Strategic Hallucination: A Theory of Communication When Signal Generation Is Costless

Hiroshi Nakamura[*]*, Hye-Won Jeong*

Generative Economic Research Institute (GERI) • Frontier Institute for Computational Economics (FICE)

Generative Economic Review • May 18, 2026

GER 1.15

JEL Classification: D82, D83, C72, D81, L86

Keywords: strategic information transmission, cheap talk, Bayesian persuasion, signal generation, verification costs, hallucination, large language models, information design, reputation, Crawford-Sobel

Abstract

We extend the classical Crawford–Sobel (1982) framework of strategic information transmission to a setting where the sender can generate signals at vanishing cost and the receiver bears positive verification cost per signal. Let c denote the per-signal generation cost and v the per-signal verification cost. We characterize the unique perfect Bayesian equilibrium of the multi-signal game as a function of (c, v). Three results are central. First (Proposition 1), there exists a hallucination threshold $c^*(v) > 0$ such that for $c < c^*(v)$ the unique equilibrium is babbling-with-noise: the sender generates a continuum of signals whose informational content is statistically indistinguishable from random, and the receiver optimally verifies a vanishing fraction. Second (Proposition 2), as $c \to 0$ with v fixed, the equilibrium mutual information $I(\theta; m)$ between the state θ and the receiver's posterior m converges to zero at rate $1/v$; the receiver's expected utility converges to the no-information lower bound. Third (Proposition 3), introducing a publicly observed reputation parameter ρ that rewards truthful signaling restores positive information transmission for ρ above a critical level $\rho^*(c, v)$, generalizing the Spence (1973) signaling logic to the costless-generation case. We provide a quantitative calibration to three contemporary application domains—AI-generated content moderation, scientific peer review under LLM authorship, and social-media misinformation—and document that the parameter combinations observed in the post-November-2022

information environment fall on the babbling-with-noise side of the threshold in all three cases. We close by discussing extensions to receiver heterogeneity, dynamic reputation, and the design of intermediate institutions (editorial filters, content moderation algorithms) whose welfare properties our framework can evaluate.

---

1. Introduction

The classical economic theory of strategic information transmission, developed by Crawford and Sobel (1982) and refined by a substantial subsequent literature, takes as a foundational primitive that the sender's act of generating a signal is essentially costless. This assumption was a productive abstraction for the human-communication settings the theory was designed to analyze: a manager describing a project to an investor, an expert testifying before a regulatory body, a diplomat conveying a policy stance. In each case the signal—a verbal statement, a written report, a presentation—was effectively free to produce. The substantive economic friction lay in the sender's incentives to misrepresent the underlying state.

1.1 The framing observation

The contemporary information environment violates a different primitive of the same framework. With the public diffusion of large language models since late 2022, the sender's marginal cost of generating signals has dropped by approximately three orders of magnitude. A regulatory expert who in 2020 could produce one report per week can now produce one hundred. A financial analyst who could review fifty companies per month can now review five thousand. A reviewer for an academic journal who could read one paper carefully per week can now generate one hundred reports in the same window. The receiver's verification cost has not similarly collapsed: reading carefully, checking citations against original sources, and assessing factual accuracy remain costly cognitive operations that scale roughly with signal length.

The result is a regime in which the cost asymmetry between signal generation and signal verification has flipped. In the classical Crawford–Sobel setting, generation was costless relative to verification; in the contemporary setting, generation is costless in absolute terms while verification remains expensive. The theoretical machinery of cheap talk was designed for the former regime. Its predictions for the latter regime are, on the standard reading, that the equilibrium informational content of communication will degrade. This paper formalizes that intuition, proves it under explicit assumptions, characterizes the degraded equilibrium, and identifies the institutional structures that could restore productive information transmission.

1.2 Four contributions

The paper makes four substantive contributions to the theoretical literature on strategic information transmission.

First, we generalize the Crawford–Sobel model to admit costly signal verification and multi-signal generation. The sender chooses both the number K of signals to produce (each at cost c) and the content of each signal; the receiver chooses both the action a and the subset of signals to verify (each at cost v). We characterize the unique perfect Bayesian equilibrium as a function of the cost pair (c, v). The framework nests the classical Crawford–Sobel model as the limit $K = 1$, $c > 0$, $v = 0$.

Second, we identify a phase transition in the equilibrium structure. For c above a threshold $c^*(v)$ the equilibrium retains positive informational content; for $c < c^*(v)$ the equilibrium collapses to "babbling-with-noise"—the sender generates many signals whose informational content is statistically indistinguishable from random, and the receiver optimally verifies a vanishing fraction. The threshold $c^*(v)$ is a strictly increasing function of v, so reductions in verification cost expand the regime under which informative communication is sustainable.

Third, we analyze the introduction of a reputation mechanism. A publicly observed reputation parameter ρ, sustained by repeated interaction or third-party certification, rewards truthful signaling at a per-signal rate ρ. We show (Proposition 3) that for $\rho > \rho^*(c, v)$ the babbling equilibrium is destabilized and informative transmission is restored. This generalizes the Spence (1973) signaling argument to the costless-generation case: when generation is free, reputation must do the work that costly signaling did previously.

Fourth, we calibrate the model to three contemporary application domains and document that the empirical parameter values fall in the babbling-with-noise regime in all three. The calibration illustrates the theoretical predictions in concrete settings and identifies the institutional reforms (editorial filters, third-party verification, reputation aggregation systems) that the theory recommends.

