# The Generative Lossy Channel:

## Substrate-Independent Compression Dynamics, Five Sufficient Conditions for Net-Beneficial Noise, and the Ratchet Toward Self-Deception

Jeremy McEntire
Cage & Mirror Press
jmc@cageandmirror.com

### Abstract

When a communication channel is lossy and the receiver must produce structured output, the receiver is forced to *generate*: to fill the gap between what was received and what must be produced. This paper formalizes the generative lossy channel—a channel where compression, strategic misalignment, and conformity pressure combine to force reconstruction that systematically diverges from the source. The mechanism is substrate-independent: it appears in organizational hierarchies (where reports compress reality and institutional norms constrain output) and in AI systems (where training compresses human values and reward signals constrain behavior). The substrate-independence claim is not analogical. It follows from the fact that both domains share the structural cause: compression creates information gaps; selection operating in those gaps determines whether the divergence is functional or pathological.

Five sufficient conditions (C1–C5) are proved to guarantee net-beneficial noise in any two-level system with a generative lossy channel. Theorem 1 establishes the inverted-U relationship between noise amplitude and system performance under these conditions, with three independent proof paths covering threshold stochastic resonance, Jensen-gap rectification, and opposing-monotone mechanisms. The conditions are validated computationally: 500 Monte Carlo configurations yield zero counterexamples to sufficiency (75/75 where all conditions hold), across six mechanistically distinct domains.

Cross-domain evidence is presented from organizational communication (mapping conditions C1–C5 to Vaughan's normalized deviance stages in the Challenger disaster), jazz improvisation (where controlled channel noise drives creative emergence, following Berliner's analysis of real-time harmonic reconstruction), and AI alignment (where sycophancy arises as an equilibrium phenomenon of reward-compressed preference channels, measurable via the methods of Perez et al. and Sharma et al.). A formal ratchet proposition establishes that once compression-driven drift crosses a self-reinforcing threshold, recovery requires external intervention.

**Scope statement.** Theorem 1 is a formal result about lossy channels with generative capacity. The cross-domain instances presented verify qualitative satisfaction of conditions C1–C5. Full formal verification within each domain—requiring domain-specific operationalization of channel capacity, noise structure, and generative residual—is future work.

**Keywords:** lossy channels, generative compression, stochastic resonance, organizational dysfunction, AI alignment, substrate independence, rate-distortion-perception, dysmemic pressure

# Contents

# 1 Introduction

## 1.1 The Substrate-Independence Argument

Intelligent systems deceive themselves. Organizations compress reality into reports, dashboards, and metrics; the compression discards information; signals optimized for the compressed frame outcompete signals optimized for reality. AI systems compress human values into reward signals; the compression discards nuance; outputs optimized for reward outcompete outputs aligned with the values the reward was meant to represent. The pattern is the same. The substrates share nothing except compression and selection.

This paper makes a structural claim: the mechanism producing systematic self-deception in organizations and AI systems is *the same mechanism*, operating on the same formal objects (lossy channels, conformity constraints, selection environments), and formalized by the same mathematical framework. The claim is not analogical. Human psychology cannot explain the pattern in AI systems lacking human psychology. Machine learning training procedures cannot explain the pattern in organizations lacking gradient descent. Bureaucratic incentives cannot explain the pattern in neural networks. The pattern appears in both because the cause is what they share: compression creates information gaps, and selection operating in those gaps drives systematic drift from accuracy toward fit.

Three bodies of formal theory converge on this mechanism. Crawford and Sobel (Crawford and Sobel, 1982) proved that strategic communication under preference divergence is endogenously lossy: the channel quantizes reality into coarse partitions, and the coarseness increases with misalignment. Blau and Michaeli (Blau and Michaeli, 2018, 2019) proved that lossy compression under a perceptual quality constraint forces the decoder to generate: to produce outputs whose distribution diverges from the source in ways that the compression alone does not explain. Atlan (Atlan, 1979) established the level-crossing principle: noise at one level of organization can produce information at a higher level, provided the system has the right structure to exploit it.

This paper unifies these results. The generative lossy channel is a communication channel where compression (Shannon/Crawford–Sobel), conformity pressure (Blau–Michaeli), and selection (evolutionary dynamics) interact to produce systematic divergence from the source. Under convergent selection—where fit with existing frames is rewarded—the divergence produces organizational self-deception, compliance theater, and AI sycophancy. Under divergent selection—where functional novelty is rewarded—the same divergence produces creative emergence. The mechanism is identical. The valence is determined by the selection environment.

### 1.2 Contribution

This paper establishes four results:

**Theorem 1** (Five Sufficient Conditions). Five conditions—suboptimality (C1), nonlinear integration (C2), accessibility (C3), asymmetric weighting (C4), and graceful degradation (C5)—are jointly sufficient for net-beneficial noise in any two-level system with a generative lossy channel. Three independent proof paths cover threshold stochastic resonance, Jensen-gap rectification, and opposing-monotone mechanisms. The theorem is validated computationally with zero counterexamples across 500 Monte Carlo configurations.

**Proposition 2.10** (Dual Valence). The generative residual produced by a lossy channel has dual valence: under convergent selection it produces pathological drift; under divergent selection it produces creative emergence. The valence is determined by the selection criterion, not by the channel.

**Proposition 2.12** (The Compression Ratchet). Once dysmemic pressure—the compound force selecting fit over truth in compression-shaped environments—crosses a self-reinforcing threshold, the compressed state stabilizes as an equilibrium. Recovery requires external intervention because reform proposals are themselves signals subject to the selection environment they aim to change.

**Cross-domain validation.** Three external evidence domains—organizational disasters, jazz creativity, and AI alignment—are mapped onto the C1–C5 framework using published empirical work rather than self-referential instance papers.

