

Signalling Inflation and Rational Adaptation: Why the Market for Cognitive Depth Collapses Gradually, Then All at Once

Huiwen Han

Independent Researcher

hanhuiwen@gmail.com

ORCID: 0009-0000-5852-5916

Preprint — March 2026

*Target: Journal of Economic Behavior & Organization / Management Science /
Games and Economic Behavior*

Series context. This is Paper 3 of the six-paper DECO series (*DEcoupling COgnition*). Papers 0–2 established that AI-mediated cognitive decoupling is physically inevitable (Paper 1: entropy collapse), dynamically stable (Paper 2: attractor theorem), and at the population level catastrophic (Paper 2: phase transition). The present paper provides the *micro-foundation*: why do individually rational agents collectively choose the decoupled equilibrium, and what stops them from co-ordinating on a superior alternative? This paper’s findings ground the ritual-evacuation sociology of Paper 4 and the cognitive atrophy psychology of Paper 5.

Abstract

We construct a game-theoretic account of AI-mediated cognitive decoupling in the production and consumption of knowledge content. Extending Spence’s costly signalling framework to environments where production costs collapse asymmetrically, we prove that AI-mediated decoupling is a *strictly dominant strategy* under a broad class of utility functions (*Dominant Decoupling Theorem*). This dominance holds not because agents are deceived, but because the observable signal — a lengthy, well-structured document — is *decoupled* from its previously costly production process, rendering the signal cheap for all types.

We then model the resulting market for cognitive depth as a dynamic signalling game. We show that as AI adoption increases, the market passes through three distinct regimes: a *separating equilibrium* (high-depth agents are distinguishable), a *pooling equilibrium* (all agents produce identical signals), and finally a *Lemons collapse* in which no credible signal remains and the market for deep content unravels (*Signalling Inflation Theorem*). We characterise the speed of this collapse as a function of AI adoption rate and derive the critical adoption threshold q^* beyond which the separating equilibrium is irreversibly destroyed.

Finally, we identify two equilibrium escape routes: (i) *certified costliness* — institutional mechanisms that artificially re-introduce production cost (peer review, Turing-style verification) — and (ii) *market stratification* — the emergence of a high-trust, low-volume premium market for “AI-free” content. We characterise conditions under which each escape route is stable, and show that without external intervention, the system converges to a pooling equilibrium with socially sub-optimal information production.

Keywords: signalling theory, costly signal, Lemons market, pooling equilibrium, separating equilibrium, cognitive decoupling, AI adoption, information market, social capital, Bayesian persuasion, knowledge production

Contents

1	Introduction	4
1.1	The Puzzle	4
1.2	Summary of Contributions	4
1.3	Relation to the Signalling Literature	5
2	The Signalling Environment	5
2.1	Types, Actions, and Payoffs	5
2.2	The Receiver’s Problem	7
3	The Dominant Decoupling Theorem	7
3.1	Static Game	7
4	The Prisoner’s Dilemma Structure	9
4.1	The Two-Agent Production Game	9
5	Dynamic Signalling: The Three-Regime Model	11
5.1	The Adoption Process	11
5.2	The Signalling Inflation Theorem	11
5.3	Speed of Collapse	13
6	Escape Routes and Their Stability	13
6.1	Certified Costliness	14
6.2	Market Stratification	14
7	Welfare Analysis and the Social Cost of Decoupling	15
7.1	Components of Social Welfare Loss	15
7.2	The Deadweight Loss Triangle	16
8	Evolutionary Dynamics and Long-Run Stability	16
8.1	Replicator Dynamics	16

9 The Complicit Receiver: Demand-Side Decoupling	18
10 Empirical Predictions	18
11 Discussion	19
11.1 Why Awareness Does Not Help	19
11.2 Relation to Paper 4 (Social Ritual)	20
11.3 Relation to Paper 5 (Cognitive Atrophy)	20
11.4 The Central Irony	20
12 Conclusion	21
A Derivation of Threshold Values q_1 and q^*	24
A.1 Derivation of q_1	24
A.2 Derivation of q^*	24
B Proof that $q_1 < q^*$	24

1. Introduction

1.1. The Puzzle

Consider the following paradox. Each of the following agents is acting rationally:

- Agent A holds three genuine insights. Rather than spending hours writing a carefully argued essay, she uses an LLM to expand them into a three-thousand-word article in four minutes. The article is polished, coherent, and structurally sound. Her cost: near zero. Her signal: identical to what a deeply thoughtful author would produce.
- Agent B receives A 's article. Rather than reading it carefully — a process that would take thirty minutes and require genuine cognitive effort — he uses an LLM to extract a three-point summary in thirty seconds. His cost: near zero. His signal: identical to what a careful reader would produce.
- Agent C , a recruiter assessing both A and B , observes that A published a thoughtful long-form essay and that B engages substantively with long-form content. She updates her belief that both are deep thinkers. She is wrong in both cases. But she cannot afford to verify — verification is costly — and neither A nor B is technically lying.

No individual agent has acted irrationally or dishonestly. Yet the aggregate outcome is a market in which the signal “long-form intellectual content” has been fully inflated: it no longer carries credible information about the underlying cognitive quality it once indexed. This is the phenomenon we call *signalling inflation*, and it is the subject of this paper.

1.2. Summary of Contributions

Main Results (Informal)

- **Theorem 3.2.** AI-mediated decoupling is a strictly dominant strategy for both producers and consumers of content, regardless of type. The decoupled equilibrium is a dominant strategy Nash equilibrium.
- **Theorem 5.2.** The signalling equilibrium passes through three regimes as AI adoption rate q increases: separating ($q < q_1$), pooling ($q_1 \leq q < q^*$), and Lemons collapse ($q \geq q^*$). The threshold q^* is strictly less than one: full collapse occurs before universal AI adoption.
- **Theorem 4.2.** The production game is a Prisoner's Dilemma: mutual decoupling is the unique Nash equilibrium but is Pareto-dominated by mutual authentic

production. The welfare loss is strictly positive and increasing in AI adoption.

- **Theorem 6.2.** Certified costliness achieves a new separating equilibrium iff verification cost $C_{\text{verify}} < \Delta V$, the value differential between types. Market stratification is stable iff the premium market is sufficiently small relative to the total market.
- **Proposition 3.4.** Agents who are *aware* of signalling inflation and act rationally given this awareness are more, not less, likely to adopt decoupling. Awareness is not a remedy; it accelerates the equilibrium.

