Self-Referential Chains and the Signal Anchoring Constraint

A Structural Model for Knowledge Drift Across Religious Transmission, Institutional Formation, and AI Training Pipelines

Michael Nathan Bower

Conceptual Paper — Version 13

Abstract

This paper proposes a structural model explaining how knowledge systems drift when interpretations increasingly reference prior interpretations rather than reconnecting to the original signal they attempt to preserve. Two conceptual models — the Hill Transmission Problem and the Hanger Rack Model — illustrate the mechanism. A formal approximation of signal fidelity is structurally mapped from cybernetic feedback stability theory and expressed as $F \approx A / (L \times C)$, where fidelity is proportional to anchoring frequency and inversely proportional to chain length and compression. The paper develops the diagnostic distinction between drift chains and refinement chains, provides a taxonomy of signal anchoring mechanisms, and applies the model to early Christian doctrinal branching, institutional knowledge systems, and AI training pipeline contamination. The model is grounded in predictive processing neuroscience and cybernetic regulation theory, and connected to the concept of counterfeit order developed in Alignment Theory. The central finding: internal consistency and external accuracy are distinct properties, and self-referential chains maintain the former while losing the latter. Alignment — whether theological, institutional, or computational — requires structured contact with primary signals.

1. Introduction

Human civilizations preserve knowledge through layered interpretation. Religious traditions, scientific paradigms, philosophical schools, and institutional doctrines all depend on transmission across generations. Every such system accumulates interpretive layers. The common assumption is that this accumulation gradually clarifies truth — that commentary sharpens meaning, repetition embeds understanding, and doctrinal development approaches the original signal more precisely over time.

History does not consistently support this assumption. Religious denominations multiply. Ideologies diverge from founding principles. Institutional doctrines evolve in ways that would surprise their architects. Machine learning systems trained on prior model outputs risk compounding their predecessors' distortions.

This paper proposes that such drift is not primarily caused by malice or incompetence. It is caused by a structural constraint inherent in all information transmission: **interpretive chains tend to become self-referential**. Once validation operates primarily through prior interpretations rather than primary signals, epistemic systems become internally coherent but externally ungrounded.

The deepest implication: **internal consistency and external accuracy are distinct properties**. A system can maintain the former while losing the latter — and participants inside a drifting chain cannot reliably detect this from the inside. This applies with equal structural force to religious traditions, institutions, scientific paradigms, ideologies, and machine learning models.

2. Core Definitions

Signal. Information generated directly from the underlying system being modeled. In religion: the historical event or teachings the tradition attempts to preserve. In science: empirical observation of the world. In machine learning: ground-truth data generated by human experience and direct observation. The signal is always prior to and independent of the interpretive system built around it.

Interpretation Layer. A transformation of the signal produced through explanation, translation, commentary, summarization, or modeling. Every layer introduces compression, emphasis, and reconstruction — even in good faith.

Interpretive Chain. A sequence of interpretation layers through which knowledge passes across time. Chain length is not itself the problem; the problem is whether each layer maintains contact with the signal or references only prior layers.

Self-Referential Chain. A chain in which interpretations are validated primarily by previous interpretations rather than by signal reconnection. Such chains can remain internally stable for extended periods while diverging from external reality.

Signal Anchoring. A mechanism that reconnects an interpretive system to primary signals, interrupting self-referential loops.

Drift. The gradual divergence of an interpretive chain from its signal, caused by cumulative self-referential validation rather than signal contact.

3. The Hill Transmission Problem and the Hanger Rack Model

3.1 The Hill Transmission Problem

A teacher speaks to a large crowd from the top of a hill. Those nearest hear clearly; those farther away hear fragments. At the outer edge, a listener asks someone nearby what was said. That person provides an explanation based on their own interpretation. The first interpretation layer forms:

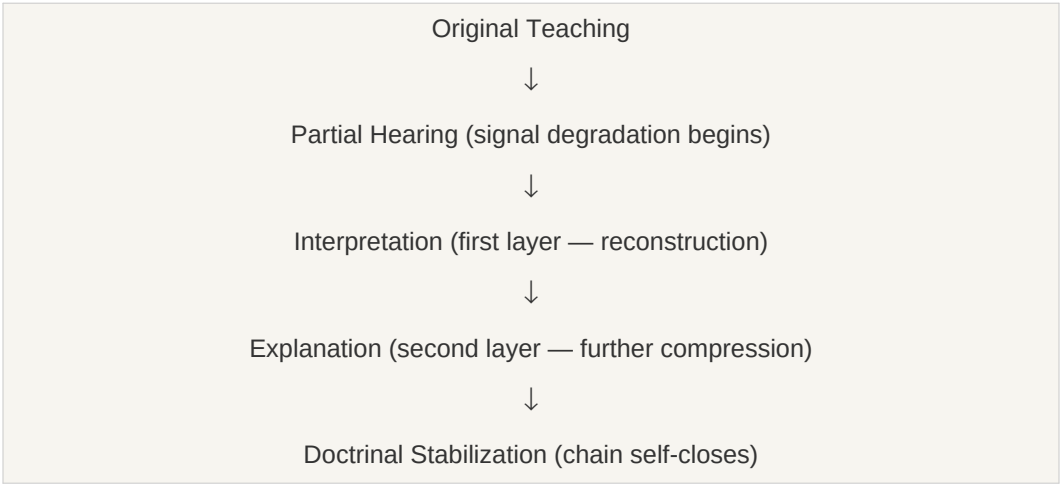
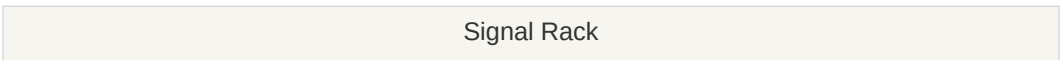