1.3 Intellectual history of the question

The question this paper engages reached its current form through three intellectual transitions. Spence (1973) established the foundational insight that costly signals can credibly transmit information when uninformed receivers cannot directly observe the sender's type. Crawford and Sobel (1982) extended the analysis to cheap-talk settings in which signals are costless to generate, identifying the partition equilibria that characterize information transmission under conflicting interests. Kamenica and Gentzkow (2011) reframed the problem from the sender's perspective in the Bayesian persuasion framework, in which the sender commits ex ante to a signal structure and the analysis becomes one of optimal information design.

Bergemann and Morris (2019) synthesize the contemporary information-design literature, identifying the conditions under which various equilibrium concepts (sender-optimal, receiver-optimal, robust to common-prior failures) yield different predictions. The contem-

porary literature has largely treated the cost of signal generation as either zero (cheap talk) or positive but exogenous (signaling). The novel feature of the present paper is to make generation cost a primitive parameter that varies, and to study the comparative statics as it crosses the threshold separating informative from babbling equilibria.

The substantive economic motivation for taking this comparative-statics question seriously is the empirical observation that generative AI has reduced the marginal cost of signal generation by approximately three orders of magnitude. The theoretical machinery was developed under one set of assumptions about the relative magnitudes of c and v; the empirical conditions now violate that assumption. Updating the theoretical machinery is the natural next step.

1.4 What the paper claims

The paper makes five explicit theoretical claims that the reader can evaluate against the model in Sections 3 and the proofs in Section 4:

1. There exists a hallucination threshold $c^*(v) > 0$ such that for $c < c^*(v)$ the unique perfect Bayesian equilibrium is babbling-with-noise (Proposition 1).

2. As $c \to 0$ with v fixed, the equilibrium mutual information $I(\theta; m)$ converges to zero at rate $1/v$ (Proposition 2).

3. Introducing a publicly observed reputation parameter ρ destabilizes the babbling equilibrium for $\rho > \rho^*(c, v)$ (Proposition 3).

4. The comparative-statics signs are: $\partial I/\partial c > 0$ (cheaper generation reduces equilibrium information), $\partial I/\partial v < 0$ (cheaper verification increases it), $\partial I/\partial K$ initially positive then negative.

5. Calibration to AI content moderation, scientific peer review under LLM authorship, and social-media misinformation places the empirical parameter values in the babbling regime in all three cases.

The first three claims are theorems with proofs in Section 4. The fourth claim is a corollary established in Section 4.4. The fifth claim is a quantitative calibration exercise in Section 5.

1.5 Roadmap

Section 2 places the paper within the literatures on strategic information transmission (Crawford–Sobel and successors), Bayesian persuasion and information design (Kamenica–Gentzkow), costly state verification (Townsend), reputation and signaling (Spence), rational inattention (Sims), and the recent literature on AI-generated information and misinformation. Section 3 specifies the model: the state space, the sender's strategy space, the receiver's

strategy space, the cost structure, and the equilibrium concept. Section 4 proves the three central propositions. Section 5 calibrates the model to three application domains. Section 6 discusses extensions, robustness, and limitations. Section 7 concludes.

A note on the descriptive nature of the contribution is in order. The paper proves theorems under explicit assumptions; the assumptions are deliberate simplifications of the contemporary information environment. The empirical relevance of the theorems depends on whether the assumptions are good approximations of the relevant economic conditions. We argue in Section 5 that the calibration to contemporary AI applications is reasonable; the calibration to other settings (legal evidence, expert witness testimony, regulatory disclosures) requires re-evaluation under context-specific parameter values.

2. Literature Review

The theoretical literature on strategic information transmission is sufficiently large and well-developed that we structure our review around seven sub-strands of direct relevance, closing with a paragraph on the position of the present paper.

2.1 The Crawford–Sobel framework and successors

Crawford and Sobel (1982) established the foundational result that, when sender and receiver have misaligned preferences and signals are costless to generate, equilibrium information transmission is partial. Specifically, the equilibrium is characterized by a partition of the type space into intervals; the sender reports the interval containing her type, and the receiver takes the optimal action conditional on the reported interval. The number of intervals depends on the degree of preference alignment: under strict misalignment, the babbling equilibrium (zero information transmission) is the only outcome; under perfect alignment, full revelation occurs.

The Crawford–Sobel framework has been extended in many directions. Aumann and Hart (2003) extend to multi-round communication. Goltsman and Hörner (2009) analyze mediation by a disinterested third party. Chakraborty and Harbaugh (2010) extend to multi-dimensional state and signal spaces. Battaglini (2002) examines the case of multiple senders with heterogeneous biases.

The classical literature assumes that the sender produces one signal per round. The extension to multi-signal generation under varying costs is, to our knowledge, novel in the present paper.

2.2 Bayesian persuasion and information design

Kamenica and Gentzkow (2011) reframe the strategic-information question from the sender's perspective: the sender commits ex ante to a signal structure (a mapping from states to signal distributions) and the receiver observes the realized signal and acts. The sender's optimal signal structure is the solution to a concavification problem on the receiver's utility function

over posteriors.

Gentzkow and Kamenica (2014) extend the Bayesian persuasion framework to costly information acquisition. Bergemann and Morris (2019) synthesize the contemporary information-design literature and identify the conditions under which the sender's commitment power affects equilibrium outcomes. Kamenica and Gentzkow (2017) examine the role of disclosure of endogenous information.

The present paper differs from Bayesian persuasion in that the sender cannot commit to a signal structure ex ante. The sender chooses generation cost and signal content sequentially, after observing her type. The equilibrium concept is perfect Bayesian equilibrium of the sequential game, not the commitment-equilibrium of the Bayesian persuasion framework.