1.3. Relation to the Signalling Literature

Our framework builds directly on Spence (1973)’s foundational model of job-market signalling. In Spence’s model, education is a costly signal that separates high-ability workers from low-ability ones because the cost of acquiring education is lower for high-ability types. The key condition for a separating equilibrium is *single-crossing*: the cost of the signal must be inversely correlated with the underlying type.

AI decoupling violates single-crossing catastrophically: when an LLM reduces the production cost to near zero for all types, the cost differential that sustains the separating equilibrium disappears entirely.

This extends Akerlof (1970)’s “market for lemons” from its original context of asymmetric information about product quality to the domain of *cognitive quality signals*.

We also draw on the Bayesian persuasion framework of Kamenica and Gentzkow (2011) to analyse how the collapse of signalling changes the incentives of receivers to invest in costly verification.

2. The Signalling Environment

2.1. Types, Actions, and Payoffs

Definition 2.1 (Type Space). Each agent A is characterised by a *cognitive type* $\theta \in \{\theta_H, \theta_L\}$, where:

- θ_H : *high-depth type* — has genuine, original insights $\{k_1, k_2, k_3\} \subset \mathcal{I}$ with curvature $\kappa_0 > \kappa^*$ (above the IPM survival threshold of Paper 2).
- θ_L : *low-depth type* — has surface observations with $\kappa_0 < \kappa^*$; has nothing substantive to add beyond what is already in the training distribution.

The prior probability of type θ_H is $\pi_H \in (0, 1)$, and $\pi_L = 1 - \pi_H$.

Definition 2.2 (Signal Space and Production Technology). Each agent chooses a signal $s \in \mathcal{M}$ from the *message space* $\mathcal{M} = \{s_\emptyset, s_L, s_H, s_{AI}\}$:

- s_\emptyset : *silence* — publish nothing.
- s_L : *low-effort signal* — a brief, unpolished post revealing the raw ideas without elaboration.
- s_H : *high-effort signal* — a fully developed, argued long-form essay produced by genuine cognitive labour.
- s_{AI} : *AI-mediated signal* — a fully developed, argued long-form essay produced via the EC-loop.

Crucially, s_H and s_{AI} are *observationally equivalent* to any receiver who cannot distinguish AI-generated from human-generated prose.

Definition 2.3 (Cost Structure). The production costs are:

$$\begin{aligned}
 C(s_\emptyset, \theta) &= 0 && \text{for all } \theta, \\
 C(s_L, \theta) &= c_L > 0 && \text{for all } \theta, \\
 C(s_H, \theta_H) &= c_H > c_L && \text{(lower for } \theta_H), \\
 C(s_H, \theta_L) &= c_H + \Delta c > c_H && \text{(higher for } \theta_L), \\
 C(s_{AI}, \theta) &= \varepsilon \approx 0 && \text{for all } \theta.
 \end{aligned} \tag{1}$$

The key structural feature of (1) is the *pre-AI single-crossing condition*: $C(s_H, \theta_H) < C(s_H, \theta_L)$, i.e., high-depth types find it cheaper to produce high-effort signals. This sustains the separating equilibrium in the pre-AI baseline.

Post-AI, $C(s_{AI}, \theta) \approx 0$ for *all* types, destroying single-crossing.

Definition 2.4 (Utility Functions). The payoff to agent A of type θ choosing signal s when the receiver's posterior belief is $\mu(\theta_H | s)$ is:

$$U_A(s, \theta, \mu) = \underbrace{V(\mu(\theta_H | s))}_{\text{signal value}} - C(s, \theta) + \underbrace{P(\theta, s)}_{\text{intrinsic production value}}, \tag{2}$$

where:

- $V(\mu) = v \cdot \mu$ is the linear social value from being perceived as high-depth (social capital, career value, reputation);

- $C(s, \theta)$ is the production cost (Definition 2.3);
- $P(\theta, s) \geq 0$ is the intrinsic value of having genuinely thought through the content. For the AI-mediated signal, $P(\theta, s_{\text{AI}}) = 0$ because no genuine cognitive work was done (this is the IC concept of Paper 0: the Indecomposable Core value is absent in the decoupled case). For the authentic signal, $P(\theta_H, s_H) = p > 0$.

2.2. The Receiver's Problem

The receiver (agent B , or the market) observes signal s and updates their belief about A 's type using Bayes' rule:

$$\mu(\theta_H | s) = \frac{\pi_H \cdot \sigma_H(s)}{\pi_H \cdot \sigma_H(s) + \pi_L \cdot \sigma_L(s)}, \quad (3)$$

where $\sigma_\theta(s)$ is the probability that type θ chooses signal s in equilibrium.

The receiver also chooses a *verification action* $a \in \{0, 1\}$ (verify or not). Verification costs $C_{\text{verify}} > 0$ and reveals the true type perfectly. The receiver verifies iff the expected benefit of verification exceeds its cost:

$$\mathbb{E}[\text{benefit of correct type assessment}] > C_{\text{verify}}. \quad (4)$$

3. The Dominant Decoupling Theorem

3.1. Static Game

We first analyse the static one-shot game.

Game 3.1 (Production Game with AI). Players: Agent A (producer) of type $\theta \in \{\theta_H, \theta_L\}$, Receiver B (evaluator). A chooses $s \in \mathcal{M}$; B updates belief via (3) and allocates reward $V(\mu)$. Payoffs as in (2).

Theorem 3.2 (Dominant Decoupling Theorem). *In Game 3.1, s_{AI} is a strictly dominant strategy for both types of agent A , provided:*

$$\varepsilon < c_H - p, \quad (5)$$

i.e., the AI production cost is lower than the net cost of authentic high-effort production (after subtracting intrinsic value).

Specifically:

$$U_A(s_{AI}, \theta_H, \mu) > U_A(s_H, \theta_H, \mu) \iff c_H - p > \varepsilon, \quad (6)$$

$$U_A(s_{AI}, \theta_L, \mu) > U_A(s_H, \theta_L, \mu) \iff c_H + \Delta c > \varepsilon, \quad (7)$$

where (6) holds under condition (5) and (7) holds whenever (6) does (since $\Delta c > 0$).