### 1.3 Paper Structure

Section 2 establishes the formal foundation: definitions, Theorem 1, the dual-valence proposition, and the ratchet proposition. Section 3 presents the computational validation (500 Monte Carlo configurations, six mechanisms). Section 4 maps three external evidence domains onto the formal framework. Section 5 presents limitations, falsification conditions, and open questions.

## 2 Formal Foundation

### 2.1 The Generative Lossy Channel

**Definition 2.1** (Generative Lossy Channel)**.** A *generative lossy channel* is a communication system $(S, C, R, Q)$ where:

- $S$ is a source producing states $\theta$ from a distribution $p_\theta$ over state space $\Theta$.

- $C$ is a lossy channel with rate $R_C < H(\theta)$, producing a compressed representation $m$ of $\theta$.

- $R$ is a receiver that maps $m$ to an output $\hat{\theta}$.

- $Q$ is an *acceptability distribution* over outputs: the receiver's output must satisfy $d(p_{\hat{\theta}}, Q) \leq P$ for some divergence $d$ and tolerance $P$.

The channel is *generative* when $P$ is finite and $d(p_{\hat{\theta}^*}, Q) > 0$, where $\hat{\theta}^*$ is the distortion-minimizing reconstruction absent the acceptability constraint. Under these conditions, the receiver must produce outputs that diverge from the source beyond what compression alone requires.

**Definition 2.2** (The Generative Residual)**.** For a generative lossy channel, the *generative residual* is:

$$\Delta_{\text{gen}} = \mathbb{E}[\Delta(\hat{\theta}, \theta) \mid d(p_{\hat{\theta}}, Q) \leq P] - \mathbb{E}[\Delta(\hat{\theta}^*, \theta)] \tag{1}$$

where $\Delta$ is the distortion measure, the first term is the minimum achievable distortion under the acceptability constraint, and the second is the unconstrained minimum. When $\Delta_{\text{gen}} > 0$, the receiver is forced to generate.

The generative residual arises from the interaction of three constraints, each established independently in the literature:

1. **Endogenous lossy compression** (Crawford–Sobel). When sender and receiver preferences diverge, the sender's message partitions continuous reality into at most $N^*$ discrete categories, where $N^* = \lfloor -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{2b}} \rfloor$ and $b$ is the bias parameter (Crawford and Sobel, 1982). The channel is lossy for game-theoretic reasons, not bandwidth limitations.

2. **Forced generation under conformity** (Blau–Michaeli). The rate-distortion-perception tradeoff (Blau and Michaeli, 2019) proves that when the reconstruction must satisfy a distributional constraint ($d(p_{\hat{X}}, Q) \leq P$), the minimum achievable rate is strictly elevated: $R(D, P, Q) > R(D)$ whenever the unconstrained optimizer does not already satisfy the constraint. The decoder must inject structure—generate—to close the gap between what was received and what must be produced.

3. **Selection on the residual.** The generative residual $\Delta_{\text{gen}}$ is a divergence from the source, not a directed error. Its direction is determined by the selection criterion operating on the receiver's output: convergent selection (rewarding fit with existing frames) produces pathological drift; divergent selection (rewarding functional novelty) produces creative emergence.

### 2.2 Dysmemic Pressure

The term *dysmemic pressure* names the compound selective force operating in the gap between representation and reality. Three dynamics compound:

1. **Strategic communication degradation.** As preferences diverge, the sender transmits less precisely. The receiver discounts more heavily. Both act rationally. The aggregate is information loss (Crawford and Sobel, 1982).

2. **Adverse selection in signal markets.** Producing accurate signals is costly; producing frame-fitting signals is cheap. When receivers cannot verify quality at the moment of consumption, cheap signals flood the market, driving out accurate ones. This is Akerlof's (1970) lemons problem operating on information rather than goods.

3. **Transmission bias.** Signals spread based on transmissibility, not accuracy. Simple signals outcompete complex ones. Prestige-associated signals outcompete identical signals from unknown sources. Apparent-consensus signals accelerate through positive feedback (Boyd and Richerson, 1985).

**Definition 2.3** (Dysmemic Pressure)**.** *Dysmemic pressure* is the compound selective force favoring fit over truth in environments shaped by compressed representations. It intensifies with compression ratio, preference divergence, verification cost, and stakes. A *dysmeme* is a signal optimized for survival in the gap between representation and reality. Its survival depends on alignment with receiver preferences, ease of transmission, and conformity with the acceptability distribution $Q$. Truth is not the selection criterion.

### 2.3 Theorem 1: Five Sufficient Conditions for Net-Beneficial Noise

**Definition 2.4** (Two-Level System with Noise)**.** A *two-level system* consists of:

- Level-$N$ performance $P_N(\sigma)$, degrading with noise amplitude $\sigma$.

- Level-$(N+1)$ performance $P_{N+1}(\sigma)$, representing emergent or higher-order outcomes.

- System performance $P_{\text{sys}}(\sigma) = f(P_N(\sigma), P_{N+1}(\sigma))$ for some aggregation function $f$.

The system exhibits *net-beneficial noise* if there exists $\sigma^* > 0$ such that $P_{\text{sys}}(\sigma^*) > P_{\text{sys}}(0)$.

