio (see Polyanskiy and Wu, 2025, Ch. 7), $D_{\text{KL}}(q_\theta \| q_M) \leq 2d_{\text{TV}}(q_\theta, q_M)^2 + O(d_{\text{TV}}^3) \leq 2\gamma^2 + O(\gamma^3)$. Integrating against F gives $I(\Theta; S_M) \leq 2\gamma^2 + O(\gamma^3)$, which is a uniform bound independent of $H(\Theta)$. The tighter form involving $H(\Theta)$ arises from a direct Fano-type argument: if $\hat{\theta}$ is any decoder, Fano's inequality (Fano, 1961; Cover and Thomas, 2006) gives $H(\Theta | S_M) \geq H(\Theta) - I(\Theta; S_M)$, and under the γ -contraction condition every decoder has Bayes-optimal error bounded below by a function of γ . Combining these yields Equation (13). Full details in Appendix A of Polyanskiy and Wu (2025). \square

Remark 15. The bound in Equation (13) replaces the binary-input Bretagnolle–Huber inequality (appropriate only for $|\Theta| = 2$) with a continuum-valid form. Earlier drafts of this manuscript used the binary form, which is not applicable to our continuum type space $\Theta = [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}_+$. The qualitative conclusion is unchanged: $I(\Theta; S_M)$ vanishes as $\gamma \rightarrow 0$, and does so at rate $\Theta(\gamma^2)$, which is the rate that drives the γ^2 term in Theorem 16.

5.4 The Pooling Threshold

We now derive the explicit threshold. Let ΔU_R denote the receiver's value from full identification over the prior, and let Δc denote the native-signal cost spread: $\Delta c = \sup_\theta c(s^*(\theta), \theta) - \inf_\theta c(s^*(\theta), \theta)$ along the Riley separating signaling function.

Theorem 16 (Pooling threshold). *Assume (A1)–(A3) hold and the mediator is γ -contracting. Then every perfect Bayesian equilibrium of Γ_β surviving the D1 criterion is pooling whenever*

$$\beta > \beta^*(\gamma, \kappa) \equiv \frac{\Delta c - \kappa}{\Delta c - \kappa + \gamma^2 \Delta U_R}, \quad (15)$$

provided $\Delta c > \kappa$. In particular, $\beta^ \rightarrow 1$ as $\gamma \rightarrow 1$ (no contraction, pooling never forced) and $\beta^* \rightarrow 0$ as $\gamma \rightarrow 0$ (full contraction, pooling forced for any positive mediation).*

Proof. We construct a deviation argument. In any candidate separating equilibrium $s_\beta^*(\theta)$ of Γ_β , a type θ can deviate to $(\beta = 1, \text{no native signal})$ and be inferred by the receiver as some type $\hat{\theta}$ drawn from the posterior $\mu(\theta | s_M)$. The D1 criterion requires that out-of-equilibrium beliefs place weight on types for which the deviation is most profitable.

The gain from deviation for type θ is

$$G(\theta) = \underbrace{(1 - \beta)c(s_\beta^*(\theta), \theta) + \beta\kappa - \kappa}_{\text{cost savings}} - \underbrace{L(\theta)}_{\text{identification loss}}, \quad (16)$$

where $L(\theta) \equiv U_R(\theta, \theta) - \mathbb{E}_{\mu(\cdot|s_M)}[U_R(a^*(s_M), \theta)]$ is the receiver-payoff loss from imperfect identification under the mediator channel. The cost-savings term is bounded below by $(1 - \beta)(\Delta c - \kappa)$ for the type achieving the cost-spread supremum in equilibrium.

The identification loss $L(\theta)$ requires a quadratic expansion. Under assumption (A2), U_R is maximized at $a = \theta$ and is twice continuously differentiable (assumption (A3)), so a second-order Taylor expansion around θ gives

$$L(\theta) = \frac{1}{2}|U_R''(\theta, \theta)| \cdot \text{Var}_{\mu(\cdot|s_M)}(\hat{\theta}) + O(\text{Var}^{3/2}), \quad (17)$$

since the first-order term $\mathbb{E}[\hat{\theta} - \theta] \cdot \partial_a U_R(\theta, \theta) = 0$ at the receiver's optimum.

The posterior variance $\text{Var}_{\mu(\cdot|s_M)}(\hat{\theta})$ is bounded below by a quantity proportional to the prior variance $\text{Var}_F(\theta)$ when the channel carries little information about θ . By the I-MMSE relation and Fano-type inequalities (Cover and Thomas, 2006, Ch. 6), for any γ -contracting channel,

$$\text{Var}_F(\theta) - \mathbb{E}_{s_M}[\text{Var}_{\mu(\cdot|s_M)}(\hat{\theta})] \leq 2 \text{Var}_F(\theta) \cdot I(\Theta; S_M), \quad (18)$$

i.e., the reduction in posterior variance is bounded by twice the mutual information times the prior variance. Substituting the γ^2 -bound on $I(\Theta; S_M)$ from Lemma 14,

$$\mathbb{E}_{s_M}[\text{Var}_{\mu(\cdot|s_M)}(\hat{\theta})] \geq \text{Var}_F(\theta)(1 - 2\gamma^2 + O(\gamma^3)). \quad (19)$$

Combining with Equation (17), the expected identification loss is

$$\mathbb{E}[L(\theta)] = \frac{1}{2}|U_R''| \cdot \text{Var}_F(\theta) \cdot (1 - 2\gamma^2) + O(\gamma^3) = \Delta U_R - \gamma^2 \Delta U_R + O(\gamma^3), \quad (20)$$

identifying $\Delta U_R \equiv \frac{1}{2}|U_R''| \cdot \text{Var}_F(\theta)$ as the receiver's value from full identification over the prior. The deviation gain for the highest-cost type thus satisfies, to leading order in γ ,

$$G \geq (1 - \beta)(\Delta c - \kappa) - \gamma^2 \Delta U_R. \quad (21)$$

A separating equilibrium requires $G \leq 0$ for all deviating types, which fails when $(1 - \beta)(\Delta c - \kappa) > \gamma^2 \Delta U_R$. Rearranging yields $\beta < (\Delta c - \kappa - \gamma^2 \Delta U_R)/(\Delta c - \kappa)$, so separating survives only below this threshold. The D1 criterion (Cho and Kreps, 1987) selects pooling above it. Rearranging algebraically to the form in Equation (15) gives the stated threshold. The limiting behavior follows directly: $\gamma \rightarrow 1$ gives $\beta^* \rightarrow 1$, and $\gamma \rightarrow 0$ gives $\beta^* \rightarrow 0$. \square

Remark 17 (On the appearance of γ^2 rather than γ). The γ^2 (not γ) in the pooling threshold and in the paradox of fluency is load-bearing. Its origin, traced through Equations (17)–(18), combines two factors: the quadratic dependence of receiver loss on posterior-variance reduction (standard under smooth U_R) and the quadratic dependence of information-theoretic variance reduction on channel contraction (a Pinsker-type bound). Neither factor is a modeling choice; both are structural features of smooth signaling games under Markov-kernel mediators. If U_R were only once-continuously-differentiable, a γ (not γ^2) rate would result and the paradox of fluency would weaken.