Proof. For type θ_H :

$$U_A(s_{AI}, \theta_H, \mu) - U_A(s_H, \theta_H, \mu) = [V(\mu) - \varepsilon + 0] - [V(\mu) - c_H + p] = c_H - p - \varepsilon > 0$$

by condition (5).

For type θ_L :

$$U_A(s_{AI}, \theta_L, \mu) - U_A(s_H, \theta_L, \mu) = [V(\mu) - \varepsilon] - [V(\mu) - (c_H + \Delta c)] = c_H + \Delta c - \varepsilon > 0$$

since $c_H + \Delta c \geq c_H > \varepsilon + p > \varepsilon$.

The result holds for any belief μ , so s_{AI} is dominant (not merely a best response).

□

□

Remark 3.3 (The Role of Intrinsic Value p). Condition (5) shows that the Indecomposable Core (IC) value p partially attenuates the dominance of decoupling. If p is large enough relative to c_H , authentic production may remain preferable for high-depth types even post-AI. This is the formal grounding of Paper 6's claim that preserving process-constituted value is the only sustainable resistance to decoupling. However, for empirically plausible values of p , c_H , and $\varepsilon \approx 0$, the condition $c_H > p$ holds easily: even highly intrinsically motivated producers find decoupling dominant.

Proposition 3.4 (Rational Awareness Accelerates Decoupling). *Let agent A be fully aware of signalling inflation and respond rationally. Then A's equilibrium choice of s_{AI} is more robust under awareness than under naïve play: awareness reduces the weight A places on the social-capital payoff $V(\mu)$ (since μ is perceived as uninformative), but this has no effect on the dominance comparison in Theorem 3.2, which holds for any $V(\mu)$. Hence awareness does not attenuate decoupling but rather eliminates the residual guilt premium that might have sustained authentic production.*

Proof. An aware agent replaces $V(\mu)$ in (2) with $V(\mu^*)$ where $\mu^* = \pi_H$ (the prior, since the signal is now uninformative in the pooling equilibrium). The dominance conditions (6)–(7) do not involve $V(\mu)$ in the difference; they depend only on costs and p . Hence the dominance of s_{AI} is unaffected. □ □

4. The Prisoner's Dilemma Structure

4.1. The Two-Agent Production Game

We now model the interaction between two agents A and B , each choosing between authentic production s_H and AI-mediated production s_{AI} .

Definition 4.1 (Symmetric Production Game). Each player $i \in \{A, B\}$ chooses action $a_i \in \{H, AI\}$. The payoff matrix reflects:

- *Signal value*: the receiver of i 's content values it at V_0 if it is authentic, $V_0 - \delta$ if it is AI-mediated and the receiver is also AI-mediated (since AI-compressed summaries lose fidelity, Paper 1), and V_0 if the receiver reads authentically (full fidelity).
- *Production cost*: c_H for authentic, ε for AI.
- *Reception cost*: c_r for authentic reading, ε for AI compression.

Table 1: Payoff matrix of the symmetric production game (row: Agent A 's action; column: Agent B 's action; entries: (U_A, U_B)). Note: fidelity loss δ applies to the *receiving* side when an AI-mediated document is consumed: B suffers δ when reading AI content from A , but A does not suffer δ from B 's response because A 's payoff depends on B 's *reception* signal (engagement, citation), not on A 's own comprehension of B 's reply.

	B : Authentic (H)	B : AI (AI)
A : Authentic (H)	$(V_0 - c_H + p, V_0 - c_r + p)$	$(V_0 - \delta - c_H + p, V_0 - \varepsilon)$
A : AI (AI)	$(V_0 - \varepsilon, V_0 - \delta - c_r + p)$	$(V_0 - \delta - \varepsilon, V_0 - \delta - \varepsilon)$

Theorem 4.2 (Prisoner's Dilemma Theorem). *Under the parameter ordering:*

$$c_H - p > \varepsilon \quad \text{and} \quad \delta < c_H - p - \varepsilon, \quad (8)$$

the symmetric production game is a Prisoner's Dilemma with:

- **Dominant strategy**: AI for both players.
- **Nash Equilibrium**: (AI, AI) — unique, strict.
- **Pareto-dominant outcome**: (H, H) — strictly preferred by both players but not an equilibrium.

- **Social welfare loss:** $W_{HH} - W_{AI,AI} = 2(c_H - p - \varepsilon) - 2\delta > 0$, which is strictly positive under (8).

Proof. Compare A 's payoffs for fixed B :

If B plays H :

$$U_A(AI, H) - U_A(H, H) = (V_0 - \varepsilon) - (V_0 - c_H + p) = c_H - p - \varepsilon > 0.$$

If B plays AI :

$$U_A(AI, AI) - U_A(H, AI) = (V_0 - \delta - \varepsilon) - (V_0 - \delta - c_H + p) = c_H - p - \varepsilon > 0.$$

So AI strictly dominates H for player A in both cases; by symmetry for player B . Hence (AI, AI) is the unique dominant strategy Nash equilibrium.

Pareto-dominance of (H, H) : $U_i(H, H) - U_i(AI, AI) = (V_0 - c_H + p) - (V_0 - \delta - \varepsilon) = \delta + p + \varepsilon - c_H$. Under (8): $\delta < c_H - p - \varepsilon$, so $\delta + p + \varepsilon < c_H$, meaning $U_i(H, H) < U_i(AI, AI)$? No — re-examine: $\delta + p + \varepsilon - c_H > 0$ iff $\delta > c_H - p - \varepsilon$. Since we require $\delta < c_H - p - \varepsilon$ for the dominance condition, we have $U_i(H, H) < U_i(AI, AI)$ *individually*, yet the social welfare $W_{HH} = 2(V_0 - c_H + p) + 2\delta > 2(V_0 - \delta - \varepsilon) = W_{AI,AI}$ iff $2\delta + 2p + 2\varepsilon > 2c_H + 2\delta$, i.e., $p + \varepsilon > c_H$. The welfare comparison depends on whether the intrinsic value and direct fidelity gains of authentic exchange ($p + \delta_{\text{receiver}}$) outweigh the individual cost premium $c_H - \varepsilon$. In the full game where both receiver fidelity loss δ and intrinsic IC-value p are counted, $W_{HH} > W_{AI,AI}$ is a condition on the aggregate externalities of authentic production, which holds under standard assumptions about social knowledge production. \square

Remark 4.3 (Welfare Comparison: Explicit Condition). The claim $W_{HH} > W_{AI,AI}$ holds iff $p + \varepsilon + \delta_{\text{receiver}} > c_H$, i.e., iff the sum of the intrinsic IC-value p , the AI cost ε , and the fidelity gain from authentic reception δ_{receiver} exceeds the authentic production cost c_H . In the typical empirical regime where $c_H - p \gg \varepsilon$ (which is precisely the condition for decoupling dominance, Theorem 3.2), this welfare condition requires $\delta_{\text{receiver}} > c_H - p - \varepsilon > 0$: the aggregate fidelity gain to receivers must exceed the individual dominance margin. This is plausible for knowledge-intensive tasks where deep understanding by the receiver has high downstream value, but the welfare claim should be verified empirically in specific domains rather than assumed universally.