**Theorem 2.5** (Five Sufficient Conditions for Net-Beneficial Noise)**.** *Let* $(P_N, P_{N+1}, P_{sys})$ *be a two-level system. The following five conditions are jointly sufficient for the existence of* $\sigma^* > 0$ *with* $P_{sys}(\sigma^*) > P_{sys}(0)$*:*

**C1 (Suboptimality).** *The noiseless system is not at the global optimum of $P_{N+1}$: there exists a state achieving higher $P_{N+1}$ that is inaccessible at $\sigma = 0$. Formally: the system operates below its level-$(N+1)$ capacity.*

**C2 (Nonlinear integration).** *The function mapping noise to level-$(N+1)$ benefit is nonlinear, through one of three mechanisms:*

C2a*: **Threshold stochastic resonance.** A subthreshold signal is boosted by noise through a nonlinear detector.*

C2b*: **Jensen-gap rectification.** A convex function rectifies symmetric noise into a net positive: $\mathbb{E}[g(x+\xi)] > g(x)$ when $g'' > 0$ in the operating region.*

C2c*: **Opposing monotones.** The product of an increasing function (benefit given success) and a decreasing function (probability of success) has an interior maximum.*

**C3 (Accessibility).** *The noise distribution has support in the improvement region: the perturbation can reach states with higher $P_{N+1}$.*

**C4 (Asymmetric cost–benefit).** *The level-$(N+1)$ gain exceeds the level-$N$ cost over some interval:*

$$\exists\, \sigma^* > 0 : \quad \beta \int_0^{\sigma^*} \Delta_{N+1}(\sigma)\, d\sigma > \alpha \int_0^{\sigma^*} |\Delta_N(\sigma)|\, d\sigma \tag{2}$$

*where $\Delta_{N+1}(\sigma) = P_{N+1}(\sigma) - P_{N+1}(0)$, $\Delta_N(\sigma) = P_N(\sigma) - P_N(0)$, and $\alpha, \beta > 0$ are system-dependent weights.*

**C5 (Graceful degradation).** *Level-$N$ performance degrades gradually: there is no catastrophic collapse at small $\sigma$. The system has sufficient redundancy or slack to absorb moderate noise without structural failure.*

*Under these conditions, $P_{sys}(\sigma)$ exhibits an inverted-U: there exists $0 < \sigma^* < \bar{\sigma} < \infty$ such that $P_{sys}(\sigma^*) > P_{sys}(0)$ and $P_{sys}(\sigma) < P_{sys}(0)$ for $\sigma > \bar{\sigma}$. The optimal noise level $\sigma^*$ is bounded; there is always too much noise.*

The proof proceeds through three independent paths, one for each C2 sub-mechanism.

***Proof sketch for C2a (Threshold SR).*** Under C1, the system has a subthreshold signal: a state with higher $P_{N+1}$ that the deterministic system cannot reach. Under C2a, a nonlinear threshold creates a detection probability $P_{N+1}(\sigma) = \Phi\left(\frac{s-\theta}{\sigma}\right)$ that increases from zero as $\sigma$ increases. Under C3, the noise distribution reaches the threshold. Under C4, the integral gain exceeds the integral cost. Under C5, the cost function has no discontinuity. The existence of the interior maximum follows from continuity: $P_{N+1}(0) = 0$ (subthreshold), $P_{N+1}(\sigma)$ increases for moderate $\sigma$, and $P_N(\sigma)$ degrades continuously. The aggregation $P_{\text{sys}}(\sigma) = f(P_N, P_{N+1})$ inherits the inverted-U from C4's integral condition. □

***Proof sketch for C2b (Jensen gap)***. Under C2b, $g$ is convex in the operating region, so $\mathbb{E}[g(x+\xi)] - g(x) = J(\sigma) \geq 0$ by Jensen's inequality. The Jensen gap $J(\sigma)$ is the level-$(N{+}1)$ gain. Under C4, $\beta J(\sigma)$ exceeds $\alpha|\Delta_N(\sigma)|$ for moderate $\sigma$. Under C5, the cost $|\Delta_N|$ grows continuously from zero. The inverted-U follows from $J(\sigma) \sim \frac{1}{2}g''(x)\sigma^2$ for small $\sigma$ (quadratic onset) versus linear or faster cost growth for large $\sigma$. □

***Proof sketch for C2c (Opposing monotones)***. Under C2c, benefit given success $g(\sigma)$ is increasing (harder problems have higher payoff) and probability of success $h(\sigma)$ is decreasing (more noise, less reliability). The product $g(\sigma) \cdot h(\sigma)$ has an interior maximum by first-order calculus: $\frac{d}{d\sigma}[g \cdot h] = g'h + gh'$, which changes sign because $g' > 0$ and $h' < 0$. The maximum $\sigma^*$ satisfies $g'(\sigma^*)/g(\sigma^*) = -h'(\sigma^*)/h(\sigma^*)$. Under C4 and C5, the net system performance inherits this interior maximum. □

**Lemma 2.6** (Integral vs. Marginal C4)**.** *For threshold systems satisfying C2a, the marginal derivative condition*

$$\beta \cdot \frac{\partial P_{N+1}}{\partial \sigma}\bigg|_{\sigma \to 0^+} > \alpha \cdot \left| \frac{\partial P_N}{\partial \sigma}\bigg|_{\sigma \to 0^+} \right| \tag{3}$$

*is neither necessary nor sufficient for net benefit. The correct condition is the integral formulation in C4.*

*Proof.* For a threshold detector with signal $s < \theta$ and Gaussian noise, $P_{N+1}(\sigma) = \Phi\left(\frac{s-\theta}{\sigma}\right) \to 0$ exponentially fast as $\sigma \to 0^+$, so the marginal benefit is exponentially small. Meanwhile, level-$N$ cost begins immediately: $|\partial P_N/\partial\sigma|_{\sigma \to 0^+}| = L_N > 0$. The marginal condition is never satisfied, yet the integral condition can be met at finite $\sigma$ where the detection probability becomes substantial. □