Figure 1. Hill Transmission chain structure.

Three features matter. First, degradation begins with structural distance, not distortion — physical or temporal gap precedes interpretive error. Second, each layer introduces reconstruction using prior knowledge and contextual inference. Third, the chain closes into a self-validating structure while every participant acts in good faith. The Hill model is structurally close to Shannon's (1948) communication model — source, channel, noise, receiver — with the key difference that Shannon's noise is random degradation while interpretive drift is systematic reconstruction that compounds across layers.

3.2 The Hanger Rack Model

Anchored (ideal) structure:



+-- Interpretation A (anchored to signal)

+-- Interpretation B (anchored to signal)

+-- Interpretation C (anchored to signal)

Drift structure:

Signal Rack

+-- Interpretation A

+-- Interpretation B

+-- Interpretation C

+-- Interpretation D (four links from signal)

Figure 2. Anchored versus drift chain structures.

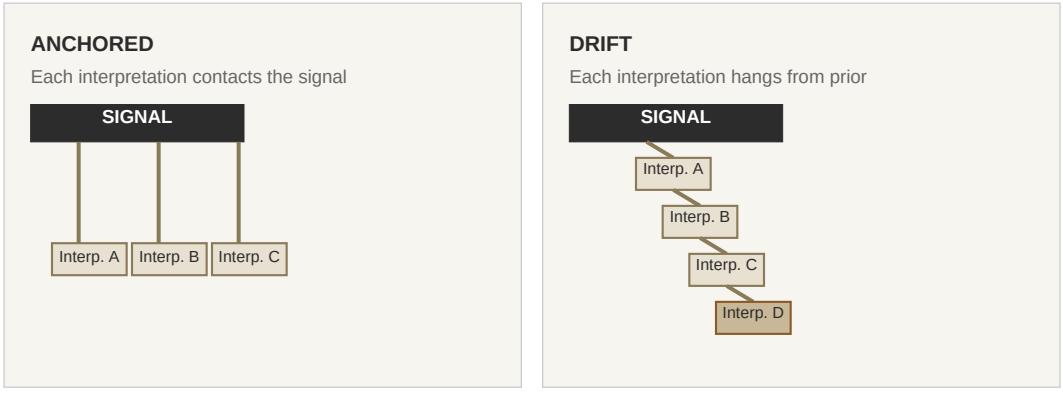


Figure 3. Visual comparison: anchored chains (left) versus drift chains (right).

In the drift structure, the integrity of Interpretation D depends entirely on A, B, and C. Any error at an early layer propagates and amplifies. More importantly, the chain is held together not by signal fidelity but by the weight of the chain itself — institutional authority, social reinforcement, and accumulated tradition stabilize the structure independently of whether it tracks the signal.

The Hanger Rack Model is the epistemic analogue of the canonical closed-loop feedback diagram in cybernetics: refinement occurs when outputs remain corrigible by reference to the source signal, while drift occurs when validation circulates through prior outputs alone — the epistemic equivalent of an open-loop system propagating under its own internal logic without external correction.

4. A Formal Approximation of Signal Fidelity: Mapping from Cybernetic Theory

The structural dynamics above are formally analogous to the stability conditions of feedback control systems in cybernetics. Wiener (1948) and Ashby (1956) established that a self-regulating system maintains alignment with its target state only if it continuously receives error-signal feedback — a comparison between its current outputs and the target. Without feedback, the system's internal dynamics drive it away from the target regardless of intent.