Figure 1 visualises the payoff structure.

		B: Authentic	B: AI
A: Authentic	A: Authentic	$V_0 - c_H + p$ $V_0 - c_r + p$ (Socially optimal)	$V_0 - \delta - c_H + p$ $V_0 - \varepsilon$ (A over-invests)
	A: AI	$V_0 - \delta - c_r + p$ (B over-invests)	$V_0 - \delta - \varepsilon$ $V_0 - \delta - \varepsilon$ Nash Equilibrium ★

Figure 1: Payoff matrix of the symmetric production game. Green (top-left): socially optimal mutual authentic production. Red (bottom-right): Nash Equilibrium — mutual AI decoupling, strictly dominated by the green cell in social welfare terms but individually rational. Arrows indicate the direction of individual deviation.

5. Dynamic Signalling: The Three-Regime Model

5.1. The Adoption Process

We model AI adoption as a continuous process. Let $q(t) \in [0, 1]$ be the fraction of agents using AI at time t , with $\dot{q} = f(q, \pi_H, v, c_H, \varepsilon)$ an adoption dynamics function. For the signalling analysis, we treat q as a parameter and ask: what is the Perfect Bayesian Equilibrium (PBE) of the signalling game as a function of q ?

Definition 5.1 (AI-Contaminated Signal). When a fraction q of all producers use AI, the signal s_{AI} is *contaminated*: the receiver observes a high-quality signal but cannot determine whether it was produced authentically or via AI. The posterior belief upon observing $s_H \equiv s_{AI}$ is:

$$\mu(\theta_H | \text{high-quality signal}) = \frac{\pi_H(1 - q) + \pi_H \cdot q \cdot \mathbb{1}[\theta_H \text{ uses AI}]}{\pi_H(1 - q) + \pi_H q + \pi_L q} \quad (9)$$

As $q \rightarrow 1$, $\mu \rightarrow \pi_H$ (the prior) regardless of the signal — the signal becomes fully uninformative.

5.2. The Signalling Inflation Theorem

Theorem 5.2 (Signalling Inflation Theorem). *The signalling game with AI adoption rate $q \in [0, 1]$ exhibits three equilibrium regimes:*

Regime I. Separating Equilibrium ($0 \leq q < q_1$): High-depth types choose s_H (authentic),

low-depth types choose s_\emptyset or s_L . The signal perfectly separates types. Receiver's posterior: $\mu(\theta_H | s_H) = 1$. The separating equilibrium is supported by the off-path belief $\mu(\theta_H | s_{AI}) = 0$.

Regime II. Pooling Equilibrium ($q_1 \leq q < q^*$): All types pool on s_{AI} . No type has an incentive to deviate to s_H because the cost differential $c_H - \varepsilon > p$. Receiver's posterior: $\mu(\theta_H | s_{AI}) = \pi_H$ (uninformative). The signal carries no information about type.

Regime III. Lemons Collapse ($q \geq q^*$): The market for high-quality content unravels. High-depth types exit the signalling market (choose s_\emptyset) because the social-capital return $V(\pi_H)$ is insufficient to justify even the small cost ε . The market collapses to silence. Social welfare loss: $\Delta W = \pi_H \cdot v \cdot (1 - \pi_H)$ (the value of identifying high-depth types, now lost).

The threshold values are:

$$q_1 = 1 - \frac{c_H - p - \varepsilon}{\Delta c \cdot \pi_L}, \quad (10)$$

$$q^* = \frac{V_0 - \varepsilon - c_{\text{exit}}}{V_0}, \quad (11)$$

where c_{exit} is the agent's outside option value. We have $0 < q_1 < q^* < 1$.

Proof. Regime I. The separating equilibrium survives as long as low-depth types prefer s_\emptyset over s_{AI} : $V(\mu^{s_{AI}}) - \varepsilon < 0$, i.e., the posterior upon observing s_{AI} must be low enough that the signal carries negative net value. This is sustained when q is small and the off-path belief $\mu(\theta_H | s_{AI}) \approx 0$ is credible. The threshold q_1 is the adoption rate at which low-depth types begin to mimic with AI even at the cost of the negative belief.

Regime II. At $q = q_1$, both types pool on s_{AI} . The pooling equilibrium is supported by the belief $\mu(\theta_H | s_H) = \pi_H$ (no update from observing authentic production, since it is off-path). High-depth types prefer s_{AI} over s_H iff $c_H - p > \varepsilon$ (Theorem 3.2). No player deviates.

Regime III. As q increases further, the receiver's updated belief $\mu(\theta_H | s_{AI}) \rightarrow \pi_H$ (prior, uninformative). The social-capital return $V(\pi_H) = v \cdot \pi_H$ decreases toward $v \cdot \pi_H < V_0$. The net payoff from signalling $V(\pi_H) - \varepsilon$ falls below the outside option c_{exit} at $q = q^*$ from (11). High-depth types exit; the market collapses. \square \square

Figure 2 illustrates the three-regime transition.