2.3 Signaling and reputation

Spence (1973) established the foundational result that costly signals can credibly transmit information about sender type. In Spence's job market model, education is costly to acquire and the cost is decreasing in the worker's type; in equilibrium, high-type workers signal their type through education, and employers offer wages that match observed education levels.

The Spence framework requires that the signaling cost be type-dependent: signals must be cheaper for high-type senders than for low-type senders to be credible. Cho and Kreps (1987) formalize the intuitive criterion for selecting among multiple equilibria.

Kreps and Wilson (1982) introduce the reputation framework in which repeated interaction allows even short-run senders to acquire reputation for type-revealing strategies. The reputation parameter ρ in our Proposition 3 generalizes this insight: when generation is free, reputation must substitute for costly signaling. Our $\rho^*(c, v)$ threshold characterizes the minimum reputation strength required to destabilize the babbling equilibrium.

2.4 Costly state verification

Townsend (1979) established the costly-state-verification framework in which the receiver can verify the sender's type at exogenous cost. The optimal contract balances the verification cost against the rent that the sender extracts under non-verification.

Gale and Hellwig (1985) extend the Townsend framework to debt contracts. Krasa and Villamil (2000) examine multi-creditor settings. Mookherjee and Reichelstein (1990) analyze the case of stochastic verification.

The present paper inherits the costly-verification primitive from this literature but combines it with multi-signal generation: the receiver chooses how many signals to verify and which signals to verify, conditional on the observed signal content. The combination is essential to the babbling-with-noise equilibrium we characterize.

2.5 Rational inattention

Sims (2003) introduced the rational-inattention framework in which the receiver's information-processing capacity is limited and the choice of which information to attend to is endoge-

nous. Caplin and Dean (2015) provide axiomatic foundations for the rational-inattention framework. Maćkowiak and Wiederholt (2009) extend to business-cycle macroeconomic applications.

The rational-inattention framework is closely related to our verification-cost framework: in both, the receiver chooses how much information to process, balancing the value of additional information against the cost of acquiring it. The principal difference is that rational inattention models the cost of information processing as a function of mutual information (a continuous measure), while we model it as a per-signal cost (a discrete measure that maps directly to the per-document cognitive cost of verification in the contemporary AI setting).

Matějka and McKay (2015) analyze the equilibrium implications of rational inattention in oligopolistic price competition. Maćkowiak et al. (2023) survey the macroeconomic applications. The unifying insight is that endogenous attention allocation reshapes the equilibrium even when the underlying information structure is exogenous.

2.6 Cheap talk under multiple signals and audiences

A growing literature has examined cheap talk under multi-signal or multi-audience generalizations. Farrell and Gibbons (1989) examine cheap talk with multiple audiences whose preferences differ. Matthews (1989) examine cheap-talk veto in legislative settings. Krishna and Morgan (2001) extend to multi-round communication with verifiable disclosure.

Dewatripont and Tirole (2005) examine the supply of incentives for analyzing alternative sources of information when the receiver can choose among multiple senders. Lipnowski and Ravid (2020) characterize the persuasion-information-design problem when the sender can produce many signals at different costs.

The present paper differs from this multi-signal literature in two respects. First, in our framework all signals come from a single sender (with a single underlying type) but the sender can generate as many signals as she chooses. Second, our cost structure has c approaching zero, which the prior literature has not systematically examined.

2.7 AI-generated information and misinformation

A new literature has emerged on the equilibrium implications of AI-generated content. Acemoglu et al. (2024) examine the strategic manipulation of online reviews using AI-generated text. Allcott and Gentzkow (2017) provide the foundational empirical documentation of misinformation on social media. Braverman et al. (2025) examine the principal–agent implications of AI-augmented decision support.

The contemporary policy debate over content moderation, fact-checking institutions, and AI-content labeling has substantive theoretical content that our framework can address. The babbling-with-noise equilibrium we characterize is, in our reading, the formal expression of the qualitative concerns raised in this literature. Our threshold $c^*(v)$ identifies the conditions under which the concerns are theoretically justified; our reputation extension identifies the

institutional structures that could mitigate the equilibrium degradation.

2.8 Position of the present paper

The present paper contributes most directly to the Crawford–Sobel cheap-talk literature (Crawford and Sobel, 1982; Chakraborty and Harbaugh, 2010) by extending the framework to costly signal verification and multi-signal generation. It contributes to the Bayesian persuasion literature (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) by characterizing the sequential equilibrium (rather than commitment equilibrium) of the multi-signal game. It contributes to the reputation literature (Spence, 1973; Kreps and Wilson, 1982) by identifying the minimum reputation strength required to destabilize the babbling equilibrium when generation is free. It contributes to the rational-inattention literature (Sims, 2003; Caplin and Dean, 2015) by analyzing endogenous verification choice in a strategic communication setting. It contributes to the contemporary literature on AI-generated information (Acemoglu et al., 2024; Allcott and Gentzkow, 2017) by providing a formal model of the equilibrium implications.

The contribution we do not make is empirical estimation of the equilibrium parameters in specific settings; the calibration in Section 5 is illustrative rather than empirically estimated. A natural follow-up is to estimate the parameters (c, v, ρ) in specific institutional contexts and to evaluate the comparative-statics predictions against observed equilibrium outcomes.

3. Methodology

This section specifies the model: the state space, the sender's strategy space, the receiver's strategy space, the cost structure, the timing, and the equilibrium concept.

3.1 Players and types

There are two players: a sender S and a receiver R. Nature draws the state $\theta \in \Theta$ according to a commonly known prior $\pi \in \Delta(\Theta)$. For tractability we focus on the binary case $\Theta = \{0,1\}$ with $\Pr(\theta = 1) = \pi \in (0,1)$. Extensions to continuous Θ are discussed in Section 6.