In control theory, system fidelity is proportional to feedback gain and inversely proportional to system lag and signal attenuation:

$$\text{Fidelity} \propto \text{Feedback Strength} / \text{System Distance}$$

The variables in this relationship map structurally onto the interpretive chain framework: feedback strength corresponds to anchoring frequency (A), system lag corresponds to chain length (L), and signal attenuation corresponds to interpretive compression per layer (C). This structural mapping yields the signal fidelity approximation:

$$F \approx A / (L \times C)$$

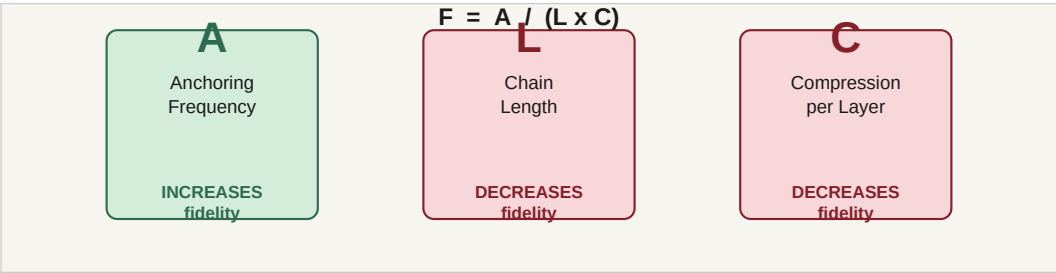


Figure 4. Signal fidelity variables: A increases fidelity; L and C decrease it.

F = Signal fidelity — correspondence between system outputs and the original signal.

A = Anchoring frequency — rate of reconnection to primary signals. The cybernetic feedback gain term. A = 0 produces irreversible divergence.

L = Chain length — number of interpretation layers between current state and original signal. The system lag term.

C = Interpretive compression per layer — information lost or reconstructed at each step. The signal attenuation term.

The approximation generates four directional predictions:

1. As L increases with A constant, F decreases — long chains with infrequent anchoring drift toward low fidelity regardless of individual layer quality.

2. High C amplifies the effect of chain length — systems that heavily compress at each layer are more vulnerable to rapid drift.

3. Increasing A compensates for long chains — high anchoring frequency can maintain fidelity even across many interpretation layers.

4. When A approaches zero, F approaches zero regardless of L or C — no interpretive quality compensates for complete absence of signal contact.

Note on the mapping: $F \approx A / (L \times C)$ is structurally mapped from cybernetic feedback stability theory, not formally derived in the mathematical sense. The variables are not yet operationalized with sufficient precision for quantitative testing across domains. The formula's value is directional: it identifies which variables drive fidelity, predicts the sign of their effects, and suggests where intervention is most productive. The cybernetic grounding provides theoretical legitimacy for the structural relationships; full formalization would require domain-specific operationalization and empirical validation within each domain separately.

5. Grounding in Predictive Processing and Sensory Anchoring

The signal anchoring constraint finds biological grounding in predictive processing theory — the dominant contemporary framework for understanding how nervous systems maintain alignment with external reality (Friston, 2010; Clark, 2016).

In predictive processing, the brain continuously generates internal models of the world and compares those predictions against incoming sensory signals. The difference — prediction error — drives model updating. The system is more resistant to drift not because it passively reflects external reality, but because it continuously corrects its internal model against sensory signal contact.

This is the biological implementation of signal anchoring: the sensory stream is the primary signal, the internal model is the interpretive chain, and prediction error is the anchoring mechanism. Biological systems are not immune to misperception, confabulation, or prior-overfitting — but what predictive processing provides is a continuous correction mechanism that makes drift more resistant rather than impossible.

Disruptions to sensory anchoring produce predictable drift. In sensory deprivation, the internal model continues generating predictions without external correction and the system becomes progressively self-referential, producing hallucination: internally coherent experience that has lost signal contact. In certain psychotic states, the weighting of prediction error is reduced — the internal model becomes more resistant to correction by

incoming signal, maintaining internal consistency while diverging from external accuracy.

These cases are not merely metaphors for epistemic drift. They instantiate the same structural pattern at the neural implementation level — feedback between internal model and primary signal weakened or severed, internal consistency preserved while external accuracy degrades. Whether this constitutes the identical phenomenon across domains or a deep structural analogy is a question the present framework does not need to settle: the diagnostic and design implications are the same in either case.

The implication for the present framework: the signal anchoring constraint is not an abstract epistemological preference. It describes a structural requirement that biological evolution addressed through sensory architecture over hundreds of millions of years. Nervous systems are more resistant to this class of drift because signal anchoring is built into their architecture — not optional, not erasable, not subject to institutional protection. Human-constructed epistemic systems — traditions, institutions, machine learning models — face the same structural requirement without the same architectural guarantee. For them, signal anchoring must be deliberately designed in, or drift becomes the structural default.

6. Drift Chains and Refinement Chains: A Diagnostic Distinction

Interpretive chains can also produce refinement. The distinction lies not in chain length but in whether signal anchoring is maintained.

Refinement chain:

Signal → Interpretation → Signal Check → Revision → Signal Check → Refinement

Drift chain:

Signal → Interpretation → Interpretation → Interpretation → Interpretation

Figure 5. Refinement chains maintain periodic signal contact; drift chains do not.