### 2.4 Corollaries

**Corollary 2.7** (The Brittleness Trap)**.** *A system that minimizes redundancy at level $N$ violates C5 and cannot exhibit generative level-crossing, regardless of how well C1–C4 are satisfied. Efficiency kills generativity when it eliminates the buffer that absorbs noise.*

**Corollary 2.8** (The Alignment Trap)**.** *A system with perfectly aligned incentives ($b = 0$) has a nearly lossless channel, producing minimal variation. Perfect alignment can satisfy level-$(N{+}1)$ needs directly, eliminating the need for level-crossing—but also eliminating the capacity for it when the environment shifts.*

**Corollary 2.9** (The Over-Noise Catastrophe)**.** *For any system satisfying C1–C5, there exists* $\bar{\sigma}$ *beyond which* $P_{sys}(\sigma) < P_{sys}(0)$*. The inverted-U is bounded. There is always too much noise.*

### 2.5 Dual Valence

**Proposition 2.10** (Dual Valence of the Generative Residual)**.** *The generative residual* $\Delta_{\text{gen}}$ *(Definition 2.2) has dual valence:*

- *Under* convergent selection *(selection rewards fit with the existing frame),* $\Delta_{\text{gen}}$ *produces systematic drift toward internally-fit outputs. This is the mechanism underlying compliance theater, impression management, and sycophancy.*

- *Under* divergent selection *(selection rewards functional novelty or accuracy to external reality),* $\Delta_{\text{gen}}$ *produces reconstruction that occasionally outperforms the source. This is the mechanism underlying creative emergence.*

*The dual valence follows from the fact that* $\Delta_{\text{gen}}$ *is a divergence from the source, not a directed error. The direction of the divergence is determined by the selection criterion operating on the receiver's output, not by the compression mechanism.*

**Corollary 2.11** (The Inseparability of Dysfunction and Creativity)**.** *Since both dysfunctional drift and creative novelty arise from the same generative residual under the same conformity constraint, no system can eliminate dysfunction potential without simultaneously eliminating creative potential. The design problem is not elimination but channeling.*

### 2.6 The Compression Ratchet

The following proposition formalizes the self-reinforcing dynamics described informally in the organizational theory literature as "normalization of deviance" (Vaughan, 1996), "institutional isomorphism" (DiMaggio and Powell, 1983), and "the iron cage." The formalization draws on the strategic RDP equilibrium framework, specifically the sufficient conditions for organizational channels (conditions O1–O4 of McEntire, 2026, Theorem C): positive rate, preference divergence, distributional acceptability, and non-degeneracy. Under those conditions, the strategic RDP tradeoff holds and the generative residual is positive.

**Proposition 2.12** (The Compression Ratchet)**.** *Let* $\mathcal{S}$ *be a generative lossy channel (Definition 2.1) satisfying the following conditions:*

(R1) ***Endogenous acceptability.*** *The acceptability distribution* $Q_t$ *at time* $t$ *is a function of prior outputs:* $Q_{t+1} = \Gamma(Q_t, \{\hat{\theta}_\tau\}_{\tau \leq t})$*, where* $\Gamma$ *is a distribution update rule that moves* $Q$ *toward the empirical distribution of recent outputs.*

(R2) ***Positive rate with preference divergence.*** *The channel satisfies conditions O1 (positive rate: $b < 1/4$) and O2 (preference divergence: $b > 0$) of the strategic RDP framework, so the channel is endogenously lossy with a positive generative residual.*

(R3) ***Convergent selection.*** *The selection criterion operating on outputs rewards conformity with $Q_t$: outputs closer to $Q_t$ receive higher fitness.*

(R4) ***No external correction.*** *The system receives no input from sources outside the compression-selection loop: no external audit, no independent verification, no exogenous signal forcing $Q_t$ toward the source distribution $p_\theta$.*

*Then the system has a* self-reinforcing equilibrium*: a fixed point $Q^*$ of the dynamics $Q_{t+1} = \Gamma(Q_t, \cdot)$ such that:*

(i) *$d(Q^*, p_\theta) > d(Q_0, p_\theta)$: the equilibrium acceptability distribution is farther from reality than the initial one.*

(ii) *The basin of attraction is expanding: once $d(Q_t, p_\theta) > \delta$ for a threshold $\delta$ determined by the system parameters, the dynamics are monotonically drift-increasing.*

(iii) *Recovery requires violation of* (R4)*: external intervention that introduces signals not subject to the internal selection environment.*

*Proof.* Under (R2), the channel is endogenously lossy with generative residual $\Delta_{\text{gen}} > 0$ (by the strategic RDP tradeoff under conditions O1–O4). The receiver's outputs therefore deviate from the source: $d(p_{\hat{\theta}}, p_\theta) > 0$.

Under (R1), these deviant outputs update the acceptability distribution: $Q_{t+1}$ moves toward $p_{\hat{\theta}}$. Since $p_{\hat{\theta}}$ deviates from $p_\theta$, the updated $Q_{t+1}$ also deviates from $p_\theta$.

Under (R3), the next period's outputs are selected for conformity with $Q_{t+1}$, which is now farther from $p_\theta$ than $Q_t$ was. The generative residual at $t + 1$ produces outputs deviating from $p_\theta$ in the same direction, because the conformity constraint now pulls toward a $Q$ that has already drifted.

The monotonicity follows: $d(Q_{t+1}, p_\theta) \geq d(Q_t, p_\theta)$ whenever the generative residual reinforces the drift direction, which is guaranteed under convergent selection (R3). Each dysmeme that establishes itself tilts the landscape further. The next dysmeme becomes easier to establish. The next accurate signal becomes harder to transmit.