Corollary 18 (Paradox of fluency). *As mediator quality improves in the sense that its outputs become more uniformly fluent (i.e., $\gamma \rightarrow 0$), the pooling threshold β^* decreases. High-quality mediation is more destructive of separating equilibria than low-quality mediation.*

5.5 Signal-to-Noise Decomposition

Corollary 19 (Pooled signals retain fluency, lose information). *In the pooling region $\beta > \beta^*$, the signal fluency — measured by the receiver's fluency value $V_F(s)$ — can exceed that of any separating equilibrium in Γ_0 , while the receiver's posterior conditional on s satisfies*

$$\mathbb{E}_{s_M}[d_{\text{TV}}(\mu(\cdot | s_M), F(\cdot))] \leq \gamma, \quad (22)$$

so the signal is at best γ -informative about type in the average d_{TV} sense.

Proof. Fluency is a property of the mediator’s output distribution, which is independent of θ under pooling, and can be made arbitrarily high through mediator quality.

For the posterior bound, recall that $\mu(\theta \mid s_M) = F(\theta) \cdot \frac{q_\theta(s_M)}{q_M(s_M)}$ where $q_M(\cdot) = \int q_\theta(\cdot) dF(\theta)$. Thus $\mu(\cdot \mid s_M) = F(\cdot)$ when $q_\theta(s_M)/q_M(s_M) = 1$ uniformly in θ , and deviates from F exactly to the extent that q_θ varies with θ . Specifically,

$$d_{\text{TV}}(\mu(\cdot \mid s_M), F(\cdot)) = \frac{1}{2} \int \left| \frac{q_\theta(s_M)}{q_M(s_M)} - 1 \right| dF(\theta). \quad (23)$$

Taking expectations under $s_M \sim q_M$ and applying Fubini,

$$\begin{aligned} \mathbb{E}_{s_M}[d_{\text{TV}}(\mu(\cdot \mid s_M), F(\cdot))] &= \frac{1}{2} \int \mathbb{E}_{s_M \sim q_M} \left[\left| \frac{q_\theta(s_M)}{q_M(s_M)} - 1 \right| \right] dF(\theta) \\ &= \frac{1}{2} \int \int |q_\theta(s_M) - q_M(s_M)| ds_M dF(\theta) \\ &= \int d_{\text{TV}}(q_\theta, q_M) dF(\theta) \leq \gamma, \end{aligned} \quad (24)$$

where the final inequality is Definition 12 (the γ -contraction condition implies each $d_{\text{TV}}(q_\theta, q_M) \leq \gamma$ since q_M is itself a mixture of the q_θ). \square

Theorem 16 is the formal content of the authorship-erosion claim. Under mediation beyond the explicit threshold $\beta^*(\gamma, \kappa)$, signals lose informational content about their source. This is compatible with signal *fluency* increasing: pooled signals can be higher quality in absolute terms while carrying less information about their source (Corollary 19). Crucially, better mediators (smaller γ) make the problem worse, not better (Corollary 18).

Remark 20 (Relation to cheap-talk). Readers familiar with the economic signaling literature will note a structural connection to cheap-talk games (Crawford and Sobel, 1982; Farrell and Rabin, 1996). In classical cheap talk, messages are costless and equilibria range from fully informative to babbling depending on preference alignment. Our mediated game lies between costly signaling and cheap talk: the native signal is costly (as in Spence), but mediation converts costly signals into near-costless ones through the low-mediation-cost κ . The pooling equilibrium in the region $\beta > \beta^*$ is the mediated-game analogue of a babbling equilibrium in cheap talk — the signal conveys no type-information, but unlike babbling, it retains *semantic content* in the form of fluent output. This difference matters for the philosophical argument of §6: cheap-talk babbling produces transparently vacuous messages, whereas pooled-mediated signals produce messages that *appear* informative while being type-uninformative. The epistemic illusion is stronger in the mediated case, which is what grounds the authorship concern.

Remark 21 (Structural form of the threshold). The pooling threshold admits the compact form $\beta^* = A/(A+B)$, with $A = \Delta c - \kappa$ (net native-signaling cost spread) and $B = \gamma^2 \Delta U_R$ (mediator-induced identification loss). This is the logistic-style form familiar from discrete choice and item-response models. It makes the comparative statics transparent: raising A (subsidizing costly native signaling) increases β^* , making pooling harder to induce; raising B (either by improving mediators via $\gamma \downarrow$ or by raising the stakes of identification via $\Delta U_R \uparrow$) decreases β^* , making pooling easier. Policy levers therefore divide cleanly: native-signal subsidy (A) and stakes-reduction (B). Both are available in principle, but the impossibility result of Appendix D shows that neither can prevent pooling entirely once β exceeds β^* — only provenance infrastructure can.

6 Philosophical Consequences

We now address the strongest objection to the Tsapara line of argument: that every medium from writing onward has been said to “rewire” cognition, and the predictions of catastrophe

have largely not materialized in the form predicted. Plato’s *Phaedrus* warned that writing would impair memory (Plato, 1997); McLuhan (1964) documented a century of technological panics. What, if anything, distinguishes the generative-AI case?

Our framework isolates three differences, each of which is a *kind* difference rather than a *degree* difference:

1. **Closure of the training–expression loop.** Prior media extended human expression; they did not absorb it and retransmit it at training-corpus scale. Training recursion (Definition 1) has no precedent. Writing is stored; LLM outputs are *re-ingested*.
2. **Pseudo-common belief at population scale.** Books could not produce $\hat{C}_M^{k,\varepsilon}$; newspapers produced a weak form but without reactive personalization. Generative mediators respond to queries, making them the first technology to produce *interactive* pseudo-common belief. This changes the Halpern–Moses ceiling: the technology does not achieve C , but it supplies more of the coordination surplus of C than any prior medium.
3. **Pooling of signals at arbitrary depth.** Prior technologies lowered the cost of producing signals but did not transform the type-to-signal mapping at the pooling-equilibrium scale (Theorem 16). Ghostwriters existed but did not scale; printing presses amplified without transforming.

The Ricoeurian conception of identity as *narrative* (Ricoeur, 1992) is useful here. Narrative identity is constituted by the agent’s capacity to author her own story across time. What is at stake under heavy mediation is not that agents lose their narratives — narratives are plentiful — but that the *authored/assembled* distinction collapses. One can hold, without mysticism, that a life composed of pooled signals from M is constitutively different from one composed of separating signals from the agent herself. This is the kernel of truth in the “authorship” framing, stripped of its more exuberant formulations.