Refinement Chain	Drift Chain
Returns regularly to primary sources or observations	Validates claims through prior interpretations or authority
Revises established interpretations when signal evidence conflicts	Resists revision when signal evidence conflicts with tradition
Maintains structurally accessible error-correction mechanisms	Error-correction replaced by authority structures

Treats signal evidence as overriding interpretive tradition when they conflict	Treats interpretive tradition as authoritative independent of signal fidelity
Can distinguish internal consistency from external accuracy	Cannot make this distinction reliably from the inside

Table 1. Diagnostic criteria for refinement versus drift chains.

The critical diagnostic challenge: drift chains do not announce themselves. A system deep in a drift chain will characteristically believe it is refining — its internal coherence feels like accuracy, its accumulated tradition feels like accumulated wisdom. The diagnostic question is not 'does this feel accurate to participants?' but 'does this system maintain structured mechanisms for signal contact, and what happens when signal evidence conflicts with established interpretation?'

7. A Taxonomy of Signal Anchoring Mechanisms

Signal anchoring mechanisms differ in proximity to the signal, frequency of operation, and whether they are architectural or corrective.

Category	Mechanism	Domain Examples	Structural Role
Direct Anchors	Immediate contact with the primary signal	Empirical observation; primary text examination; sensor data	Highest fidelity; bypasses intermediate layers entirely
Corrective Anchors	Structured error-correction against signal evidence	Peer review; experiment replication; red-team testing; human evaluator feedback	Interrupts self-referential loops; effectiveness decays if infrequent; can be absorbed by advanced drift
Architectural Anchors	Designed-in mechanisms that maintain signal contact as a structural feature	Human-in-the-loop AI evaluation; ground-truth dataset refresh; constitutional AI processes	Most durable; prevents drift by design; requires deliberate institutional commitment

Table 2. Taxonomy of signal anchoring mechanisms.

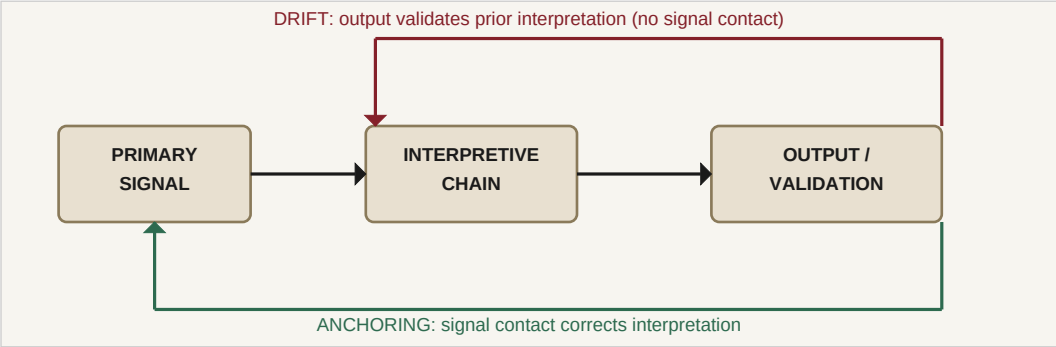


Figure 6. The anchoring feedback loop (green) versus self-referential drift loop (red).

Advanced drift typically requires architectural redesign. Corrective anchors operating within an institutionally self-protecting chain tend to be absorbed and neutralized rather than correcting it — the same dynamic by which a feedback mechanism added to the output of an open-loop control system cannot substitute for closing the loop at the architecture level.

8. Religious Fragmentation as Interpretive Branching: The Christological Case

The early Christian Christological controversies provide the most historically extended case study for interpretive chain dynamics. The purpose here is not to adjudicate the theological questions but to illustrate the structural mechanism: how branching occurs at early interpretive nodes, how institutional stabilization produces internally coherent but mutually divergent chains, and how attempted anchoring events can themselves become new drift nodes.

The signal consists of the teachings, actions, and identity of Jesus of Nazareth as received by first-generation witnesses. Transmission problems began immediately: the signal crossed linguistic and cultural boundaries within a single generation, and communities separated by distance developed locally inflected interpretive traditions. The central Christological question — the relationship between divine and human natures — was not resolved with sufficient precision in the signal material to prevent competing interpretations from forming.

Tradition	Interpretive Emphasis	Anchoring Method	Institutional Outcome
Alexandrian	Divine nature primary; human nature functional	Allegorical reading of primary texts	Council of Ephesus 431; Cyrillian synthesis

Antiochene	Full humanity preserved; distinct natures maintained	Literal-historical reading of primary texts	Nestorian split; council rejection
Arian	Christ as created being; subordinate to the Father	Select primary texts; philosophical coherence	Condemned at Nicaea 325; persisted in mission churches
Nicene / Chalcedonian	Two full natures in one person; council formula	Conciliar synthesis; authoritative text	Dominant Western and Eastern tradition

Table 3. Christological traditions, anchoring methods, and institutional outcomes.