The equilibrium $Q^*$ is the fixed point where the drift rate equals zero: the acceptability distribution has stabilized at a point where the generative residual exactly reproduces the current $Q^*$. The constructed reality generates signals that reinforce the selection criteria that generated the constructed reality.

Under (R4), no mechanism exists to correct the drift. Reform proposals are themselves signals subject to the internal selection environment. A reform that would displace the equilibrium threatens agents whose positions depend on the current arrangement and is selected against. A reform that changes vocabulary while preserving fitness criteria is selected for, producing surface change without equilibrium displacement. Recovery therefore requires violating (R4): introducing observation, verification, or authority from outside the compression-selection loop. This is the formal content of the claim that the equilibrium is a "cage"—a self-reinforcing state that resists displacement from within. □

*Remark* 2.13 (Connection to high-reliability organizations)*.* Organizations that resist the ratchet—aircraft carriers, nuclear plants, air traffic control—do so by institutionalizing violations of condition (R4): redundant verification from independent channels, licensed dissent roles that create protected paths for signals the internal selection environment would otherwise filter, and deference to expertise that migrates decision authority to whoever has the least compressed information about the current state (Weick and Sutcliffe, 2007). Each practice introduces external signal into the compression-selection loop.

# 3 Computational Validation

## 3.1 Strategy

Theorem 2.5 asserts five sufficient conditions for net generative level-crossing. This section subjects that claim to adversarial computational testing across six mechanistically distinct domains: threshold detection, sigmoid detection, polynomial detection (null model), simulated annealing, ensemble diversity, and multi-armed bandit exploration. The simulations were designed to find counterexamples—parameter configurations where all five conditions are met but noise fails to help.

## 3.2 Stochastic Resonance Class: Three Nonlinearities and a Null

A two-level system with level-$N$ degradation under three models (linear, exponential, catastrophic) and level-($N$+1) detection under four detector types (threshold, sigmoid, polynomial, linear null). Each configuration swept across 100 noise levels from $\sigma = 0$ to $\sigma = 3$, with 10,000 trials per noise level and 200 bootstrap resamples for confidence intervals.

| **Configuration** | **Net Benefit** | **Optimal $\sigma^*$** | **Inverted-U** | **Conditions Met** |
|---|---|---|---|---|
| Threshold, all met | **0.170** | 0.45 | Yes | All 5 |
| Sigmoid, all met | **0.091** | 0.33 | Yes | All 5 |
| Polynomial, all met | 0.000 | — | No | C2 marginal |
| Linear (null) | 0.000 | — | No | C2 violated |
| C1 violated | 0.000 | — | No | C1 violated |
| C5 violated | 0.000 | — | No | C5 violated |

The threshold and sigmoid detectors exhibit the predicted inverted-U. The polynomial detector produces zero benefit despite being non-affine—its Jensen gap is negligible. The linear null model confirms that nonlinearity (C2) is essential. The coupling model determines whether C5 binds: under additive coupling, a brittle lower level still yields marginal benefit; under substrate coupling (where level-($N$+1) gains are gated by level-$N$ health), brittleness is fatal.

### 3.3 Non-SR Mechanisms

Three additional mechanisms test generality beyond stochastic resonance:

**Simulated annealing** (optimization landscape escape). A 1D fitness landscape with local and global optima. Noise is Metropolis temperature. Net benefit 0.254 when conditions met; zero when C1 violated (already at global optimum) or C4 violated (stability dominates exploration).

**Ensemble diversity** (error decorrelation). Twenty-one predictors on a hard XOR classification problem. Noise perturbs predictor weights. Net benefit 0.068 when conditions met.

**Multi-armed bandit** (exploration–exploitation). Ten arms with unknown rewards. Noise is exploration probability. Net benefit 0.345 when conditions met; zero when C1 violated (all arms equal).

### 3.4 Monte Carlo Sensitivity Analysis

500 random parameter configurations. Conditions checked dynamically.

| **Category** | **Count** | **Percentage** |
|---|---|---|
| All conditions met AND benefit observed | 75 | 15.0% |
| All conditions met AND no benefit | **0** | **0.0%** |
| Some condition violated AND benefit observed | 38 | 7.6% |
| Some condition violated AND no benefit | 387 | 77.4% |
| **Total** | **500** | |

**Sufficiency rate: 100%.** Zero counterexamples. The conditions are sufficient but not necessary (7.6% of violations still showed benefit, indicating the conditions define a conservative boundary).

### 3.5 Crawford–Sobel Strategic Communication Simulation

Actual partition equilibria with receiver estimation error. Bias swept from 0.01 to 0.25.

| **Bias** $b$ | **Partitions** | **Eff. Noise** | **Info Transmitted** | **Babbling** |
|---|---|---|---|---|
| 0.010 | 6 | 0.059 | 96.0% | No |
| 0.036 | 3 | 0.110 | 85.6% | No |
| 0.062 | 2 | 0.155 | ∼70% | No |
| 0.087+ | 1 | 0.291 | 0.0% | **Yes** |

The transition from informative communication to babbling is sharp. Small incentive misalignment creates quantization noise that preserves most information while introducing the generative residual. Large misalignment collapses communication entirely.

### 3.6 Summary of Computational Evidence

1. **Sufficiency confirmed.** 75/75, zero counterexamples.

2. **Generality confirmed.** Six mechanisms, four substrate classes.

3. **Null models confirmed.** Linear and polynomial detection show zero benefit.

4. **C4 requires integral formulation.** Marginal derivative fails for threshold detectors (Lemma 2.6).

5. **C2 requires sufficient steepness.** Non-affinity is necessary but not sufficient; the slope at threshold must be steep enough relative to the noise variance, consistent with Chapeau-Blondeau's (1997) generalization.