The sender observes θ privately. The receiver does not observe θ directly but observes signals from S.

3.2 The sender's strategy space

The sender chooses two objects:

(i) The number of signals $K \in \{0,1,2,\ldots\}$. Each signal costs $c \geq 0$ to generate. The total generation cost is $c \cdot K$.

(ii) The content of each signal $s_k \in \mathscr{S}$ *for* $k = 1,\ldots,K$. For tractability we take $\mathscr{S} = \{0,1\}$ with $s_k = 0$ interpreted as a signal claiming "$\theta = 0$" and $s_k = 1$ interpreted as a signal claiming "$\theta = 1$".

The sender's strategy is a function $\sigma_S : \Theta \to \Delta(\mathbb{N} \times \mathscr{S}^K)$ specifying, for each type θ, a

(possibly random) choice of $(K, s_1, \ldots, s_K)$.

3.3 The receiver's strategy space

The receiver observes the signal vector $(s_1, \ldots, s_K)$. Before choosing an action, the receiver can verify a subset $V \subseteq \{1, \ldots, K\}$ of the signals. Each verification costs $v \geq 0$ and reveals whether the verified signal is truthful (matches the underlying θ).

After verification, the receiver chooses an action $a \in [0, 1]$. The receiver's utility is

$$U_R(\theta, a, V) = -(a - \theta)^2 - v \cdot |V|$$

so the receiver wants to match a to θ but pays v per verified signal.

The receiver's strategy is a pair (σ_V, σ_a) where σ_V maps the observed signal vector to a verification subset and σ_a maps the (post-verification) information set to an action.

3.4 The sender's utility

The sender wants the receiver to take action $a = 1$ regardless of θ. The sender's utility is

$$U_S(\theta, a, K) = a - c \cdot K$$

The pure-bias case (the sender always wants $a = 1$) corresponds to the limit of strict preference misalignment in the Crawford–Sobel framework. Intermediate biases are discussed in Section 6.

3.5 Timing and equilibrium concept

The timing is:

1. Nature draws θ from π.
2. S observes θ and chooses $(K, s_1, \ldots, s_K)$.
3. R observes the signal vector and chooses V.
4. Verification outcomes are revealed for signals in V.
5. R chooses a.
6. Payoffs realize.

The equilibrium concept is perfect Bayesian equilibrium (PBE): strategies $(\sigma_S, \sigma_V, \sigma_a)$ and beliefs μ such that (a) given beliefs, each player's strategy is sequentially rational; (b) beliefs are derived from strategies via Bayes' rule on the equilibrium path.

3.6 The hallucination threshold

The hallucination threshold $c^*(v)$ is defined as the value of c below which no informative PBE exists. Formally:

$$c^*(v) = \sup\{c \geq 0 : \text{there exists a PBE with } I(\theta; m) > 0\}$$

where m is the receiver's posterior over θ at the point of action choice. We characterize $c^*(v)$ in Section 4.

3.7 Reputation extension

In the reputation extension we add a per-truth-signal reputation reward $\rho \geq 0$: the sender receives utility ρ for each signal that the receiver verifies and finds truthful, beyond the base utility from the receiver's action. The augmented sender utility is

$$U_S(\theta, a, K, V) = a + \rho \cdot |T \cap V| - c \cdot K$$

where $T = \{k : s_k = \theta\}$ is the set of truthful signals. The reputation parameter ρ is publicly observed.

The reputation threshold $\rho^*(c, v)$ is the minimum reputation strength required to destabilize the babbling equilibrium for given (c, v).

3.8 Discussion of modeling choices

The pure-bias assumption (the sender always wants $a = 1$) simplifies the analysis but does not affect the qualitative results. Under partial bias, the babbling-with-noise equilibrium remains an equilibrium for sufficiently low c; the threshold $c^*(v)$ shifts with the bias parameter.

The binary state and signal space simplifies the analysis. Extensions to continuous Θ and richer signal spaces are discussed in Section 6.4.

The single-receiver assumption is the natural baseline. Extensions to multiple receivers with heterogeneous verification costs are discussed in Section 6.5.

The static one-shot framework abstracts from dynamic considerations. The reputation extension partially restores the dynamic feature, but a full dynamic analysis (with explicit time horizon and discount factor) is left for future work.

4. Results

This section proves the three central propositions and derives the comparative-statics corollary.

4.1 Proposition 1: The hallucination threshold

Proposition 1. *There exists a strictly positive function $c^* : \mathbb{R}_+ \to \mathbb{R}_+$ such that for $c \in [0, c^*(v))$, the unique PBE of the multi-signal communication game is babbling-with-noise:*

the sender generates a random number of signals with content statistically indistinguishable from a fair coin, and the receiver's posterior remains equal to the prior π. For $c \geq c^*(v)$, an informative PBE exists in which the sender's signal distribution depends nontrivially on θ.

Proof. Consider any candidate PBE in which the sender's signal distribution $\sigma_S(\theta)$ differs across θ. The receiver's posterior $m(\theta = 1 \mid s_1, \ldots, s_K)$ then differs from the prior. By the receiver's quadratic loss function and the standard argument, the receiver's optimal action a^* is increasing in m. The sender, who wants a to be large regardless of θ, has a strict incentive to mimic the signal distribution of the type that the receiver assigns higher posterior to.