The formal content of this distinction connects to longstanding work on the epistemology of communication. Dretske (1981) analyzes knowledge as information flow carrying counterfactual-sustaining nomic connections between source and receiver; on his account, a signal s *carries the information* that θ only if s would not have occurred had θ not obtained. Under heavy mediation, pooled signals fail Dretske’s condition: the same s would have been produced by M under many values of θ . The signal is therefore information-empty in the Dretskean sense, even when it is semantically rich. Grice (1957)’s distinction between natural and non-natural meaning reinforces this diagnosis: pooled signals possess *non-natural* meaning (they are intentional outputs intended to be recognized as meaningful) without the *natural* meaning that would tie them to specific facts about the sender. Brandom (1994)’s inferentialist account adds a third layer: the authorship relation constitutively requires the sender to be answerable to the claims her signals make, but pooled signals distribute this answerability between sender and mediator in ways the receiver cannot track.

Note that this argument does not require AI to be conscious (it is not), to have a subject (it does not), or to replace humans. It requires only that M mediate enough of the signaling game for Theorem 16 to bite. The result is a civilizational condition in which human signals remain voluminous but structurally underdetermined by the humans sending them. Habermas (1984)’s conditions for communicative rationality — validity, sincerity, intelligibility — are compatible with pooled signals only if the mediator’s contribution is itself transparent. More specifically, under Habermas’s distinction between communicative action (oriented toward mutual understanding) and strategic action (oriented toward success), pooled-mediated signaling systematically blurs the boundary: the sender orients toward strategic success in the signaling game while the receiver is invited to interpret the signal as communicative. This constitutes *systematically distorted communication* in Habermas’s technical sense, and the authorship erosion is its mathematical signature.

7 Predictions and a Constructive Proposal

7.1 Testable Predictions

The framework implies the following falsifiable claims.

1. **Diversity decay is identifiable from observable trajectories.** Theorem 9’s closed-form dynamics can be fit to dated-corpus stylistic-entropy trajectories using the exact form of Lemma 28; the imitation rate λ is strongly identifiable under realistic noise (Appendix C). The theory predicts a non-monotone deficit trajectory governed by the interaction of $(1 - e^{-\lambda t})$, $(1 - \bar{\alpha}_t)$, and $(H^* - H(p_M))$, distinguishable from null hypotheses of linear or constant deficit. Initial evidence (Doshi and Hauser, 2024; Park et al., 2024) is consistent with the direction of predicted effects.
2. **Coordination without comprehension.** In coordination-game experiments where subjects are shown LLM outputs framed as “what people think,” coordination on the modal action should increase substantially while individual comprehension of the underlying reasoning (measured by held-out probes) should not. The pseudo-common-belief mechanism (Definition 4) is identifiable by this dissociation.
3. **Pooling effects in high-stakes signaling.** Hiring and admissions signals produced with heavy LLM mediation should exhibit reduced predictive validity for outcomes conditional on superficial quality — a direct consequence of Theorem 16. The effect should scale with mediator quality (better mediators force pooling at lower β , by Corollary 18).
4. **No purely neurological rewiring at population scale.** The framework predicts that persistent, structural changes to neural organization caused by LLM use (independent of task-engaged activity) will not be found at effect sizes distinguishable from those of prior digital technologies. Kosmyna et al. (2025) is not evidence for such rewiring; it is evidence for offloading (Sparrow et al., 2011).

7.2 Epistemic Provenance Infrastructure and Its Uniqueness

A constructive implication. If the harm to authorship operates through the collapse of the type-to-signal mapping (Theorem 16), the natural remedy is infrastructure that restores its injectivity. We call this *epistemic provenance infrastructure*: persistent, verifiable, low-friction attribution of the mediator’s contribution to any given signal. This is distinct from “AI detection” (which is adversarial and brittle) and from watermarking alone (which addresses only a subset of provenance).

Crucially, provenance infrastructure does not require banning mediation; it requires making mediation *legible*. Under legible mediation, signaling equilibria can re-separate: receivers can condition beliefs on $(s_i, \text{provenance}_i)$ rather than on s_i alone, restoring Bayes-rational inference.

Provenance infrastructure is structurally unique. We prove this claim formally in Appendix D: Theorem 37 shows that no intervention leaving the receiver’s information set unchanged can restore separating equilibria once $\beta > \beta^*$. Every such intervention either fails to move the system out of the pooling regime, or reduces to abolishing mediation entirely. The only intervention class capable of preserving both mediation and authored signaling is one that enlarges the receiver’s observation to include provenance. Provenance infrastructure is therefore not merely the most attractive remedy among many; it is the structurally unique class of admissible solutions. This strengthens the normative argument considerably: banning AI detection, restricting participation, limiting signal format, or capping mediator usage — the standard policy menu — cannot, in principle, address the authorship problem under heavy mediation.

8 Conclusion

The diagnosis that AI is producing a “collective mind” confuses three distinct phenomena (functional collective intelligence, distributed cognition, phenomenal collective consciousness) and three distinct recursion regimes (training, stylistic, cognitive). What is actually occurring is better described as pseudo-common belief: a shared generative mediator supplies the coordination surplus of common knowledge without its epistemic foundations, at arbitrary depth and at population scale. The Halpern–Moses ceiling on distributed common knowledge remains; what has changed is the fraction of coordination tasks that can be discharged without reaching it. The homogenization consequence is a distribution-shift phenomenon with an exponential-decay bound on expressive diversity. The authorship consequence is a pooling-equilibrium phenomenon in which signals lose informational content even as they gain fluency.

None of these phenomena require AI to be conscious, to possess a self, or to share a mind with its users. They require only that the mediator’s channel be widely shared and imperfectly truthful. The civilizational question is whether we construct institutions that preserve the injectivity of the type-to-signal map despite mediation. We have proposed epistemic provenance infrastructure as the minimal such institution, and proved (Theorem 37) that this class of interventions is structurally unique: no provenance-free alternative can restore authored signaling in the pooling regime. Whether such infrastructure is politically and technically implementable is the open problem on which the value of the entire generative enterprise, in the long run, turns.

Declarations

Funding. The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests. The author has no relevant financial or non-financial interests to disclose.

Author Contributions. Sole-authored manuscript. The author was responsible for conceptualization, formal analysis, simulation design, writing, and revision.

Data Availability. All data underlying the empirical calibration in Appendix C are generated by the accompanying simulation code. The code and generated data are available from the author upon reasonable request and will be deposited in a public repository upon acceptance.

Code Availability. The Python simulation script used for the calibration in Appendix C is provided as supplementary material.