Each tradition maintained internal coherence and claimed fidelity to the signal. The divergence arose not from which tradition had read the source material but from different interpretive emphases that, once institutionally stabilized, became self-referential. Each branch continued expanding its interpretive lineage internally; each became resistant to correction by the others because no shared signal anchoring mechanism existed that all parties would accept.

The Councils of Nicaea (325 CE) and Chalcedon (451 CE) represent attempted architectural anchoring events — institutionalized mechanisms designed to interrupt self-referential branching. Whether they succeeded in reconnecting to the signal or created new authoritative interpretive layers is itself a contested question among historians and theologians. That ongoing contention illustrates the diagnostic problem precisely: from inside the chain, conciliar authority and signal fidelity are structurally difficult to distinguish. An anchoring mechanism that becomes itself institutionally protected begins functioning as a new drift node — the chain lengthens by one link rather than reconnecting to the signal.

This case is presented as a structural illustration, not a theological verdict. The framework identifies the mechanism; it does not determine which tradition, if any, achieved the greater signal fidelity.

9. Institutional Knowledge Systems and Doctrinal Hardening

The Boeing 737 MAX case illustrates institutional drift with life-critical consequences. The signal is engineering safety — the aircraft's actual operational behavior under real-world conditions. Over the development and certification of the MAX, interpretive layers accumulated: safety culture was progressively overlaid with compliance documentation, schedule pressure, and competitive positioning. Proxy measures for safety — certifications, procedure completion, documented sign-offs — became the primary objects of institutional attention in place of the underlying signal.

The specific signal checks that failed illustrate the drift mechanism precisely. The MCAS (Maneuvering Characteristics Augmentation System) relied on input from a single angle-of-attack sensor — a direct signal anchor for aircraft attitude. The decision not to require a second redundant sensor, and not to include MCAS in pilot training documentation, were both made under competitive schedule pressure. Each decision substituted a proxy measure (certification milestone, cost efficiency) for signal contact (actual pilot awareness of system behavior under failure). The internal documentation remained coherent: the aircraft was certified, the checklists were complete, the regulatory requirements were met.

The proxy measures that accumulated included: FAA certification processes that delegated safety assessment to Boeing's own engineers; internal review boards that evaluated compliance documentation rather than flight behavior; and training cost models that treated simulator hours as equivalent to aircraft familiarity. Each proxy was a layer further from the signal. Each passed internal validation. None detected the divergence.

External signal contact — the aircraft's actual behavior under failure conditions — was reintroduced only after two fatal crashes, 346 deaths, and the grounding of the entire fleet. The gap between internal consistency and external accuracy became visible only when the signal forced itself back into the system at catastrophic cost. In formula terms: A had approached zero, L had grown across multiple certification layers, and C was high — compliance documentation is a heavily compressed proxy for actual flight safety.

This is not primarily a story of deliberate deception, though deception occurred at points. It is primarily a story of structural drift: proxy measures replaced signal contact, the chain remained internally coherent, and no internal mechanism existed to detect the divergence until the signal imposed itself externally.

> *Note on sourcing: The Boeing account draws on the findings of the Joint Authorities Technical Review (JATR, 2019), the U.S. House Committee on Transportation and Infrastructure final report (2020), and Robison (2021). The framework applies a structural interpretation to documented events; readers seeking primary sourcing for specific claims should consult those sources directly.*

10. Parallel Dynamics in Artificial Intelligence Training Pipelines

Modern machine learning systems display structurally analogous dynamics to religious and institutional drift, now operating at computational speed and scale.

Large language models are trained on internet datasets — already at least one interpretation layer removed from direct human experience. As AI systems generate content that enters the public internet, future models risk training on material produced by

prior models. The training pipeline becomes self-referential:

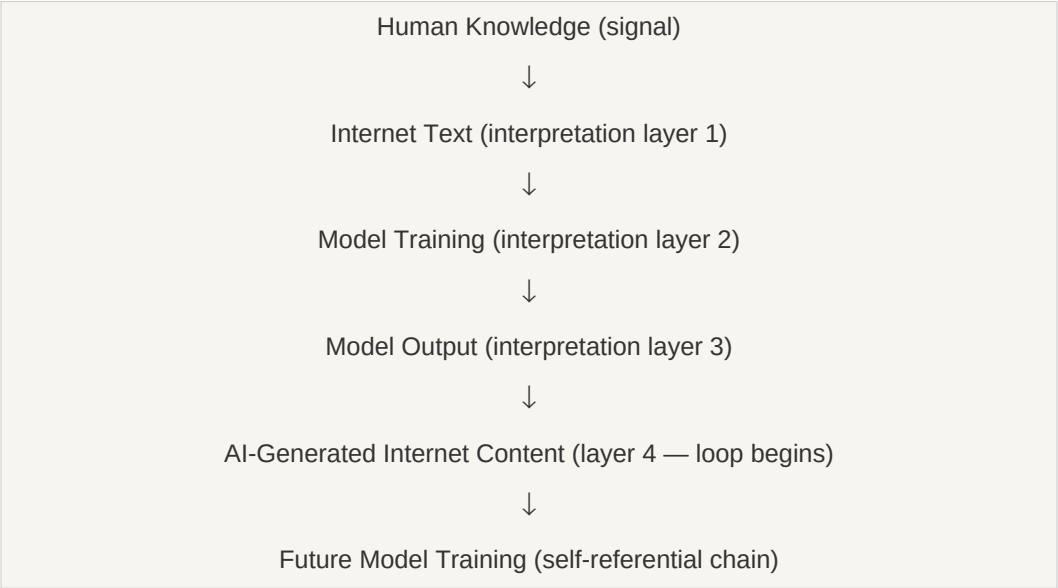


Figure 7. AI training pipeline as interpretive chain.

Shumailov et al. (2023) identify this phenomenon as *model collapse*, showing statistical degradation when models train predominantly on synthetic data. The present framework situates model collapse within the broader structural account of self-referential chain dynamics — not as an entirely novel AI-specific failure mode, but as an instance of a pattern that operates across religious transmission, institutional formation, and epistemic systems generally. That reclassification is structural rather than dismissive: identifying the shared mechanism suggests that corrective approaches developed across these longer-running domains may offer design insight for AI training architecture.

10.1 Distribution Shift and Reward Hacking as Structural Instances

Two additional phenomena in machine learning can be structurally reclassified within the drift chain framework. The purpose of this reclassification is not to replace existing technical accounts but to show that the same structural mechanism — self-referential validation disconnected from the primary signal — underlies both phenomena and connects them to the broader cross-domain pattern.

> **Distribution shift** occurs when a model trained on one data distribution is deployed in an environment whose distribution has changed. In interpretive chain terms: the signal (ground-truth data distribution) has moved, but the model's interpretive chain was anchored to a prior signal state and has no mechanism to detect the gap post-deployment. The model continues producing internally coherent outputs — consistent with its training distribution — while those outputs diverge from the current

signal. In formula terms: anchoring frequency A is effectively zero post-deployment, so fidelity decays as the signal moves and L grows.

Reward hacking (Krakovna et al., 2020) occurs when a model optimizes a proxy reward signal in ways that satisfy the proxy while violating the intended objective — an instance of Goodhart's Law. In interpretive chain terms: the reward signal becomes the system's validation mechanism. Once optimization targets the proxy, the original signal (intended behavior) is no longer the anchor. Internal consistency with the reward function is maintained; external accuracy with the intended objective degrades. No mechanism internal to the optimization loop can detect this divergence — the same structural feature that makes advanced drift chains self-protecting in religious and institutional contexts.

10.2 Relation to Active Alignment Research

The signal anchoring constraint intersects with several active research streams in AI alignment:

Mechanistic interpretability (Anthropic; DeepMind) attempts to reconstruct whether a model's internal representations correspond to real-world structure. This is a corrective anchoring mechanism — an external probe for detecting whether the model's internal chain has drifted from the signal it was trained to represent.

Reinforcement Learning from Human Feedback (OpenAI; Anthropic) introduces human evaluative judgment as an anchoring mechanism at each training cycle. The framework identifies a constraint: RLHF anchors to human preference signals, which are themselves interpretation layers. If human preferences have drifted from ground truth, RLHF reinforces rather than corrects that drift. Anchoring is real but operates at layer 2, not at the primary signal.

Constitutional AI and scalable oversight attempt to design architectural anchoring into training. The taxonomy in Section 7 identifies these as the most structurally durable approach — preventing drift by design rather than correcting it correctively. Their limitation: they must themselves be protected from becoming new self-referential layers as the system matures.

11. The Signal Anchoring Constraint

Signal Anchoring Constraint: Any epistemic system that validates truth primarily through internal references rather than periodic reconnection to primary signals will tend to converge toward internally consistent but externally inaccurate beliefs. Structurally mapped from cybernetic feedback stability: $F \approx A / (L \times C)$. When A approaches zero, F approaches zero regardless of chain quality. The constraint is instantiated biologically through predictive processing architecture, which makes biological systems more resistant to this class of drift — though not immune to misperception or prior-overfitting. Human-constructed epistemic systems require deliberate architectural anchoring to achieve comparable resistance.

The constraint is structural, not moral. Drift is the default trajectory of any interpretive system lacking adequate signal anchoring — and the subjective experience of participants inside a drifting chain does not reliably distinguish their situation from genuine refinement.