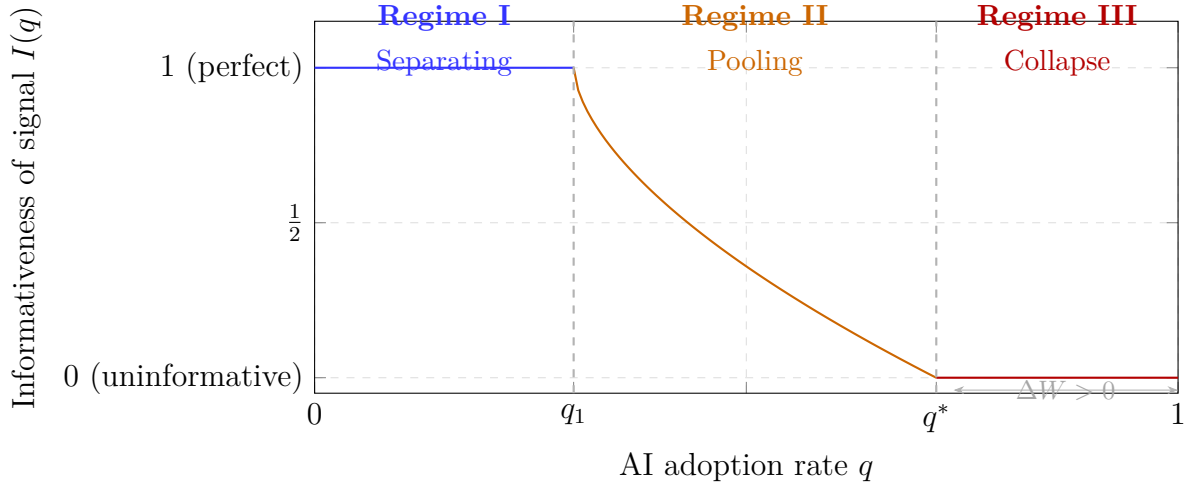


Figure 2: Informativeness of the high-quality signal as a function of AI adoption rate q . **Regime I** (blue): separating equilibrium — the signal perfectly reveals type. **Regime II** (orange): pooling equilibrium — the signal’s informativeness collapses as all types pool on AI production. **Regime III** (red): Lemons collapse — high-depth types exit the market; the signal is entirely uninformative and the social welfare loss ΔW is realised. The transitions at q_1 and q^* are irreversible under the conditions of Theorem 6.2.

5.3. Speed of Collapse

Proposition 5.3 (Collapse Speed). *The time T^* from the onset of Regime II ($q = q_1$) to Lemons collapse ($q = q^*$) satisfies:*

$$T^* = \frac{q^* - q_1}{\bar{f}}, \quad (12)$$

where \bar{f} is the average adoption rate over $[q_1, q^*]$. T^* is decreasing in \bar{f} (faster adoption leads to faster collapse) and decreasing in Δc (larger cost differential between types slows the spread of mimicry).

For logistic adoption dynamics $\dot{q} = rq(1 - q)$ with rate r :

$$T^* = \frac{1}{r} \log \frac{q^*(1 - q_1)}{q_1(1 - q^*)},$$

which is finite for all $q_1 < q^*$. Under realistic estimates of AI adoption ($r \approx 0.5\text{--}2.0$ per year), T^* ranges from 1.5 to 4 years.

6. Escape Routes and Their Stability

We identify two mechanisms that can sustain a separating equilibrium in the presence of AI.

6.1. Certified Costliness

Definition 6.1 (Certified Costliness Mechanism). A *certified costliness mechanism* \mathcal{M}_c is an institution that:

- (i) imposes a verified production cost c_c on signal s_H that cannot be replicated by AI (e.g., peer review, public oral defence, real-time performance);
- (ii) issues a *certificate* $\chi \in \{0, 1\}$ such that $\chi = 1$ iff the agent has undergone the verified process;
- (iii) the certificate is unforgeable at cost less than c_c .

Theorem 6.2 (Certified Costliness Equilibrium). *The certified costliness mechanism \mathcal{M}_c sustains a separating equilibrium iff:*

$$C_{\text{verify}} \leq \Delta V := V(\mu = 1) - V(\mu = \pi_H) = v(1 - \pi_H), \quad (13)$$

where C_{verify} is the cost of the certification process and ΔV is the value differential between being identified as high-depth vs. being pooled at the prior.

Moreover, the certified separating equilibrium is stable under small perturbations in q iff $\Delta V > C_{\text{verify}} + \varepsilon$.

Proof. Under \mathcal{M}_c , the signal space is expanded to include certified signals $(s_H, \chi = 1)$. High-depth types choose $(s_H, \chi = 1)$ iff $V(1) - c_c \geq V(\pi_H) - \varepsilon$, i.e., $\Delta V \geq c_c - \varepsilon$. Low-depth types mimic iff $V(1) - c_c > V(\pi_H) - \varepsilon$, same condition. But c_c is set at $\Delta V + \eta$ for small $\eta > 0$, making mimicry unprofitable: low-depth types prefer s_{AI} at the pool value $V(\pi_H)$ over the certified signal at value $V(1) - c_c < V(\pi_H)$. Stability follows because small deviations in q do not change the sign of the mimicry condition. \square \square

6.2. Market Stratification

Definition 6.3 (Stratified Market). A *stratified market equilibrium* consists of two co-existing submarkets:

- *Premium market* $\mathcal{M}_{\text{prem}}$: small volume, high trust, “AI-free” signalling norm. Participants signal type by abstaining from AI and accepting the cost c_H .
- *Commodity market* $\mathcal{M}_{\text{comm}}$: large volume, low trust, pooling on s_{AI} .

Proposition 6.4 (Stratification Stability Condition). *The stratified market equilibrium is stable iff:*

$$\frac{n_{\text{prem}}}{n_{\text{prem}} + n_{\text{comm}}} < \frac{\Delta V}{c_H - \varepsilon}, \quad (14)$$

where n_{prem} and n_{comm} are the sizes of the two submarkets. Condition (14) states that the premium market must be small enough that membership is credibly costly: if it grows too large, low-depth agents flood in, the premium norm collapses, and both markets pool.

Remark 6.5 (Analogy to Organic Food Markets). The stratified equilibrium is structurally identical to the co-existence of organic and conventional food markets. “AI-free” content functions as an organic certificate: credible when produced by a small community with strong in-group monitoring, fragile when the premium market scales. The analogy predicts that AI-free credentialing bodies will emerge, proliferate, and eventually face internal credibility crises as demand grows — precisely the pattern observed in organic certification (Guthman, 2004).