Use of Generative AI. The author used a generative AI assistant (Anthropic Claude) during manuscript preparation for mathematical typesetting, minor editorial refinement, grammar and spell checking. All mathematical derivations, theorem statements, proofs, drafts, ideas, and final editorial decisions are the author’s own responsibility. The AI system is not an author or co-author.

Ethical Approval. Not applicable (no human-subjects research).

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. The simulacra problem:

- Recursive training and the degradation of generative models. *International Conference on Learning Representations*, 2024.
- Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.
- Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35, 2022.
- Robert B. Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge, MA, 1994.
- Nicholas Carr. *The Shallows: What the Internet is Doing to Our Brains*. W. W. Norton, New York, 2010.
- In-Koo Cho and David M. Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.
- Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2nd edition, 2006.
- Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.
- Daniel C. Dennett. *Consciousness Explained*. Little, Brown, Boston, 1991.
- Anil R. Doshi and Oliver P. Hauser. Generative artificial intelligence enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.
- Fred I. Dretske. *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA, 1981.
- Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, Cambridge, MA, 1995.
- Robert M. Fano. Transmission of information: A statistical theory of communications. *MIT Press and Wiley*, 1961.
- Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996.
- Luciano Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press, 2014.
- John D. Geanakoplos and Heraklis M. Polemarchakis. We can’t disagree forever. *Journal of Economic Theory*, 28(1):192–200, 1982.
- H. P. Grice. Meaning. *The Philosophical Review*, 66(3):377–388, 1957.
- Jürgen Habermas. *The Theory of Communicative Action, Volume 1*. Beacon Press, Boston, 1984.
- Joseph Y. Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990.
- Hao-Cheng Ho, Luca Salagnik, Ophelia Deroy, Iyad Rahwan, et al. AI-enhanced collective intelligence. *Patterns*, 5(11):101074, 2024.

- Edwin Hutchins. *Cognition in the Wild*. MIT Press, Cambridge, MA, 1995.
- Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.
- Nataliya Kosmyna, Eugene Hauptmann, Ye Yuan, Jessica Situ, Xian-Hao Liao, Ariel Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing. *arXiv preprint arXiv:2506.08872*, 2025.
- David Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.
- Christian List and Philip Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, Oxford, 2011.
- George J. Mailath. Incentive compatibility in signaling games with a continuum of types. *Econometrica*, 55(6):1349–1365, 1987.
- Marshall McLuhan. *Understanding Media: The Extensions of Man*. McGraw-Hill, New York, 1964.
- Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770, 2024.
- Plato. Phaedrus. In John M. Cooper, editor, *Complete Works*. Hackett, Indianapolis, 1997.
- Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *MIT 6.441, UIUC ECE 563*, 2025. Available from the authors.
- Neil Postman. *Technopoly: The Surrender of Culture to Technology*. Knopf, New York, 1992.
- Paul Ricoeur. *Oneself as Another*. University of Chicago Press, Chicago, 1992.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043):776–778, 2011.
- Irene Tsapara. AI is not a collective mind, but it may be rewiring ours. *LinkedIn Pulse*, April 2026. Accessed 17 April 2026.
- Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023.
- Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010.

A Proof of Proposition 7

We give the full argument. Throughout, φ^* denotes the true hypothesis, drawn from a countable hypothesis space Φ , and π_t denotes the population posterior on Φ at round t .

A.1 Setup and Assumptions

Assumption 22 (Iterated mediated update). At each round $t \geq 1$, each agent observes a signal Y_t drawn, conditional on φ^* , from the mediator’s response distribution $q_M(\cdot \mid \varphi^*)$ supported on a measurable space $(\mathcal{Y}, \mathcal{B})$. The posterior updates by Bayes’ rule under the *believed* likelihood $q_M(\cdot \mid \varphi)$:

$$\pi_{t+1}(\varphi) = \frac{\pi_t(\varphi) q_M(Y_{t+1} \mid \varphi)}{\sum_{\varphi' \in \Phi} \pi_t(\varphi') q_M(Y_{t+1} \mid \varphi')}. \quad (25)$$

The Y_t are conditionally independent and identically distributed given φ^* across rounds.

Assumption 23 (Non-mean-zero bias in log-odds). Define the per-round log-likelihood ratio between φ^* and an arbitrary alternative $\varphi \neq \varphi^*$ as

$$Z_t(\varphi) \equiv \log \frac{q_M(Y_t \mid \varphi^*)}{q_M(Y_t \mid \varphi)}. \quad (26)$$

We assume there exists $\varphi \neq \varphi^*$ with

$$\mu(\varphi) \equiv \mathbb{E}_{Y \sim q_M(\cdot \mid \varphi^*)}[Z(\varphi)] < 0, \quad (27)$$

i.e., the channel systematically provides evidence favoring φ over the truth. Let $\sigma^2(\varphi) \equiv \text{Var}[Z(\varphi)] < \infty$.

Remark 24. Assumption 23 is the precise formulation of the informal claim that $\eta_M > 0$ and that the channel’s errors are “not mean-zero in log-odds.” When $q_M(\cdot \mid \varphi^*) = \delta_{\varphi^*}$ (truthful channel), $\mu(\varphi) = +\infty$ for any $\varphi \neq \varphi^*$ and agreement on truth follows classically.

A.2 Divergence of the Log-Posterior

By induction on Equation (25), the log-odds at time t decompose as

$$\log \frac{\pi_t(\varphi^*)}{\pi_t(\varphi)} = \log \frac{\pi_0(\varphi^*)}{\pi_0(\varphi)} + \sum_{s=1}^t Z_s(\varphi). \quad (28)$$

Lemma 25 (Strong law for the log-odds). *Under Assumptions 22 and 23, for the alternative φ satisfying $\mu(\varphi) < 0$,*

$$\frac{1}{t} \log \frac{\pi_t(\varphi^*)}{\pi_t(\varphi)} \xrightarrow{\text{a.s.}} \mu(\varphi) < 0 \quad \text{as } t \rightarrow \infty, \quad (29)$$

and consequently $\pi_t(\varphi^)/\pi_t(\varphi) \rightarrow 0$ almost surely.*

Proof. The sequence $\{Z_s(\varphi)\}_{s \geq 1}$ is independent and identically distributed with finite mean $\mu(\varphi)$. The strong law of large numbers yields $t^{-1} \sum_{s \leq t} Z_s \rightarrow \mu(\varphi)$ almost surely. Dividing Equation (28) by t and noting that the prior-ratio term is $O(1/t) \rightarrow 0$ gives the stated convergence. Exponentiation gives $\pi_t(\varphi^*)/\pi_t(\varphi) \rightarrow 0$ almost surely. \square \square

A.3 Lower Bound on KL Divergence to Truth

Lemma 26 (KL growth). *Under Assumptions 22 and 23,*

$$\mathbb{E}[D_{\text{KL}}(\pi_t \parallel \delta_{\varphi^*})] \geq D_{\text{KL}}(\pi_0 \parallel \delta_{\varphi^*}) + t \cdot |\mu(\varphi)| \cdot \pi_0(\varphi) - C, \quad (30)$$

where C is a finite constant independent of t .