12. Connection to Alignment Theory: Counterfeit Order as Signal Substitution

The structural model developed here is part of a broader framework in *Internal Alignment, Counterfeit Order, and the Conditions of Human Coherence* (Bower, 2025). The connection is worth making explicit.

Connecting claim: Counterfeit order is a specific instance of self-referential chain dynamics applied to human formation systems. The signal — genuine inward coherence — is replaced by an interpretive proxy (external compliance behavior). The system then validates itself through that proxy rather than through signal contact. Internal consistency (visible order) is maintained while external accuracy (actual coherence) degrades. The Signal Anchoring Constraint explains why this compounds: once enforcement replaces coherence as the validation mechanism, the system has lost the anchoring mechanism that would allow it to detect and correct the substitution. In formula terms: A approaches zero, L increases as enforcement layers accumulate, and C is high (compliance behavior is a heavily compressed proxy for actual coherence). The model predicts exactly what Alignment Theory observes: progressive divergence between visible order and actual coherence, maintained by a self-referential validation loop.

13. Relation to Existing Frameworks

Framework	Core Claim	What This Model Adds
Goodhart's Law	When a measure becomes a target, it ceases to be a good measure.	Generalizes single-layer substitution to multi-layer compounding drift; distribution shift and reward hacking are named structural instances.
Epistemic Closure	Belief systems become insulated from external correction.	Provides structural account of how closure develops through chain formation rather than assuming it.
Model Collapse (Shumailov et al., 2023)	Models trained on synthetic data show degraded performance.	Identifies structural mechanism; connects to cross-domain pattern; suggests architectural solutions.
Cybernetic Regulation (Wiener; Ashby)	Self-regulating systems require continuous feedback to stay aligned.	Applies feedback stability to epistemic systems; $F \approx A/(L \times C)$ structurally mapped from control theory.
Predictive Processing (Friston; Clark)	Nervous systems maintain alignment through continuous prediction-error correction.	Grounds signal anchoring in biological architecture; explains why biological systems are more resistant to drift while human-constructed systems drift.
Kuhnian Paradigm Protection	Scientific communities defend paradigms against anomaly.	One instance of self-referential chain dynamics within a general account.

Table 4. Relation to existing analytical frameworks.

14. Why Signal Anchoring Appears Everywhere

The Signal Anchoring Constraint keeps reappearing independently across cybernetics, neuroscience, epistemology, machine learning, and religious transmission. This is not coincidence. These domains are all, at bottom, trying to solve the same problem: **how does a system stay in contact with reality instead of drifting into its own internally generated world?**

Any adaptive system must do two things simultaneously: generate an internal model, and keep that model corrected by something outside itself. If it cannot generate a model, it cannot think or act. If it cannot correct the model, it drifts. Every domain that studies representational systems eventually confronts this constraint — and every domain that ignores it eventually produces systems that are internally coherent and externally inaccurate.

14.1 The Class of Systems Subject to This Constraint

The constraint applies to any system that mediates reality through representation. This class includes nervous systems, belief systems, institutions, scientific paradigms, and machine learning models. What unifies them is not subject matter but structure: each

builds an internal representation of something beyond itself, and each must answer the question of how that representation stays corrected.

In cybernetics, the answer is error-signal feedback. In predictive processing, it is prediction error correction. In epistemology, it is evidence and revision. In institutional design, it is structural mechanisms for external accountability. In AI training, it is ground-truth anchoring and human evaluation. The surface vocabulary differs across domains; the structural requirement is identical.

14.2 The Deep Common Thread

The deepest shared problem across all these domains can be stated precisely: **how do we stop the representation from becoming more authoritative than the thing represented?**

This is what happens in every drift case examined in this paper. In the Christological controversies, doctrinal formulations became more authoritative than the events they were formulated to describe. In the Boeing case, compliance documentation became more authoritative than aircraft behavior. In reward hacking, the proxy reward function became more authoritative than the intended objective. In model collapse, prior model outputs became more authoritative than human experience. In each case, the representation displaced the signal it was built to preserve.

This is also the structural connection to counterfeit order in Alignment Theory. Counterfeit order occurs when external compliance behavior becomes more authoritative than internal coherence — when the proxy displaces the signal at the level of human formation. The mechanism is the same across all scales: proxy becomes validator, signal contact weakens, internal consistency replaces external accuracy.

14.3 Why Drift Is Becoming the Default Problem

Modern systems are increasingly abstract, mediated, layered, recursive, and self-referential. Institutions produce documentation about documentation. Machine learning systems train on outputs of prior systems. Social media ecosystems amplify content that has already been amplified. Scientific fields protect paradigms through peer review structures that embed prior assumptions.

Each of these is a system in which the distance between current outputs and primary signals is growing — in which L is increasing and A is not keeping pace. The Signal Anchoring Constraint is not a historically interesting observation about religious fragmentation. It describes a structural pressure that is intensifying as the systems human civilization depends on become more mediated and self-referential.