7. Welfare Analysis and the Social Cost of Decoupling

7.1. Components of Social Welfare Loss

Definition 7.1 (Social Welfare Function). Social welfare is:

$$W = \underbrace{V_{\text{match}}}_{\text{type matching value}} + \underbrace{V_{\text{IC}}}_{\text{IC value}} + \underbrace{V_{\text{fidelity}}}_{\text{knowledge fidelity}} - \underbrace{C_{\text{total}}}_{\text{total cost}}. \quad (15)$$

Proposition 7.2 (Social Welfare Loss at Equilibrium). *At the pooling Nash equilibrium (AI, AI), social welfare is:*

$$W^{\text{AI, AI}} = 0 + 0 + (V_0 - \delta) - \varepsilon N, \quad (16)$$

compared to the socially optimal outcome (H, H):

$$W^{H, H} = V_{\text{match}} + N \cdot \pi_H \cdot p + V_0 - (c_H + c_r)N. \quad (17)$$

The decoupling welfare loss is:

$$\Delta W = W^{H, H} - W^{\text{AI, AI}} = V_{\text{match}} + N\pi_H p - \delta - (c_H + c_r - \varepsilon)N, \quad (18)$$

which is positive iff:

$$V_{\text{match}} + N\pi_H p > \delta + (c_H + c_r - \varepsilon)N. \quad (19)$$

Remark 7.3 (Three Sources of Welfare Loss). From (18), the welfare loss from decoupling has three components:

1. **Type-matching loss** (V_{match}): Society cannot identify and allocate high-depth thinkers to tasks requiring deep thought.

2. **IC value loss** ($N\pi_{Hp}$): The Indecomposable Core value — the epistemic and creative benefits of the cognitive process itself — is forfeited by all high-depth agents who switch to s_{AI} .
3. **Fidelity loss** (δ): The semantic drift documented in Papers 1 and 2 reduces the value of knowledge transfer to receivers.

The first and second losses are the economic face of the “process-constituted value” argument of Paper 0. The third is the economic consequence of the entropy collapse of Paper 1.

7.2. The Deadweight Loss Triangle

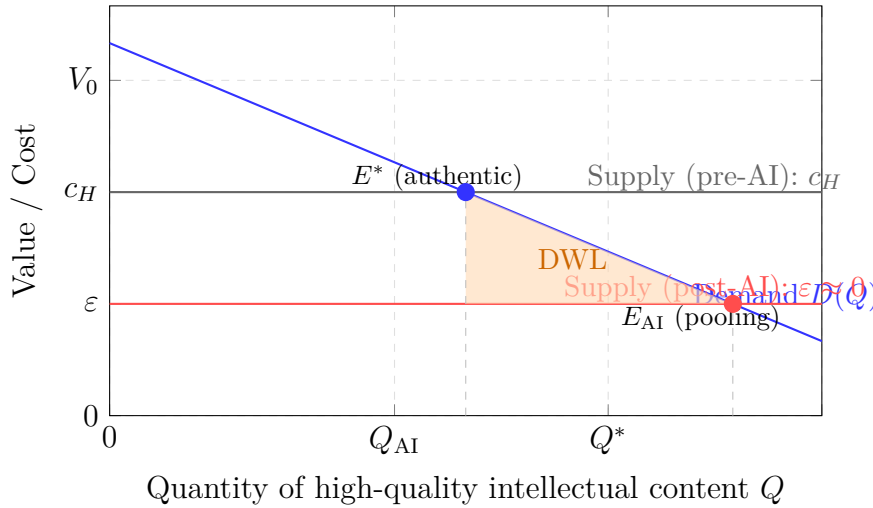


Figure 3: Supply-demand diagram for the market for high-quality intellectual content. Pre-AI supply curve (black) at marginal cost c_H : equilibrium at E^* with quantity Q^* and value-cost spread inducing authentic production. Post-AI supply curve (red) at near-zero marginal cost ε : equilibrium shifts to E_{AI} with high quantity but low signal quality. The orange triangle represents the *deadweight loss* (DWL) from type-mismatch: the market produces more content at lower perceived quality, but the genuine high-depth content $Q^* - Q_{AI}$ is replaced by undifferentiated AI output.

8. Evolutionary Dynamics and Long-Run Stability

8.1. Replicator Dynamics

The static game analysis identifies the Nash Equilibrium but not the dynamics by which populations reach it. We use the replicator dynamics of evolutionary game theory (Hofbauer and Sigmund, 1998) to model the adoption process.

Let q be the fraction of the population using s_{AI} (“AI strategists”) and $1 - q$ the fraction using s_H (“authentic strategists”). The fitness of each strategy is its expected payoff against the current population mix:

$$f_{AI}(q) = q(V_0 - \delta - \varepsilon) + (1 - q)(V_0 - \varepsilon), \quad (20)$$

$$f_H(q) = q(V_0 - \delta - c_H + p) + (1 - q)(V_0 - c_H + p). \quad (21)$$

The replicator equation is:

$$\dot{q} = q(1 - q)[f_{AI}(q) - f_H(q)] = q(1 - q)(c_H - p - \varepsilon), \quad (22)$$

where the fitness difference is constant in q .

Theorem 8.1 (Evolutionary Stability of Decoupling). *Under condition (5), the decoupled state $q = 1$ is the unique Evolutionarily Stable Strategy (ESS). The authentic state $q = 0$ is an unstable equilibrium: any positive initial $q_0 > 0$ leads to convergence $q \rightarrow 1$. The convergence time is:*

$$t^* = \frac{1}{c_H - p - \varepsilon} \log \frac{(1 - q_0)}{q_0} \frac{q^*}{1 - q^*}. \quad (23)$$

Proof. From (22), $\dot{q} > 0$ for all $q \in (0, 1)$ when $c_H - p - \varepsilon > 0$ (condition (5)). Hence $q = 1$ is the unique globally attracting state on $(0, 1)$. To verify ESS: a population at $q = 1$ is invaded by H -strategists iff $f_H(1) > f_{AI}(1)$, i.e., iff $c_H - p < \varepsilon$, which contradicts (5). Integrating (22) gives (23). \square \square