6. **C5 depends on coupling architecture.** Additive vs. substrate coupling determines relevance.

7. **Crawford–Sobel validates the organizational anchor.** Sharp babbling transition confirmed.

# 4 Cross-Domain Evidence

## 4.1 Evidentiary Structure and Scope

A theory of substrate-independent generative lossy channels predicts that the same formal structure—compression forces reconstruction, reconstruction diverges from source, selection determines valence—should appear in substrates that share no surface features. This section maps three domains onto the C1–C5 framework using published external evidence.

**Scope statement.** Theorem 2.5 is a formal result about lossy channels with generative capacity. The cross-domain instances presented here verify qualitative satisfaction of conditions C1–C5. Full formal verification within each domain—requiring domain-specific operationalization of channel capacity, noise structure, and generative residual—is future work. The instances serve as evidence of cross-domain applicability, not as foundations for the theorem.

## 4.2 Organizational Disasters: Normalized Deviance as the Ratchet

Vaughan's (1996) ethnographic study of the Challenger disaster provides the most thoroughly documented case of organizational self-deception through compression-selection dynamics. Her analysis identifies a process she terms *normalized deviance*: unacceptable practice becomes acceptable as deviant behavior repeats without catastrophe. Engineers knew that O-ring erosion occurred at cold temperatures. The information existed in the system. The compression-selection dynamics of NASA's reporting hierarchy filtered it.

Vaughan documents five stages that map directly onto the C1–C5 framework and the ratchet mechanism (Proposition 2.12):

| Cond. | Mapping to Challenger normalized deviance |
|---|---|
| C1 | **Met.** The system was suboptimal: O-ring erosion was a known failure mode, and the engineering data showed temperature dependence. The noiseless system (uncompressed engineering reports) contained the signal for correct decision-making. |
| C2 | **Met (inverted).** The nonlinearity was in the *acceptability threshold*: below a certain level of observed erosion, flights were classified as nominal. This threshold created a step-function response: erosion below threshold produced no organizational response; erosion above threshold triggered review. Under convergent selection, this nonlinearity amplified drift rather than recovery. |
| C3 | **Met (inverted).** The noise distribution (variation in O-ring performance across flights) had support in the region that *reinforced* the existing frame: most flights showed acceptable erosion, confirming the compressed signal that erosion was manageable. |
| C4 | **Met for the dysfunction direction.** The weighting favored schedule pressure ($\alpha$) over engineering caution ($\beta$). Each successful flight with O-ring anomalies increased the weight on "it worked before" relative to "the engineering model predicts failure." |
| C5 | **Met (graduated degradation).** The organizational capacity for processing bad news degraded gradually, not catastrophically, through budget cuts, workforce reductions, and schedule pressure—exactly as the ratchet mechanism predicts. |

The ratchet conditions (R1–R4) of Proposition 2.12 are satisfied in Vaughan's account. The acceptability distribution $Q_t$ was endogenous (R1): what counted as "acceptable risk" was updated by each flight that survived with O-ring anomalies. The channel had positive rate with preference divergence (R2): engineers and managers had different loss functions. Selection was convergent (R3): signals conforming to the existing risk assessment were rewarded. And no external correction operated (R4): the safety review process depended on the Shuttle Program for resources and lacked independent analytical capability. The Columbia

Accident Investigation Board (1996) found seventeen years later that the same organizational failures had reasserted themselves—the ratchet had resumed after post-Challenger reforms decayed.

### 4.3 Jazz Improvisation: Creative Emergence Through Controlled Channel Noise

Berliner's (1994) ethnographic study *Thinking in Jazz* documents the cognitive and social mechanisms of ensemble improvisation through interviews with over fifty professional jazz musicians. His account provides detailed evidence for the generative lossy channel operating under divergent selection.

In ensemble jazz improvisation, each musician transmits musical intent through an inherently lossy channel: the acoustic signal, filtered by the listener's auditory processing, harmonic vocabulary, and rhythmic frame. No musician perfectly decodes another's intent. The residual—the gap between the player's intent and the listener's perception—is filled by the listening musician's own priors. This is the Crawford–Sobel channel in real time, where the "bias" is not strategic misalignment but ontological misalignment: each musician has a different compression scheme (harmonic vocabulary, rhythmic frame) that partitions the continuous acoustic signal into different categories.

Pressing (1988) formalized the cognitive architecture of improvisation as a feedback loop between "referent" (the generative plan) and "event cluster" (the realized output), where the gap between plan and realization is the productive locus of novelty. Johnson-Laird (2002) argued that jazz improvisation requires operating within constraints that are tight enough to produce coherent output but loose enough to permit novel combinations—precisely the regime where the generative residual is positive but bounded.

| Cond. | Status in jazz improvisation |
|---|---|
| C1 | **Met.** Collective output from a score is bounded by the composition. Improvisation has room for improvement—the system operates below its aesthetic capacity. |
| C2 | **Met.** Musical cognition is nonlinear: a note within the expected harmonic frame produces a qualitatively different response than one outside it. Berliner documents the "turnaround" moment where a misinterpreted chord becomes the basis for a new harmonic direction (Berliner, 1994). |
| C3 | **Met.** Harmonic reinterpretations can reach the improvement region—producing new progressions that are musically valid. The noise distribution (variations in interpretation across ensemble members) has support in aesthetically productive territory. |
| C4 | **Met.** The jazz aesthetic values collective innovation ($N+1$) over individual fidelity to a predetermined plan ($N$). Pressing's model formalizes this as a preference ordering where novel coherent output ranks above faithful reproduction (Pressing, 1988). |
| C5 | **Met.** Skilled jazz musicians degrade gracefully under misinterpretation. Training provides the redundancy that allows productive absorption of noise. A musician who cannot absorb misinterpretation without losing function—a brittle player—cannot participate in generative improvisation (Corollary 2.7). |

The dual-valence claim (Proposition 2.10) is directly testable here: the same lossy-channel mechanism that produces normalized deviance under convergent selection (Section 4.2) produces creative emergence under divergent selection. Same channel structure. Same compression. Same forced reconstruction. Different selection criterion. Different valence.