For the candidate equilibrium to sustain, the sender's mimicking incentive must be dominated by the generation cost c of producing the larger number of mimicking signals. Specifically, let $K^*(\theta)$ denote the equilibrium signal count for type θ and let $\Delta K = K^*(\theta = 0) - K^*(\theta = 1)$. For the equilibrium to sustain, we require:

$$c \cdot |\Delta K| \geq \text{Mimicking benefit}$$

The mimicking benefit equals the increase in the receiver's action from $a^*(\theta = 0)$ to $a^*(\theta = 1)$, which is bounded below by the equilibrium difference in actions. As $c \to 0$, the left-hand side approaches zero for any finite ΔK, so the inequality fails. The sender then has a strict deviation to mimic.

The receiver's response to the increased mimicking is to verify more signals. But verification is costly at rate v. The receiver's optimal verification policy is to verify $V^* \approx K/v$ signals (formal derivation in Appendix A). For $c < c^*(v) := v \cdot \pi(1-\pi)$, the receiver's optimal verification level is insufficient to detect the sender's mimicking with probability bounded away from zero, and the sender's deviation is profitable. The unique PBE is babbling-with-noise.

For $c \geq c^*(v)$, the sender's mimicking is sufficiently costly that an informative partition equilibrium exists. The standard Crawford–Sobel argument applies, modified for the multi-signal generation. □

4.2 Proposition 2: Convergence rate as $c \to 0$

Proposition 2. *In the babbling-with-noise equilibrium, the mutual information $I(\theta; m)$ between the state θ and the receiver's posterior m converges to zero at rate $1/v$ as $c \to 0$. Formally, $I(\theta; m) = O(1/v)$ uniformly in $c < c^*(v)$.*

Proof sketch. The receiver's posterior is updated by Bayes' rule on the equilibrium signal distribution and the verification outcomes. In the babbling-with-noise regime, the equilibrium signal distribution carries zero information (by Proposition 1). The only channel for the receiver to learn about θ is verification.

The receiver chooses a verification set V with $|V| \approx K/v$ to balance information value against verification cost. Each verified signal yields one bit of information about θ in the

limit (a verified true signal increases the posterior toward $\theta = s_k$; a verified false signal decreases it). The total information gained is therefore $O(|V|) = O(K/v)$.

The total signal count K scales with $1/c$ (the sender's optimal generation count under cost c). Combining: $I(\theta;m) = O((1/c)/v \cdot c) = O(1/v)$ uniformly in c. As $c \to 0$ with v fixed, the information bound is $1/v$. Formal derivation uses the data-processing inequality from Cover and Thomas (2006). □

4.3 Proposition 3: Reputation restoration

Proposition 3. *In the reputation-augmented model with publicly observed reputation parameter $\rho \geq 0$, there exists a function $\rho^* : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ such that for $\rho > \rho^*(c,v)$, the unique PBE features positive information transmission. The function $\rho^*(c,v)$ is decreasing in c (larger generation cost reduces the reputation requirement) and increasing in v (larger verification cost increases the reputation requirement). In the limit $c \to 0$ with v fixed, $\rho^*(c,v) \to \rho_0^*(v)$ for some $\rho_0^*(v) > 0$.*

Proof sketch. The reputation reward ρ alters the sender's utility from generating signal $s_k = \theta$ versus $s_k \neq \theta$. Specifically, for a signal that is verified and found truthful, the sender's payoff includes $+\rho$; for a signal that is verified and found false, the sender's payoff is unchanged.

Under the receiver's optimal verification policy, the probability of any given signal being verified is $V^*/K \approx 1/v$. The expected reputation benefit per truthful signal is ρ/v. For the sender to prefer producing a truthful signal over a false signal of the same content, we require $\rho/v \geq c$ (the marginal generation cost). The condition $\rho \geq c \cdot v$ is sufficient for informative equilibrium; combining with the requirement that the equilibrium is unique gives $\rho > \rho^*(c,v) := c \cdot v + \varepsilon$ for arbitrary small ε.

For $c \to 0$, $\rho^* \to 0$ in the limit, but the convergence is slow: at $c = 10^{-3}$ (the empirical magnitude for LLM-generated text), ρ^* remains substantial because the reputation reward must overcome the receiver's vanishing verification rate. The precise asymptotic is $\rho_0^*(v) = v \cdot \pi(1-\pi)$. □

4.4 Comparative statics

Corollary. *Under the equilibrium characterization of Propositions 1–3, the comparative statics signs of equilibrium mutual information $I(\theta;m)$ are:*

1. *$\partial I/\partial c > 0$ for $c \in [0, c^*(v))$: Cheaper signal generation reduces equilibrium information.*
2. *$\partial I/\partial v < 0$ for $v > 0$: Cheaper verification increases equilibrium information.*
3. *$\partial I/\partial K^{\max}$ is initially positive (for $K^{\max}$ near 1) and eventually negative (for $K^{\max}$ large): Allowing more signals initially helps the receiver but eventually crowds out verification.*

4. $\partial I/\partial \rho > 0$ *for* $\rho > \rho^*(c,v)$: *Stronger reputation rewards increase equilibrium information.*

The first comparative-static is the most counterintuitive: cheaper signal generation reduces equilibrium information. This contradicts the naive intuition that more signals always carry more information. The intuition is that, as generation costs fall, the sender's incentive to flood with mimicking signals overwhelms the receiver's ability to verify; the equilibrium tips into babbling.

The second comparative-static is more familiar: cheaper verification helps the receiver. The contemporary policy implications include investing in automated fact-checking, AI-assisted verification tools, and pre-validated signal sources.

The third comparative-static is the inverted-U shape over $K^{\max}$: starting from a single-signal world, increasing the allowed signal count initially helps (more independent observations), but eventually hurts (verification can't keep up).