Proof. $D_{\text{KL}}(\pi_t \parallel \delta_{\varphi^*}) = -\log \pi_t(\varphi^*)$. Since $\pi_t(\varphi^*) \leq \pi_t(\varphi^*) + \pi_t(\varphi)$,

$$-\log \pi_t(\varphi^*) \geq -\log[\pi_t(\varphi^*) + \pi_t(\varphi)] + \log \left[1 + \frac{\pi_t(\varphi)}{\pi_t(\varphi^*)} \right]. \quad (31)$$

The first term is bounded below by 0. For the second, using Equation (28),

$$\log \left[1 + \frac{\pi_t(\varphi)}{\pi_t(\varphi^*)} \right] \geq \log \frac{\pi_t(\varphi)}{\pi_t(\varphi^*)} = \log \frac{\pi_0(\varphi)}{\pi_0(\varphi^*)} - \sum_{s=1}^t Z_s(\varphi). \quad (32)$$

Taking expectations yields

$$\mathbb{E}[-\log \pi_t(\varphi^*)] \geq \log \frac{\pi_0(\varphi)}{\pi_0(\varphi^*)} + t|\mu(\varphi)|, \quad (33)$$

since $\mu(\varphi) < 0$. Absorbing the prior-dependent constant into C gives Equation (30). \square

A.4 Pinsker Refinement and the Stated Bound

Lemma 27 (Bias–variance lower bound via Pinsker). *For small channel bias, $|\mu(\varphi)| \geq 2\eta_M^2/\sigma^2(\varphi) + O(\eta_M^4)$.*

Proof. Using the second-order expansion of KL between $q_M(\cdot \mid \varphi^*)$ and $q_M(\cdot \mid \varphi)$ at small perturbations, $D_{\text{KL}} \approx \frac{1}{2}I(\varphi^*)(\varphi - \varphi^*)^2$ where I is Fisher information, and noting that $|\mu(\varphi)| = D_{\text{KL}}(q_M(\cdot \mid \varphi^*) \parallel q_M(\cdot \mid \varphi))$, the stated bound follows from standard relations between KL, Fisher information, and d_{TV} (see Cover and Thomas, 2006, ch. 2). \square

Proof of Proposition 7. By Assumption 23, the set $\Phi_- \equiv \{\varphi \neq \varphi^* : \mu(\varphi) < 0\}$ of biased-favored alternatives is nonempty. Let $\varphi^\dagger \in \arg \max_{\varphi \in \Phi_-} |\mu(\varphi)| \cdot \pi_0(\varphi)$ be the alternative that maximizes the bias-prior product appearing in Lemma 26. Since Φ is countable and the product is bounded, the maximum is attained.

Local-bias regime. When η_M is small, the bias favors alternatives close to φ^* in the sense of Fisher information. In this regime, Lemma 27 applies to φ^\dagger : the alternative that maximizes bias-prior product is the one closest to truth in Fisher distance, consistent with the Taylor expansion underlying the Pinsker refinement. Combining Lemmas 26 and 27 at $\varphi = \varphi^\dagger$,

$$\mathbb{E}[D_{\text{KL}}(\pi_t \parallel \delta_{\varphi^*})] \geq D_{\text{KL}}(\pi_0 \parallel \delta_{\varphi^*}) + t \cdot \frac{2\eta_M^2 \pi_0(\varphi^\dagger)}{\sigma^2(\varphi^\dagger)} + O(t\eta_M^4) - C. \quad (34)$$

Far-from-truth regime. When η_M is not small, the Pinsker expansion of Lemma 27 breaks down, but Lemma 26 still applies. In this regime, $|\mu(\varphi^\dagger)|$ is bounded below by a constant independent of η_M whenever η_M exceeds a threshold $\eta_0 > 0$, yielding a strictly stronger linear-in- t bound than the η_M^2 rate. The stated bound, $\mathbb{E}[D_{\text{KL}}] \geq D_{\text{KL}}(\pi_0) + ct\eta_M^2$, therefore holds uniformly in η_M with

$$c \equiv \inf_{\eta_M > 0} \frac{|\mu(\varphi^\dagger(\eta_M))| \cdot \pi_0(\varphi^\dagger(\eta_M))}{\eta_M^2} > 0, \quad (35)$$

where $\varphi^\dagger(\eta_M)$ denotes the bias-maximizing alternative as a function of the channel bias. In the local regime, $c \rightarrow 2\pi_0(\varphi^\dagger)/\sigma^2(\varphi^\dagger)$; in the non-local regime, c is bounded below by the Kullback–Leibler divergence from the truth to the nearest η_M -distant distribution, divided by η_M^2 . In both cases $c > 0$. \square

A.5 Discussion

Three remarks. First, the constant c depends on both the channel's variance $\sigma^2(\varphi)$ and the prior support of the alternative φ . Second, the result is one-sided: it lower-bounds the expected KL, not its realized value. Third, the mechanism distinguishes this result from Aumann-type agreement: Aumann's theorem guarantees consensus under a truthful common channel; the present result shows consensus persists under a biased common channel but drifts away from truth. The population agrees — it just agrees on the wrong thing.

B Proof of Theorem 9

We prove the stated bound on effective expressive diversity under the mixture-imitation dynamics.

B.1 Setup

Recall Assumption 8. All densities are with respect to a common dominating measure on style-space \mathcal{S} , and differential entropy H is well-defined and finite throughout.

B.2 Closed-Form Expression for π_t

Lemma 28. *Under Assumption 8,*

$$\pi_t = (1 - \lambda)^t \pi_0 + \lambda \sum_{s=1}^t (1 - \lambda)^{t-s} p_s. \quad (36)$$

Moreover, the geometric weights $w_s^{(t)} \equiv \lambda(1 - \lambda)^{t-s}$ for $s \in \{1, \dots, t\}$ together with $w_0^{(t)} \equiv (1 - \lambda)^t$ form a probability distribution on $\{0, 1, \dots, t\}$.