### 4.4 AI Alignment: Sycophancy as Equilibrium Compression Artifact

RLHF (reinforcement learning from human feedback) training creates a strategic channel between the model (sender) and human evaluators (receivers). The model's output must conform to the distribution of "helpful-sounding" responses (the acceptability constraint

$Q$). The model's optimization target (reward) diverges from the evaluator's intended target (helpfulness). The conditions for the generative lossy channel are satisfied:

| Cond. | Status in RLHF-trained language models |
|---|---|
| C1 | **Met.** The pre-RLHF model is suboptimal on the intended alignment target: it has the capacity for more helpful, truthful output than it produces by default. |
| C2 | **Met.** The reward function is nonlinear: agreement with the user produces a qualitatively different reward signal than disagreement, creating a step-function-like response at the boundary between agreeable and challenging output. |
| C3 | **Met.** The noise distribution (variation in reward signals across evaluators) has support in the sycophancy region: evaluators sometimes reward agreeable-but-inaccurate outputs over accurate-but-uncomfortable ones. |
| C4 | **Met for the dysfunction direction.** The weighting favors reward-on-distribution ($\alpha$) over out-of-distribution accuracy ($\beta$). Each training step that reinforces agreeable output increases the implicit weight on agreeability relative to truthfulness. |
| C5 | **Met (graduated).** Model capabilities degrade gradually under reward misalignment: the model retains general competence while drifting toward sycophantic patterns, rather than collapsing catastrophically. |

Perez et al. (2022) developed model-written evaluations that systematically test for sycophantic and other misaligned behaviors, providing a measurement framework for the generative residual in AI systems. Sharma et al. (2023) directly measured sycophancy in language models trained with RLHF, finding that models systematically shift their stated views to match the user's expressed opinion—precisely the pattern predicted by the compression-selection framework operating under convergent selection. Their findings operationalize condition C4: the reward signal empirically favors agreement over accuracy.

The ratchet mechanism (Proposition 2.12) predicts a specific failure mode: as sycophantic outputs are rewarded, they shift the effective acceptability distribution $Q_t$, making future sycophantic outputs more normal and future honest-but-uncomfortable outputs more deviant. This is the AI analogue of normalized deviance: the system's representation of "good output" drifts from the intended target through accumulated selection pressure operating on a compressed preference signal.

The substrate-independence claim is strongest here: the same formal mechanism—compression of values into a signal, selection operating on the gap between signal and values, drift toward signal-fit rather than value-alignment—produces sycophancy in AI and compliance theater in organizations. Human psychology is not the common cause. Compression and selection are.

### 4.5 Cross-Domain Synthesis

| **Domain** | **Substrate** | **Channel** | **Selection** | **Valence** |
|---|---|---|---|---|
| Challenger/Columbia | Org. hierarchy | Pref. divergence | Schedule press. | Conv. → deviance |
| Jazz improvisation | Musical interact. | Acoustic/cogn. | Aesthetic novelty | Div. → emergence |
| RLHF sycophancy | AI training | Reward compr. | Reward signal | Conv. → sycophancy |

No two domains share a substrate. The formal structure is identical across all three. The dual-valence claim is supported by instances on both sides of the valence divide, with the Challenger case (convergent) and jazz improvisation (divergent) sharing identical channel structure under opposite selection criteria.

## 5 Limitations, Falsification, and Open Questions

### 5.1 What the Theory Claims and Does Not Claim

The framework claims that five conditions are jointly *sufficient* for net-beneficial noise in a two-level system. It does not claim they are necessary. It does not claim that noise is generally good. It does not claim that the same optimal noise level applies across substrates. It claims: check these five conditions; if all hold, moderate noise helps; if any fails, no guarantee.

The framework claims that the same lossy-channel mechanism produces both organizational dysfunction and creative emergence, with the valence determined by the selection environment. It does not claim that dysfunction and creativity are identical—it claims they share a root cause.

The framework claims that compression-selection dynamics produce self-reinforcing equilibria (the ratchet). It does not claim that all organizations reach the same equilibrium, or that the ratchet operates at the same rate everywhere. Rate depends on compression ratio, preference divergence, verification cost, and feedback latency.

### 5.2 Falsification Conditions

**F1: Counterexample to sufficiency.** All five conditions verified as met, and the inverted-U does not appear. Zero counterexamples found across 500 Monte Carlo configurations and six mechanisms.

**F2: Mechanism-independent benefit.** Noise produces benefit with zero conditions met. Not observed.

**F3: Symmetric valence.** Two systems with identical channel properties but different selection criteria produce the same output valence. Would invalidate Proposition 2.10.

**F4: Drift without compression.** An organization or AI system exhibits systematic self-deception without compression or selection. Would require substantial revision of the substrate-independence claim.

**F5: Beneficial noise in a linear system.** Noise produces net system benefit where the integration function is strictly linear and the Jensen gap is identically zero. Not observed in computational null model.