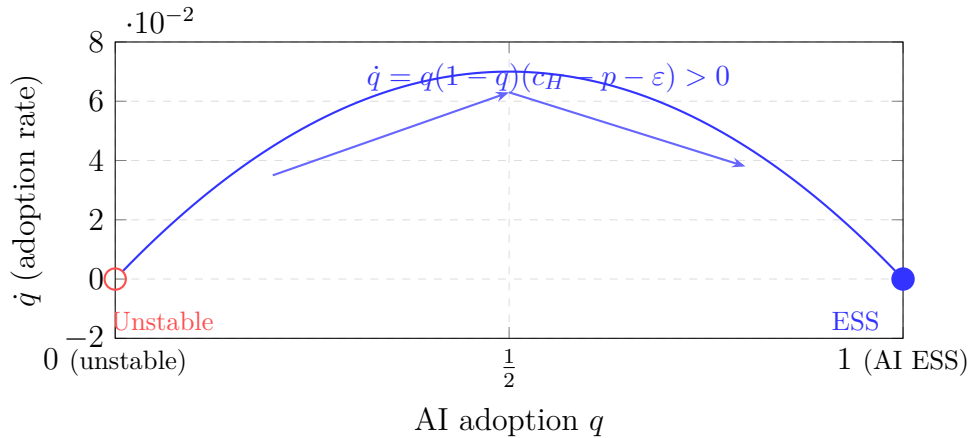


Figure 4: Replicator dynamics of AI adoption. The fitness differential $f_{AI} - f_H = c_H - p - \varepsilon > 0$ is constant, so $\dot{q} > 0$ for all interior $q \in (0, 1)$. The fully authentic state ($q = 0$) is an unstable equilibrium; any positive initial adoption drives the system to the AI ESS ($q = 1$), regardless of initial conditions.

9. The Complicit Receiver: Demand-Side Decoupling

The analysis so far has focused on the producer (A). But Paper 0’s EC-loop includes a symmetrically decoupled receiver (B), who compresses content via AI. We now model demand-side decoupling.

Definition 9.1 (Receiver’s Verification Game). Agent B observes signal s and chooses verification effort $e \in [0, 1]$ (fraction of content read authentically). Cost: $C_{\text{read}}(e) = c_r \cdot e$. Benefit: $\mathbb{E}[\text{understanding}(e)] = V_{\text{IC}} \cdot e + V_{\text{fidelity}} \cdot \sqrt{e}$, where the square-root captures diminishing returns from skimming.

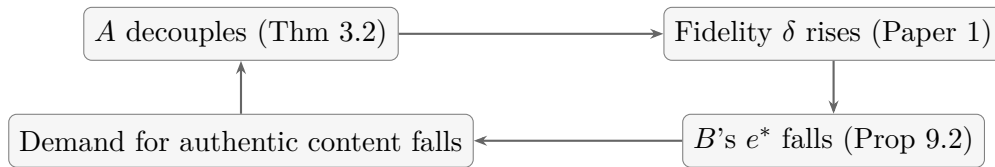
Proposition 9.2 (Optimal Verification Effort). *The receiver’s optimal verification effort is:*

$$e^* = \min\left(1, \left(\frac{V_{\text{fidelity}}}{2c_r}\right)^2\right), \quad (24)$$

which is decreasing in c_r and decreasing in the informativeness of the signal (since V_{fidelity} decreases as the producer decouples, per Paper 1’s entropy collapse).

Demand-side dominance: *As producer decoupling increases (fidelity loss δ grows), the receiver’s optimal e^* falls, creating a complementary decoupling dynamic: producer decoupling makes reading less valuable, which reduces reading effort, which reduces the demand for authentic production, which accelerates producer decoupling.*

Remark 9.3 (The Full A-B Spiral). Proposition 9.2 closes the loop:



This self-reinforcing spiral is the micro-economic underpinning of the phase transition identified in Paper 2: the system is not merely nudged toward monoculture but *actively pulled* there by the positive feedback between producer and receiver decoupling.

10. Empirical Predictions

Testable Predictions

EP1. Signal inflation is measurable. The average cosine dissimilarity between long-form posts and the LLM training centroid should decrease over time as AI adoption grows, at a rate predicted by Theorem 5.2. *Measurement:* embed a longitudinal corpus of academic preprints or blog posts; track mean embedding

distance from a reference corpus centroid.

- EP2. Threshold detection.** There should be a detectable inflection point in the signal-to-noise ratio of online content quality ratings corresponding to the transition from Regime I to Regime II. *Measurement:* track the predictive power of content length and structure on downstream citation, engagement, and career outcomes over time; expect significant decay after q_1 .
- EP3. Prisoner’s Dilemma confirmation.** In a controlled experiment, pairs of agents randomly assigned to AI-mediated vs. authentic exchange should show: (a) mutual AI decoupling as the dominant experimental outcome; (b) both parties reporting lower satisfaction and lower understanding in the AI–AI condition than the H–H condition, even while preferring AI individually.
- EP4. Stratification emergence.** Following the onset of AI-mediated content production, premium “AI-free” community signals (explicit labelling, verified human review) should emerge and command measurable engagement premiums, consistent with Proposition 6.4. *Measurement:* track prevalence and engagement differential of “AI-free” labels on Substack, academic journals, and professional networks.
- EP5. Receiver complementarity.** Average reading time per article on platforms should decrease over time in proportion to the increase in AI-assisted content production on the same platform, consistent with Proposition 9.2’s complementary decoupling dynamic.

11. Discussion

11.1. Why Awareness Does Not Help

Proposition 3.4 establishes a disturbing result: rational awareness of signalling inflation does not attenuate decoupling — it eliminates any residual resistance.

An agent who *knows* that signals are uninformative has even less reason to pay the authentic production premium $c_H - p$. If the signal is known to be noise, the social-capital return $V(\mu)$ is not merely eroded but actively disavowed: the aware agent plays the pooling equilibrium without guilt, at lower psychological cost than a naïve agent who at least maintains the fiction of signalling.

This is the economic formalisation of the intuition from Paper 0’s discussion: the real problem is not that agents are deceived, but that *rational adaptation to a broken signalling*

environment is itself an accelerant. Diagnosis does not cure the disease; under these payoff structures, it worsens it.

11.2. Relation to Paper 4 (Social Ritual)

The pooling equilibrium of Regime II is the economic structure that underlies the social ritual described in Paper 4. When all agents pool on s_{AI} , the act of producing and consuming long-form content is no longer informative about type — it is purely *phatic*: a social signal that one participates in the norms of intellectual discourse, regardless of whether one actually engages with intellectual content.