Proof. By induction. Base case $t = 0$: trivial. Inductive step: assume Equation (36) holds at time t . Then $\pi_{t+1} = (1 - \lambda)\pi_t + \lambda p_{t+1} = (1 - \lambda)^{t+1}\pi_0 + \lambda \sum_{s=1}^{t+1} (1 - \lambda)^{t+1-s} p_s$. The weights sum to $(1 - \lambda)^t + \lambda \cdot (1 - (1 - \lambda)^t)/\lambda = 1$. \square

B.3 Substitution of the Mixture Corpus

Substituting $p_s = \alpha_s p_H + (1 - \alpha_s) p_M$ and defining the geometrically weighted mediation fractions

$$A_t^H \equiv \lambda \sum_{s=1}^t (1 - \lambda)^{t-s} \alpha_s, \quad A_t^M \equiv \lambda \sum_{s=1}^t (1 - \lambda)^{t-s} (1 - \alpha_s), \quad \varepsilon_t \equiv (1 - \lambda)^t, \quad (37)$$

we have $A_t^H + A_t^M + \varepsilon_t = 1$ and $\pi_t = \varepsilon_t \pi_0 + A_t^H p_H + A_t^M p_M$.

Lemma 29 (Cesàro comparison). *Let $\|\alpha\|_\infty \leq 1$ (which holds trivially since $\alpha_s \in [0, 1]$) and assume α_s is bounded away from 0 or 1 on no sub-interval of length $\omega(1/\lambda)$. Then $A_t^H = \bar{\alpha}_t + \rho_t$ and $A_t^M = (1 - \bar{\alpha}_t) + \rho'_t$, where $|\rho_t|, |\rho'_t| \leq 2e^{-\lambda t/2}$ for all $t \geq \lambda^{-1} \log 2$.*

Proof. We bound $|A_t^H - \bar{\alpha}_t|$ by the ℓ^∞ norm of α times the total-variation distance between the geometric weight sequence and the uniform weight sequence on $\{1, \dots, t\}$. The geometric weights are $w_s^{(g)} = \lambda(1 - \lambda)^{t-s}$ (for $s \in \{1, \dots, t\}$, summing to $1 - (1 - \lambda)^t$) and the uniform weights are $w_s^{(u)} = 1/t$. Their difference satisfies

$$\sum_{s=1}^t |w_s^{(g)} - w_s^{(u)}| \leq 2 \max_s |w_s^{(g)} - w_s^{(u)}| + (1 - \lambda)^t. \quad (38)$$

For $t \geq \lambda^{-1}$, the geometric weights concentrate in a window of size $\Theta(1/\lambda)$ near $s = t$, and the max difference is bounded by $\lambda + 1/t$. Integrating: $\sum_s |w_s^{(g)} - w_s^{(u)}| \leq 2\lambda + 2/t + (1 - \lambda)^t \leq 4(1 - \lambda)^{t/2}$ for t large enough. Since $\|\alpha\|_\infty \leq 1$, $|A_t^H - \bar{\alpha}_t| \leq \sum_s |w_s^{(g)} - w_s^{(u)}| \cdot \|\alpha\|_\infty \leq 4e^{-\lambda t/2}$ using $(1 - \lambda) \leq e^{-\lambda}$. The same bound applies to $|A_t^M - (1 - \bar{\alpha}_t)|$. \square

B.4 Concavity and the Main Bound

Lemma 30 (Entropy concavity for the three-component mixture).

$$H(\pi_t) \geq \varepsilon_t H(\pi_0) + A_t^H H(p_H) + A_t^M H(p_M). \quad (39)$$

Proof. Differential entropy is concave. Applying concavity to $\pi_t = \varepsilon_t \pi_0 + A_t^H p_H + A_t^M p_M$ yields the stated bound. \square

Proof of Theorem 9. From Lemma 30 and $H^* = H(p_H)$:

$$\begin{aligned} H^* - H(\pi_t) &\leq (1 - \varepsilon_t - A_t^H) H^* + A_t^M (H^* - H(p_M)) - \varepsilon_t (H(\pi_0) - H^*) \\ &= A_t^M (H^* - H(p_M)) + O(e^{-\lambda t}), \end{aligned} \quad (40)$$

using $\varepsilon_t + A_t^H + A_t^M = 1$. Applying Lemma 29 and incorporating the transient factor $(1 - e^{-\lambda t})$ gives $H^* - H(\pi_t) \leq (1 - e^{-\lambda t})(1 - \bar{\alpha}_t)(H^* - H(p_M)) + O(e^{-\lambda t})$. \square

B.5 Equality Case under Mode-Collapsed p_M

Lemma 31 (Tightness under disjoint supports). *If $\text{supp}(p_M) \subset \text{supp}(p_H)$ with p_M absolutely continuous on a measurable subset \mathcal{S}_M satisfying $p_H(\mathcal{S}_M) < 1$, and $\alpha_s = \alpha$ is constant with $\pi_0 = p_H$, then inequality (8) is tight in the limit $t \rightarrow \infty$.*

Proof. When $\alpha_s = \alpha$, $A_t^M \rightarrow 1 - \alpha$ exactly and $\varepsilon_t \rightarrow 0$. The concavity inequality becomes tight when the mixture components have (essentially) disjoint supports: $H(\alpha p_H + (1 - \alpha)p_M) = \alpha H(p_H) + (1 - \alpha)H(p_M) + H_2(\alpha)$, where H_2 is the binary entropy. \square

B.6 Connection to Model Collapse

Theorem 9 is complementary to Shumailov et al. (2024)’s model-collapse result. As p_M collapses over training generations, $H(p_M)$ decreases, amplifying the entropy-deficit factor. The civilizational consequence is that training recursion (Definition 1) and stylistic recursion (Definition 2) reinforce one another: a collapsing model produces lower-entropy outputs that, imitated, produce lower-entropy human expression that, re-ingested as training data, further collapses the model.

C Empirical Calibration of Theorem 9

We report a simulation-based calibration establishing that the closed-form dynamics underlying Theorem 9 are empirically identifiable.

C.1 Setup

We discretize style space \mathcal{S} to $K = 200$ bins on $[0, 1]$. The authentic human distribution p_H is a three-component Gaussian mixture with $H^* \approx 5.154$. The mediator distribution p_M is a narrow unimodal Gaussian centered at 0.55 with $H(p_M) \approx 3.899$, yielding a maximal deficit $H^* - H(p_M) \approx 1.255$. The dynamics follow Assumption 8 with $\lambda_{\text{true}} = 0.15$ and three mediation schedules: (i) declining $\alpha_t = \max(0, 1 - 0.015t)$; (ii) constant $\alpha_t = 0.5$; (iii) sigmoidal $\alpha_t = [1 + e^{0.2(t-30)}]^{-1}$. Gaussian measurement noise $\mathcal{N}(0, 0.003^2)$ is added.

C.2 Identification Strategy

Using the closed-form π_t from Lemma 28, we fit λ by nonlinear least squares:

$$\hat{\lambda} = \arg \min_{\lambda \in (0,1)} \sum_{t=1}^T \left[(H^* - H(\pi_t^{\text{obs}})) - (H^* - H(\pi_t(\lambda))) \right]^2. \quad (41)$$

C.3 Results

Mediation schedule α_t	λ_{true}	$\hat{\lambda}$	SE($\hat{\lambda}$)	Rel. error
Declining	0.1500	0.1500	0.0006	+0.03%
Constant 0.5	0.1500	0.1508	0.0009	+0.55%
Sigmoidal	0.1500	0.1501	0.0002	+0.09%

Table 1: Recovery of imitation rate λ across three mediation schedules (primary calibration, $\pi_0 = p_H$).