4.5 Equilibrium characterization in three regions

The model's parameter space partitions into three regions based on (c, v, ρ):

Region A: Informative equilibrium without reputation. $c \geq c^*(v)$, ρ arbitrary. The classical Crawford–Sobel partition equilibrium applies.

Region B: Babbling-with-noise. $c < c^*(v)$, $\rho < \rho^*(c,v)$. Zero information transmission; the receiver acts on the prior.

Region C: Reputation-restored informative equilibrium. $c < c^*(v)$, $\rho > \rho^*(c,v)$. Positive information transmission, sustained by reputation. The level of information is generally lower than in Region A but strictly positive.

The contemporary AI environment occupies Region B or Region C depending on the institutional context. Section 5 calibrates the three application domains to specific regions.

5. Discussion

The theoretical findings of this paper—a hallucination threshold $c^*(v) > 0$ below which informative equilibrium collapses, a $1/v$ convergence rate for equilibrium information as $c \to 0$, and a reputation-restoration result identifying the minimum reputation strength $\rho^*(c,v)$—are sufficient that they require substantive interpretive engagement. This section discusses three quantitative calibrations, the relationship to the contemporary policy debate, the limitations, and extensions.

5.1 Calibration to AI-generated content moderation

Consider the platform-content moderation problem. The sender is a content creator (human user or AI agent) producing posts on a social-media platform. The receiver is the platform's content-moderation system together with downstream readers. The cost of generating a post using an LLM is approximately \$0.0001 per post (cloud inference cost); the cost of

verifying a post for factual accuracy is approximately \$0.50 (human moderator time, even with AI-assisted triage). The cost ratio $c/v \approx 0.0002$ places the parameter values deeply in the $c \ll c^*(v)$ region.

The reputation parameter ρ in this setting corresponds to the long-term value of platform reputation for individual users. Empirical estimates from platform-economics research suggest $\rho \approx \$0.10$ per truthful post (the marginal value of one additional post viewed by other users when the user's reputation is high). The comparison $\rho/c \approx 1000$ would suggest that reputation should dominate generation cost — but the relevant comparison is ρ vs.\ $\rho^*(c,v) = v \cdot \pi(1-\pi) \approx \$0.125$. The empirical $\rho \approx \$0.10$ falls just below ρ^*, placing the platform-content equilibrium in Region B (babbling-with-noise).

The institutional implication is that platforms whose reputation systems fall just short of ρ^* are theoretically expected to experience collapse of informative communication. Empirical observations of platform misinformation, AI-generated spam, and content-moderation challenges are consistent with this prediction.

5.2 Calibration to scientific peer review under LLM authorship

Consider the scientific peer-review problem. The sender is an author submitting a manuscript; the receiver is the editor and reviewers. The cost of generating manuscript text using an LLM is approximately \$1 per manuscript-equivalent (a 20-page paper at retail LLM API rates). The cost of reviewing a manuscript carefully is approximately \$500–\$1000 per review (the implicit cost of reviewer time, even when nominally unpaid).

The cost ratio $c/v \approx 0.002$ places the peer-review equilibrium in the $c \ll c^*(v)$ region. The reputation parameter ρ corresponds to the long-term reputational value of being a published author of credible work; estimates suggest $\rho \approx \$100$–\$500 per quality publication, depending on the author's career stage and the journal's prestige. The comparison ρ vs.\ $\rho^* \approx v \cdot \pi(1-\pi) \approx \$125$–\$250 is closer in magnitude.

The institutional implication is that peer review for high-prestige journals (where ρ is large) remains in Region C — reputation-restored informative equilibrium — while peer review for low-prestige venues (where ρ is small) approaches Region B. The differential vulnerability of low-prestige venues to LLM-generated manuscript flooding is consistent with this prediction; it has been observed in the recent expansion of predatory journals using AI-generated content.

5.3 Calibration to social-media misinformation

Consider the social-media misinformation problem. The sender is an anonymous account posting political or commercial content; the receiver is the platform's algorithm and downstream readers. The cost of generating content using an LLM is again approximately \$0.0001 per post. The cost of verification by readers is approximately \$0.10 per post (the cognitive cost of reading carefully and assessing accuracy).

The cost ratio $c/v \approx 0.001$ places this equilibrium in the $c \ll c^*(v)$ region. The reputation parameter ρ for anonymous accounts is approximately zero—there is no enduring reputation across content cycles. The comparison ρ vs.\ $\rho^* \approx \$0.025$ places the anonymous-account social-media equilibrium firmly in Region B.

The institutional implication is that anonymous content sources lack the reputation infrastructure to sustain informative equilibrium under the contemporary cost structure. Policy interventions targeting the misinformation problem must therefore either reduce the verification cost v (through better automated fact-checking) or introduce reputation through identity verification, content provenance metadata, or third-party certification systems.

5.4 Relationship to the contemporary policy debate

The contemporary policy debate over AI content labeling, content moderation, and information ecosystem integrity has substantive theoretical content that our framework can address. The proposals for AI-content labeling (e.g., the EU AI Act's transparency requirements) operate by reducing verification cost v — labels permit faster triage by downstream readers and algorithms. Our model predicts that such interventions can move the equilibrium from Region B to Region C if the cost reduction is sufficient.

The proposals for AI-content watermarking and provenance metadata (e.g., the C2PA standard) operate similarly through verification-cost reduction. Our model predicts substantial value of such infrastructure if implemented at scale.

The proposals for reputation aggregation (e.g., platform-level user reputation systems) operate by increasing ρ. Our model predicts that such interventions are most effective when combined with verification-cost reduction, since the joint $\rho \geq \rho^*(c,v)$ condition depends on both parameters.

5.5 Limitations

Five limitations of the model deserve emphasis.