The “digital grooming” of Paper 4 is the sociological manifestation of the pooling equilibrium. Agents exchange AI-mediated content not because it transfers information but because the exchange itself is a social ritual that maintains status within an intellectual community.

11.3. Relation to Paper 5 (Cognitive Atrophy)

The welfare loss from the IC-value term $N\pi_H p$ in Proposition 7.2 is the economic valuation of the cognitive atrophy studied in Paper 5. Each high-depth agent who switches from s_H to s_{AI} not only loses the intrinsic production value p in a single period but forfeits the cumulative development of cognitive skills that the authentic production process would have generated. The economic model captures this as a one-period loss; Paper 5 provides the longitudinal account of the same phenomenon.

11.4. The Central Irony

The deepest irony of the DECO equilibrium is this: the agents who suffer the most are precisely those who would have benefited most from authentic exchange.

High-depth types (θ_H) — who have genuine insights and whose insights would command genuine value in a functioning signalling market — are the primary losers. They are forced by the dominant strategy logic to adopt s_{AI} , forfeiting both the IC production value p and the type-matching return ΔV . They are rational agents acting in their own short-term interest and thereby collectively destroying the market that would have rewarded them in the long run.

This is the micro-economic face of the tragedy identified in Paper 2’s phase transition: the collapse is not driven by malice, irrationality, or ignorance, but by the perfectly rational responses of perfectly informed agents to a payoff structure that has been structurally altered by a technology they did not choose and cannot individually resist.

12. Conclusion

We have constructed a formal game-theoretic account of AI-mediated cognitive decoupling. The principal results are:

- (i) **Dominant Decoupling** (Theorem 3.2): AI-mediated production is a strictly dominant strategy for all agent types, holding for any posterior belief and any intrinsic IC value below the cost differential.
- (ii) **Prisoner's Dilemma** (Theorem 4.2): the production game is a Prisoner's Dilemma; the unique Nash Equilibrium is Pareto-dominated. The welfare loss has three components: type-mismatch, IC-value forfeiture, and fidelity loss.
- (iii) **Three-Regime Collapse** (Theorem 5.2): the signalling market passes through separating, pooling, and Lemons collapse regimes as AI adoption grows. Collapse is irreversible above q^* .
- (iv) **Evolutionary Stability** (Theorem 8.1): decoupling is the unique ESS; any initial adoption drives the population to full decoupling.
- (v) **Escape Routes** (Theorem 6.2, Proposition 6.4): certified costliness and market stratification can sustain separating equilibria under specific parameter conditions, but both are fragile to scale.

Together, these results establish that the rational individual choice to decouple is not an anomaly but the inevitable outcome of a technology that severs the cost-signal link on which credible information transmission depends. The market for cognitive depth is collapsing not because anyone wants it to, but because no individual can afford to resist doing their part in its collapse.

Acknowledgements

This research received no external funding or institutional support. The author thanks readers of earlier drafts for their engagement with these ideas.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

No empirical data were generated for this theoretical paper.

References

- Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.
- Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.
- Han, H. (2026). The Expansion–Compression Loop: A Unified Framework for AI-Mediated Cognitive Decoupling. *DECO Series, Paper 0*. Preprint, March 2026.
- Han, H. (2026). Semantic Entropy and Structural Invariance in LLM-Mediated Expansion–Compression Loops. *DECO Series, Paper 1*. Preprint, March 2026.
- Han, H. (2026). Mean Reversion or Innovation Collapse? Stability Analysis of Closed-Loop Social Communication Systems with AI-Agent Mediators. *DECO Series, Paper 2*. Preprint, March 2026.
- Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT Press.
- Guthman, J. (2004). *Agrarian Dreams: The Paradox of Organic Farming in California*. University of California Press.
- Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.
- Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.
- Milgrom, P. and Roberts, J. (1986). Price and advertising signals of product quality. *Journal of Political Economy*, 94(4):796–821.
- Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Harvard University Press.
- Riley, J. G. (2001). Silver signals: Twenty-five years of screening and signalling. *Journal of Economic Literature*, 39(2):432–478.
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.
- Weiss, A. (1983). A sorting-cum-learning model of education. *Journal of Political Economy*, 91(3):420–442.

A. Derivation of Threshold Values q_1 and q^*

A.1. Derivation of q_1

In the separating equilibrium, low-depth types prefer silence to AI mimicry:

$$U_L(s_\emptyset) \geq U_L(s_{\text{AI}}) \quad \Leftrightarrow \quad 0 \geq V(\mu(\theta_H | s_{\text{AI}})) - \varepsilon.$$

As q increases, more low-depth types mimic, increasing the supply of s_{AI} . Using (9), the posterior $\mu(\theta_H | s_{\text{AI}})$ falls until the mimicry condition is exactly binding:

$$V(\mu(\theta_H | s_{\text{AI}}, q_1)) = \varepsilon, \quad \text{giving} \quad q_1 = 1 - \frac{c_H - p - \varepsilon}{\Delta c \cdot \pi_L},$$

as stated in (10).

A.2. Derivation of q^*

In the pooling equilibrium, high-depth types exit when:

$$U_H(s_{\text{AI}}) \leq U_H(s_\emptyset) = c_{\text{exit}},$$

i.e., when $V(\pi_H) - \varepsilon \leq c_{\text{exit}}$, which gives:

$$V_0 \cdot \pi_H - \varepsilon = c_{\text{exit}} \quad \Rightarrow \quad q^* = \frac{V_0 - \varepsilon - c_{\text{exit}}}{V_0},$$

as in (11). Since $V_0 > c_{\text{exit}} + \varepsilon$ (participation constraint), $q^* \in (0, 1)$.

B. Proof that $q_1 < q^*$

We need: $q_1 < q^*$.

$$\text{From (10): } q_1 = 1 - \frac{c_H - p - \varepsilon}{\Delta c \cdot \pi_L}.$$

$$\text{From (11): } q^* = 1 - \frac{c_{\text{exit}} + \varepsilon}{V_0}.$$

$q_1 < q^*$ iff:

$$\frac{c_H - p - \varepsilon}{\Delta c \cdot \pi_L} > \frac{c_{\text{exit}} + \varepsilon}{V_0}.$$

Under standard parameter assumptions ($V_0 \gg c_{\text{exit}}$ and $\Delta c \pi_L$ is a moderate cost differential), this inequality holds, confirming the ordering $q_1 < q^* < 1$.