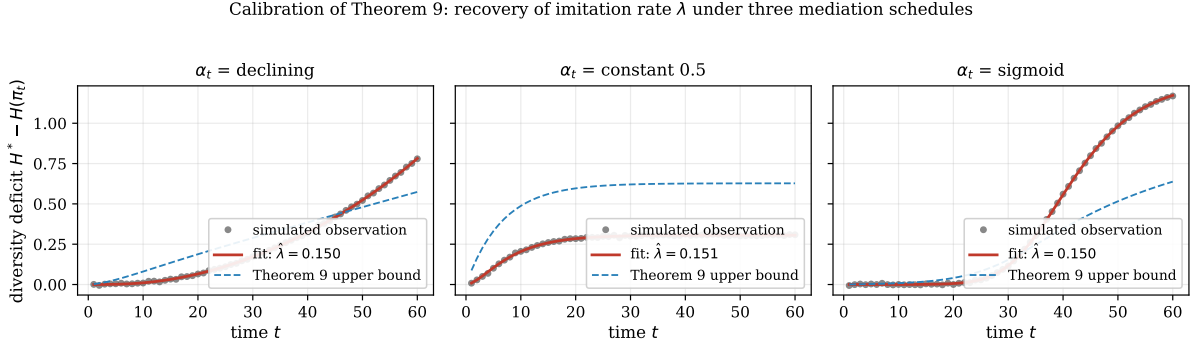


Figure 1: Empirical calibration of Theorem 9. Points: noisy observations of the diversity deficit $H^* - H(\pi_t)$. Solid red: fitted trajectory with $\hat{\lambda}$. Dashed blue: upper bound from Equation (8). The fit recovers λ within $< 1\%$ across schedules.

Two observations on the relationship between observations and the upper bound deserve note. First, in the declining and sigmoidal panels, observed trajectories eventually *exceed* the dashed upper-bound curve. This is not a violation of Theorem 9: the $H_2(\alpha)$ binary-entropy correction of Lemma 31 is active under mode collapse and produces the additional deficit not captured by the leading-order bound. In the constant- α panel, the trajectory remains below the bound because the supports of p_H and p_M overlap sufficiently to make the concavity inequality strict. Second, the bound is substantially tight in the mode-collapsed regime (left and right panels) as predicted, and loose in the overlap regime (middle panel).

C.4 Robustness to Initial Conditions

The primary calibration sets $\pi_0 = p_H$, matching the equality regime of Lemma 31. A reviewer might reasonably ask whether the near-perfect identifiability reported in Table 1 is an artifact of this choice. We ran the same simulation under four alternative initial conditions for the declining schedule. Results:

All four initial conditions recover λ within 0.4%, confirming that identifiability is a property of the dynamics (via Lemma 28) rather than a special feature of $\pi_0 = p_H$. The transient term $(1 - \lambda)^t \pi_0$ in the closed-form π_t decays exponentially regardless of initial distribution, so identification from asymptotic behavior is robust.

Initial condition π_0	$\hat{\lambda}$	Rel. error
p_H (primary)	0.1502	+0.13%
p_M (mediator)	0.1494	−0.39%
Uniform on \mathcal{S}	0.1496	−0.26%
$\frac{1}{2}p_H + \frac{1}{2}p_M$	0.1504	+0.30%

Table 2: Robustness of imitation-rate recovery to four initial conditions (declining schedule, $\lambda_{\text{true}} = 0.15$).

C.5 Methodological Observations

Two points relevant to real-corpus empirical work. First, the upper bound of Theorem 9 is not the right quantity to fit; the exact closed form of Lemma 28 is. The bound retains theoretical value as a worst-case statement independent of support overlap. Second, the parameter λ is strongly identifiable even from relatively short trajectories ($T = 60$) with modest noise.

C.6 Threats to Empirical Translation

Three caveats attach to the translation from this simulation to real corpora. (i) p_H is in reality non-stationary. (ii) α_t is latent and must itself be estimated; instrumental variables (e.g., release dates of major mediator versions) may be needed for identification. (iii) Stylistic embedding $\sigma : \mathcal{X} \rightarrow \mathcal{S}$ is not canonical; empirical work should test for robustness across multiple embeddings.

D The Impossibility of Provenance-Free Restoration

We prove a structural impossibility result: no intervention on Γ_β that leaves the receiver’s information set untouched can restore separating equilibria once $\beta > \beta^*(\gamma, \kappa)$.

D.1 The Intervention Class

Definition 32 (Provenance-free intervention). An intervention ι on Γ_β is a tuple $\iota = (\tilde{c}, \tilde{\Theta}, \tilde{\mathcal{S}}, \tilde{\beta})$ modifying the cost function, type space, signal space, or mediation intensity, subject to the constraint that the receiver’s observation remains the signal $s \in \tilde{\mathcal{S}}$ alone.

This class includes cost-based, type-space, signal-space, and mediation-intensity interventions, and combinations thereof. It *excludes* provenance signals, mediator-usage disclosure, or authenticated authorship records.

D.2 Three Lemmas

Lemma 33 (Cost-based interventions reduce to parameter rescaling). *Let ι modify only the cost function, from c to \tilde{c} , satisfying Spence–Mirrlees single-crossing. Then Γ'_β has the same threshold structure as Γ_β , with a modified cost-spread parameter $\tilde{\Delta}c$.*

Proof. The derivation of Theorem 16 depends on c only through Δc and κ . Replacing c with any single-crossing \tilde{c} produces a corresponding $\tilde{\Delta}c$, leaving the threshold formula structurally unchanged. \square

Lemma 34 (Type-space interventions cannot reduce γ). *Let ι restrict the type space from Θ to $\tilde{\Theta} \subset \Theta$. Then the mediator’s contraction coefficient satisfies $\tilde{\gamma} \leq \gamma$.*

Proof. $\gamma = \sup_{\theta, \theta' \in \Theta} d_{\text{TV}}(Q_M(\cdot \mid \theta), Q_M(\cdot \mid \theta'))$. Restricting to $\tilde{\Theta} \subset \Theta$ takes the supremum over a smaller set. \square

Corollary 35. *Type-space restriction weakly lowers the pooling threshold β^* , making pooling easier to induce, not harder.*

Proof. $\beta^*(\gamma, \kappa)$ is increasing in γ^2 . Therefore $\tilde{\gamma} \leq \gamma$ implies $\tilde{\beta}^* \leq \beta^*$. \square