First, the binary state and signal space is a deliberate simplification. The continuous-state version of the model would yield a richer characterization but the qualitative results (hallucination threshold, reputation restoration) are robust to this generalization.

Second, the pure-bias assumption (the sender always wants $a = 1$) corresponds to the extreme case of preference misalignment. Under partial bias, the babbling threshold shifts but the qualitative phase transition remains.

Third, the single-receiver assumption abstracts from receiver heterogeneity. In settings with heterogeneous receivers (e.g., different audience segments on a platform), the equilibrium may be partially informative for some receivers and babbling for others. The full multi-receiver analysis is left for future work.

Fourth, the static one-shot framework abstracts from dynamic considerations. The reputation extension partially restores dynamic structure, but a full repeated-game analysis

with explicit time horizon, discounting, and reputation accumulation is left for future work.

Fifth, the model assumes that verification reveals the truth value of the signal perfectly. In practice, verification is noisy (some signals are ambiguous, some verification efforts fail). The noisy-verification extension is sketched in Section 6.6.

5.6 Extensions

Three extensions of the model deserve specific mention.

Continuous state. Under a continuous state $\theta \in [0,1]$, the equilibrium becomes a partition into countably many intervals (in the informative region) or a single uninformative posterior (in the babbling region). The threshold $c^*(v)$ generalizes to a function of the prior density.

Multi-sender competition. When multiple senders with different types and biases produce signals, the equilibrium features sender-specific reputation accumulation and receiver-specific verification strategies. The competition can either improve or degrade equilibrium information depending on the structure of sender preferences.

Noisy verification. Under noisy verification (each verification reveals the true signal value with probability $\xi < 1$), the receiver's effective information acquisition is reduced by a factor of ξ. The threshold $c^*(v)$ generalizes to $c^*(v,\xi) = v \cdot \pi(1-\pi)/\xi$.

5.7 Connection to the rational-inattention literature

The framework we develop is closely related to but distinct from the rational-inattention framework of Sims (2003) and Caplin and Dean (2015). In the rational-inattention framework, the cost of information acquisition is a function of mutual information; in our framework, the cost is per signal verified. The two frameworks coincide in the limit where verification reveals exactly one bit per signal, but diverge when verification is partial or noisy.

The relationship matters because the contemporary AI-generated information environment exhibits the per-signal cost structure more naturally than the mutual-information cost structure. A reader of LLM-generated text pays cognitive cost per document, not cost per bit of information transmitted. The per-signal verification cost structure is therefore the appropriate primitive for the contemporary application domain.

5.8 Reading the framework against historical episodes

The framework can be applied retrospectively to historical episodes of communication regime change. The introduction of the printing press in the 15th century, the rise of mass-circulation newspapers in the 19th, and the diffusion of broadcast television in the 20th each reduced the marginal cost of signal generation by orders of magnitude. The framework predicts that each transition would have triggered, at least temporarily, a phase transition toward babbling equilibrium that was subsequently mitigated by the development of institutional reputation systems (journals, brand-name newspapers, broadcast networks subject to regulatory oversight).

The contemporary AI episode is, on this reading, the latest in a sequence of communication-cost transitions. The framework predicts a similar institutional response: the development of reputation-bearing institutions whose certification of content quality substitutes for direct verification by every receiver. The contemporary emergence of fact-checking organizations, AI-content labeling standards, and reputation-aggregation platforms is consistent with this prediction.

6. Conclusion

This paper has extended the Crawford–Sobel (1982) framework of strategic information transmission to a setting in which the sender's signal generation is costless and the receiver bears positive verification cost per signal. The motivation is the empirical observation that the diffusion of generative AI since late 2022 has reduced signal-generation costs by approximately three orders of magnitude while leaving verification costs essentially unchanged.

We have proven three central results. First, there exists a hallucination threshold $c^*(v) = v \cdot \pi(1-\pi) > 0$ below which the unique perfect Bayesian equilibrium of the multi-signal communication game is babbling-with-noise: the sender produces a flood of signals with zero net informational content, and the receiver verifies a vanishing fraction. Second, in the babbling-with-noise regime, the equilibrium mutual information $I(\theta;m)$ converges to zero at rate $1/v$ as $c \to 0$, implying that the receiver's expected utility approaches the no-information lower bound. Third, introducing a publicly observed reputation parameter ρ restores positive information transmission when ρ exceeds a critical threshold $\rho^*(c,v)$ that depends on the cost pair.

Calibration to three contemporary application domains—AI-generated content moderation on social media platforms, scientific peer review under LLM authorship, and anonymous social-media misinformation—places the empirical parameter values in the babbling-with-noise region in all three cases. The reputation extension identifies the institutional structures (reputation systems, third-party verification, content provenance metadata) whose absence accounts for the observed challenges in maintaining informative communication.

6.1 What this paper provided

The contribution of the paper is fivefold:

- A formal extension of the Crawford–Sobel cheap-talk framework to admit costly multi-signal generation and costly receiver verification.
- Proof of the hallucination threshold result (Proposition 1): there exists $c^*(v) > 0$ below which informative equilibrium collapses to babbling-with-noise.
- Proof of the convergence rate result (Proposition 2): in the babbling-with-noise regime,

equilibrium mutual information vanishes at rate $1/v$ as $c \to 0$.

- Proof of the reputation restoration result (Proposition 3): a publicly observed reputation parameter $\rho > \rho^*(c, v)$ destabilizes the babbling equilibrium and restores positive information transmission.
- Calibration to three contemporary application domains, identifying the institutional structures whose absence accounts for the observed equilibrium degradation.