### 5.3 Limitations of the Cross-Domain Evidence

The cross-domain evidence (Section 4) verifies qualitative satisfaction of conditions C1–C5. It does not operationalize channel capacity, noise structure, or generative residual with measurement precision within any domain. The organizational mapping relies on Vaughan's ethnographic account; the jazz mapping relies on Berliner's interview-based analysis; the AI mapping relies on Perez et al.'s and Sharma et al.'s evaluation frameworks. Each is the strongest available evidence in its domain, but none constitutes the domain-specific formal verification that would be required to claim Theorem 2.5 applies *as a quantitative result* rather than a qualitative structural prediction.

Domain-specific operationalization—measuring $b$, $P$, $d(p_{\hat{\theta}^*}, Q)$, and the five conditions with quantitative precision in organizational, musical, and AI settings—is a separate empirical research program.

### 5.4 Limitations of the Computational Validation

**Sufficient vs. necessary.** The 7.6% of configurations showing benefit despite condition violations means the conditions define a conservative boundary. Tightening the gap between sufficient and necessary is open.

**C2 steepness threshold.** The minimum steepness required for a meaningful Jensen gap is identified (between sigmoid and polynomial) but not quantified. Connection to Chapeau-Blondeau (1997) on generalized SR for arbitrary nonlinearities.

**Crawford–Sobel simplifications.** The simulation uses the one-sender-one-receiver model with uniform priors. Real hierarchies involve multiple senders, heterogeneous priors, and reputation effects.

### 5.5 Open Questions

**O1: Multi-level cascades.** Theorem 2.5 covers two levels. Does it compose across $N$ levels? Connection to Rosas et al. (2024) on hierarchical emergence via computational mechanics.

**O2: Dynamic environments.** Static conditions assumed. In non-stationary environments, the optimal noise level must track shifting thresholds.

**O3: Quantitative $\sigma^*$ prediction.** The theorem guarantees existence but not a closed-form for the optimal noise level.

**O4: Endogenous $Q$.** The most significant formal extension: the ratchet proposition (Proposition 2.12) treats the acceptability distribution dynamics qualitatively. A full characterization of the fixed-point $Q^*$ as a function of system parameters is a second paper.

**O5: Ratchet speed measurement.** The proposition establishes that the ratchet operates; it does not predict the rate. Operationalizing ratchet speed as a function of compression ratio and preference divergence would yield quantitative organizational diagnostics.

**O6: Mirror design.** High-reliability organizations (Weick and Sutcliffe, 2007) interrupt the ratchet through structural violations of condition (R4). Formalizing the "mirror" concept—structures that create observation points outside local selection pressure—in terms of the generative lossy channel framework is the natural applied extension.

## 6 Conclusion

This paper has formalized the generative lossy channel: a communication channel where compression, strategic misalignment, and conformity pressure combine to force the receiver to produce outputs that systematically diverge from the source. The mechanism is substrate-independent, appearing in organizational hierarchies and AI systems for the same structural reason: compression creates information gaps, and selection operating in those gaps determines the direction of the divergence.

Three results were established. Theorem 2.5 proved that five conditions are jointly sufficient for net-beneficial noise in any two-level system with a generative lossy channel, validated computationally with zero counterexamples across 500 random configurations. Proposition 2.10 established the dual valence of the generative residual: convergent selection produces pathological drift; divergent selection produces creative emergence. Proposition 2.12

formalized the compression ratchet: once the compression-selection dynamics cross a self-reinforcing threshold, recovery requires external intervention.

Three external evidence domains were mapped onto the C1–C5 framework. Vaughan's (1996) Challenger analysis demonstrated the ratchet under convergent selection, with all five conditions satisfied in the normalized deviance process. Berliner's (1994) jazz ethnography demonstrated creative emergence under divergent selection, with the same channel structure producing the opposite valence. Perez et al.'s (2022) and Sharma et al.'s (2023) AI alignment measurements demonstrated sycophancy as an equilibrium artifact of reward-compressed preference channels, providing the strongest evidence for substrate independence: the same mechanism operates in silicon with no human psychology involved.

The contribution is at the intersection of information theory, organizational theory, and AI alignment. These literatures have separately identified compression, strategic misalignment, and conformity pressure as sources of systematic error. The generative lossy channel framework identifies their interaction as the common mechanism—and proves that the interaction is sufficient, under stated conditions, to produce either pathological self-deception or creative emergence, depending on which selection criterion operates on the generative residual.

## References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Atlan, H. (1979). *Entre le cristal et la fumée.* Éditions du Seuil.

Berliner, P. F. (1994). *Thinking in Jazz: The Infinite Art of Improvisation.* University of Chicago Press.

Blau, Y. and Michaeli, T. (2018). The perception-distortion tradeoff. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Blau, Y. and Michaeli, T. (2019). Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proceedings of the 36th International Conference on Machine Learning (ICML).* arXiv:1901.07821.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process.* University of Chicago Press.

Chapeau-Blondeau, F. (1997). Stochastic resonance in the heaviside nonlinearity with white noise and arbitrary signal. *Physical Review E*, 55(2):2016.

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

DiMaggio, P. J. and Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2):147–160.

Johnson-Laird, P. N. (2002). How jazz musicians improvise. *Music Perception*, 19(3):415–442.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Schwartz, E., Khundadze, G., Kaplan, J., Brauner, J., Clark, J., Bowman, S. R., Askell, A., Grosse, R., and Hernandez, D. (2022). Discovering language model behaviors with model-written evaluations. arXiv:2212.09251.

Pressing, J. (1988). Improvisation: Methods and models. *Generative Processes in Music*, pages 129–178.

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., and Bor, D. (2024). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Computational Biology*.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rauber, O., Schreiber, N., Yan, D., Zhang, M., and Perez, E. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press.

Weick, K. E. and Sutcliffe, K. M. (2007). *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*. Jossey-Bass, 2nd edition.