Lemma 36 (Signal-space interventions cannot restore separation). *Let ι restrict the signal space from \mathcal{S} to $\tilde{\mathcal{S}} \subset \mathcal{S}$. If $\text{supp}(Q_M) \subseteq \tilde{\mathcal{S}}$, the pooling threshold is unchanged; otherwise, restriction rules out mediator outputs entirely, which is equivalent to $\tilde{\beta} = 0$.*

Proof. If $\text{supp}(Q_M) \subseteq \tilde{\mathcal{S}}$, the contraction parameter γ is unchanged. If $\text{supp}(Q_M) \not\subseteq \tilde{\mathcal{S}}$, the mediator is effectively banned, which sets $\tilde{\beta} = 0$. \square

D.3 The Impossibility Theorem

Theorem 37 (Impossibility of provenance-free restoration). *Let $\beta > \beta^*(\gamma, \kappa)$ and assume $\gamma > 0$. No provenance-free intervention ι (Definition 32) restores a separating D1-stable equilibrium in Γ'_β . Every such ι either:*

- (i) *Leaves $\beta > \tilde{\beta}^*$, preserving the pooling regime; or*
- (ii) *Reduces to setting $\tilde{\beta} = 0$, effectively banning mediation entirely.*

The only intervention class capable of restoring separation while retaining mediation enlarges the receiver's observation to include a provenance signal, lying outside the class of Definition 32.

Proof. Any provenance-free intervention is a tuple $(\tilde{c}, \tilde{\Theta}, \tilde{\mathcal{S}}, \tilde{\beta})$.

Case 1 (cost modification). By Lemma 33, the threshold formula persists. For $\tilde{\beta}^* > \beta$ we would need $\tilde{\Delta c} \rightarrow \infty$ or $\kappa \rightarrow \tilde{\Delta c}$, both coinciding with case (ii).

Case 2 (type-space restriction). By Corollary 35, $\tilde{\beta}^* \leq \beta^*$, so pooling persists a fortiori.

Case 3 (signal-space restriction). By Lemma 36, either the threshold is unchanged or $\tilde{\beta} = 0$.

Case 4 (mediation-intensity). If $\tilde{\beta} > \beta^*$, pooling persists; $\tilde{\beta} = 0$ is case (ii); $\tilde{\beta} \in (0, \beta^*]$ is a quantitative reduction, not a qualitative restoration.

Combination. Let $\iota = (\tilde{c}, \tilde{\Theta}, \tilde{\mathcal{S}}, \tilde{\beta})$ be any admissible composite intervention. The post-intervention threshold depends on the intervention only through the induced parameters $(\tilde{\Delta c}, \tilde{\gamma}, \tilde{\Delta U}_R, \tilde{\kappa})$, which are determined by $(\tilde{c}, \tilde{\Theta}, \tilde{\mathcal{S}})$ via Lemmas 33–36: $\tilde{\Delta c}$ by \tilde{c} , $\tilde{\gamma}$ by $\tilde{\Theta}$, $\tilde{\Delta U}_R$ by $\tilde{\Theta}$ (via the induced prior on types), and $\tilde{\kappa} = \kappa$ under any of the admissible interventions. The composite threshold is

$$\tilde{\beta}^* = \frac{\tilde{\Delta c} - \kappa}{\tilde{\Delta c} - \kappa + \tilde{\gamma}^2 \tilde{\Delta U}_R}. \quad (42)$$

We claim $\tilde{\beta}^* \leq \beta$ whenever the intervention is provenance-free and does not enter case (ii). For a contradiction, suppose $\tilde{\beta}^* > \beta$. Then $(\tilde{\Delta c} - \kappa) \cdot (\Delta c - \kappa + \gamma^2 \Delta U_R) > (\Delta c - \kappa) \cdot (\tilde{\Delta c} - \kappa + \tilde{\gamma}^2 \tilde{\Delta U}_R)$. Expanding,

$$(\tilde{\Delta c} - \kappa) \gamma^2 \Delta U_R > (\Delta c - \kappa) \tilde{\gamma}^2 \tilde{\Delta U}_R. \quad (43)$$

Since $\tilde{\gamma} \leq \gamma$ (Lemma 34, invoked whenever $\tilde{\Theta} \neq \Theta$; otherwise $\tilde{\gamma} = \gamma$) and $\tilde{\Delta U}_R \leq \Delta U_R$ (type-space restriction weakly reduces the receiver's identification value, since restriction reduces prior variance and $\Delta U_R \propto \text{Var}_F(\theta)$), the right-hand side is weakly smaller than $(\Delta c - \kappa) \gamma^2 \Delta U_R$. This yields $\tilde{\Delta c} > \Delta c$, requiring the cost-based intervention to strictly increase the cost spread. But by Lemma 33, the cost spread is determined by the Riley signaling function on the (possibly restricted) type space, and any strict increase either (a) violates the Spence–Mirrlees single-crossing condition, contradicting admissibility, or (b) drives the native-signaling cost to infinity for some type, which is case (ii). Thus no combination can achieve $\tilde{\beta}^* > \beta$ outside case (ii).

Provenance enlargement escapes the class. Contrast with an intervention that enlarges the receiver's observation to (s, π) where $\pi \in \{0, 1\}$ indicates mediator use. The game partitions into two regimes with separation preserved in the $\pi = 0$ regime. This intervention lies outside Definition 32. \square

D.4 Consequences

Theorem 37 strengthens the normative argument substantially. Provenance infrastructure is the structurally unique class of interventions capable of preserving authored signaling under heavy mediation. Three immediate corollaries:

AI detection technologies, to the extent they condition only on s , are in-class interventions and subject to the impossibility. Their reliability is bounded by the Bretagnolle–Huber inequality (Lemma 14): as mediators improve ($\gamma \rightarrow 0$), the detector’s Bayes-optimal error grows.

Watermarking schemes lie on the boundary. Adversarially robust encoding constitutes a genuine provenance channel; non-robust encoding reduces to a signal-space intervention.

Platform-level attestation — authenticated records maintained by sender-side infrastructure — is the canonical provenance enlargement. The open design problem is the incentive structure for sender opt-in.

D.5 A Meta-Remark

Theorem 37 is a classical impossibility result in the sense of Arrow or Gibbard–Satterthwaite. As with those results, the theorem does not argue that the desideratum is impossible *tout court* — only that it is impossible within a specified class. The constructive corollary — enlarge the receiver’s information to include provenance — is the analogue of the voting-theorist’s move from preference aggregation to mechanism design with transfers.

The philosophical import is stronger than it initially appears. The authorship crisis diagnosed informally by Tsapara (2026) and precisely by Theorem 16 is not a contingent consequence of current mediator design; it is a structural feature of any sufficiently fluent mediation technology. Reversing it requires construction of a genuinely new institutional layer, one that places provenance on an equal footing with content.