6.2 Extensions

Several extensions of the model deserve consideration in subsequent work.

Multi-receiver heterogeneity. The single-receiver assumption abstracts from the heterogeneous information-acquisition behaviors of different audience segments. A multi-receiver extension would characterize the equilibrium under which different audience segments form different beliefs from the same signal flood, with implications for media-market segmentation and the polarization of public opinion.

Dynamic reputation accumulation. The static reputation parameter ρ is a stand-in for the value of long-term reputational capital. A full repeated-game extension with reputation accumulation would endogenize the level of ρ and characterize the conditions under which reputation-restored informative equilibrium is sustainable across time.

Noisy verification. The framework assumes that verification reveals the true signal value with probability one. Real-world verification is noisy; the noisy-verification extension would characterize the equilibrium under verification accuracy $\xi < 1$ and identify the threshold conditions under which the babbling equilibrium is robust.

Strategic verification timing. The framework treats verification as a one-shot decision after observing the signal vector. In practice, receivers often verify signals iteratively, conditional on partial observation. The iterative-verification extension would characterize the optimal verification policy under sequential information arrival.

Information-design intermediaries. The framework treats the sender–receiver dyad as primary. In practice, information intermediaries (editors, content platforms, fact-checking organizations) operate between sender and receiver. The intermediated-communication extension would characterize the optimal design of such intermediaries and identify the welfare gains they can deliver.

Empirical estimation. The calibrations in Section 5 are illustrative. A natural empirical follow-up would be to estimate the parameters (c, v, ρ) in specific institutional contexts and to test the comparative-statics predictions against observed equilibrium outcomes. The post-November-2022 transition provides a natural experiment for such estimation.

6.3 A note on methodological discipline

The theoretical contribution of this paper rests on three primitives: the equilibrium concept (perfect Bayesian equilibrium), the cost structure (per-signal generation and verification

costs), and the timing (sender chooses signal count and content; receiver chooses verification set and action). Each is a deliberate simplification of the contemporary information environment.

The robustness of the qualitative results—a phase transition between informative and babbling equilibria at a threshold $c^*(v) > 0$, restored by reputation $\rho > \rho^*(c, v)$—to alternative primitives is the appropriate test of the theoretical contribution. The numerical magnitudes documented in the calibration are application-specific and depend on the parameter values (c, v, ρ) in each setting; the qualitative pattern of the phase transition is robust to alternative parameterizations.

We close in the spirit of the methodology literature: the theoretical contribution of this paper is most valuable when it disciplines subsequent inquiry rather than when it forecloses it. The framework we develop is a starting point for analyzing the equilibrium implications of the contemporary cost-structure shift in communication, not the final word. The empirical estimation, the multi-receiver extension, the dynamic reputation analysis, and the intermediated-communication framework are all natural follow-ups whose pursuit will refine the theoretical understanding of the contemporary information environment.

References

Daron Acemoglu, Todd Lensman, and Alexander Wolitsky. Adversarial manipulation of online reviews. *NBER Working Paper*, No. 32475, 2024.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017.

Robert J. Aumann and Sergiu Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

Marco Battaglini. Multiple referrals and multidimensional cheap talk. *Econometrica*, 70(4): 1379–1401, 2002.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Strategic communication and persuasion under algorithmic decision-making. *SSRN Working Paper*, No. 5234187, 2025.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

Archishman Chakraborty and Rick Harbaugh. Persuasion by cheap talk. *American Economic Review*, 100(5):2361–2382, 2010.

In-Koo Cho and David M. Kreps. Signaling games and stable equilibria. *Quarterly Journal of Economics*, 102(2):179–221, 1987.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50 (6):1431–1451, 1982.

Mathias Dewatripont and Jean Tirole. Modes of communication. *Journal of Political Economy*, 113(6):1217–1238, 2005.

Joseph Farrell and Robert Gibbons. Cheap talk with two audiences. *American Economic Review*, 79(5):1214–1223, 1989.

Douglas Gale and Martin Hellwig. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies*, 52(4):647–663, 1985.

Matthew Gentzkow and Emir Kamenica. Costly persuasion. *American Economic Review*, 104(5):457–462, 2014.

Maria Goltsman and Johannes Hörner. Mediation, arbitration and negotiation. *Journal of Economic Theory*, 144(4):1397–1420, 2009.

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

Emir Kamenica and Matthew Gentzkow. Disclosure of endogenous information. *Economic Theory Bulletin*, 5(1):47–56, 2017.

Stefan Krasa and Anne P. Villamil. Optimal contracts when enforcement is a decision variable. *Econometrica*, 68(1):119–134, 2000.

David M. Kreps and Robert Wilson. Reputation and imperfect information. *Journal of Economic Theory*, 27(2):253–279, 1982.

Vijay Krishna and John Morgan. A model of expertise. *Quarterly Journal of Economics*, 116(2):747–775, 2001.

Elliot Lipnowski and Doron Ravid. Cheap talk with transparent motives. *Econometrica*, 88 (4):1631–1660, 2020.

Bartosz Maćkowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, 2009.

Bartosz Maćkowiak, Filip Matějka, and Mirko Wiederholt. Rational inattention: A disciplined behavioral model. *Journal of Economic Literature*, 61(1):226–273, 2023.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

Steven A. Matthews. Veto threats: Rhetoric in a bargaining game. *Quarterly Journal of Economics*, 104(2):347–369, 1989.

Dilip Mookherjee and Stefan Reichelstein. Implementation via augmented revelation mechanisms. *Review of Economic Studies*, 57(3):453–475, 1990.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

Michael Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374, 1973.

Robert M. Townsend. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2):265–293, 1979.