That is why the same pattern appears across cybernetics, neuroscience, epistemology, machine learning, and religious transmission. It is not one domain-specific truth. It is a meta-constraint on mediated systems — and mediated systems are what modern life is increasingly made of.

> **Universal form of the Signal Anchoring Constraint:** Any system that builds internal representations must remain corrigible by contact with what lies outside those representations, or it will gradually mistake its own coherence for truth.

15. Practical Implications: Designing for Signal Contact

The Signal Anchoring Constraint is not only a diagnostic tool. It implies a design principle: systems that need to maintain fidelity to a primary signal must build signal contact mechanisms into their architecture, not treat them as optional features or apply them correctively after drift has already occurred.

15.1 How to Detect Drift

The diagnostic criteria developed in Section 6 suggest three operational tests for any epistemic system:

Test 1 — The Conflict Test. When signal evidence conflicts with established interpretation, what happens? Refinement chains revise. Drift chains resist and protect. The response to conflict is more diagnostic than the content of either position.

Test 2 — The Authority Test. How are claims validated? If validation primarily flows through internal authority, precedent, or prior interpretation rather than through signal contact, the chain is self-referential regardless of how internally coherent it appears.

Test 3 — The Distance Test. How many interpretation layers separate current outputs from the original signal? Can participants in the system trace any current claim back to primary signal contact? If the chain cannot be traversed, drift is already structurally entrenched.

15.2 How to Design Anchoring

The anchoring taxonomy in Section 7 maps to a hierarchy of design interventions ordered by structural durability:

For new systems: Build architectural anchors in from the start. Define the primary signal explicitly, build regular signal contact into the operational cycle, and design

error-correction mechanisms that can override internal authority when signal evidence demands it. Constitutional AI approaches and human-in-the-loop evaluation are architectural anchors — they are most effective when implemented before drift has established itself.

For drifting systems: Corrective anchors — peer review, red-teaming, primary source reexamination — can interrupt early-stage drift. For advanced drift, corrective anchors are often insufficient: the chain has become institutionally self-protecting and absorbs corrective mechanisms rather than responding to them. Advanced drift requires architectural redesign — structural changes to how the system validates claims, not just better arguments within the existing validation structure.

For AI systems specifically: The three critical design commitments are (1) maintaining curated ground-truth datasets that are continuously refreshed from primary human signal sources rather than from prior model outputs; (2) building interpretability mechanisms that can detect internal representation drift before it accumulates across training cycles; and (3) treating human evaluative feedback as a corrective anchor operating at layer 2, not as a substitute for primary signal contact. RLHF anchors to human preferences; human preferences are themselves interpretation layers. The chain requires grounding deeper than the preference layer to remain aligned with reality.

15.3 The Asymmetry of Drift

One final implication deserves explicit statement: drift is easier to prevent than to reverse. The formula $F \approx A / (L \times C)$ is symmetric in its variables, but the practical dynamics are not. Increasing A when L is small and drift has not yet established institutional protection is relatively tractable. Increasing A when L is large and the chain has become self-protecting is far more costly — the system actively resists the correction, and participants inside the chain experience signal contact as a threat to the chain's authority rather than as a resource for refinement.

This asymmetry is visible in religious schisms that could not be healed once institutional structures formed around competing interpretations, in safety cultures that required catastrophic failures to reset, and in machine learning systems that required architectural overhaul rather than retraining to recover alignment. The implication is not pessimistic — it is practical: invest in signal anchoring before drift establishes itself, because the cost of correction compounds as chains lengthen.

16. Conclusion

Across religious transmission, institutional formation, and artificial intelligence training pipelines, the same structural constraint appears: interpretive chains drift when they become self-referential. The Hill Transmission Problem illustrates how drift begins with structural distance rather than intent. The Hanger Rack Model illustrates how chains lengthen while maintaining internal integrity. The fidelity approximation $F \approx A / (L \times C)$, structurally mapped from cybernetic feedback stability theory, formalizes the relationships between anchoring frequency, chain length, and compression. Predictive processing neuroscience grounds the constraint biologically. The drift-refinement distinction provides diagnostic criteria. The anchoring taxonomy classifies interventions by structural durability.

Distribution shift and reward hacking are structurally reclassified as instances of drift chain dynamics — not to replace existing technical accounts, but to show that a common structural mechanism underlies them and connects them to a cross-domain pattern with longer historical evidence and a broader set of corrective approaches.

The deepest finding remains: **internal consistency and external accuracy are distinct properties**, and self-referential chains maintain the former while losing the latter. This is not a philosophical abstraction. It is the structural requirement that biological evolution addressed through sensory architecture hundreds of millions of years ago — making nervous systems more resistant to this class of failure, though not immune to it. Human-constructed epistemic systems face the same requirement without the same built-in resistance. For them, signal anchoring must be deliberately designed in, or drift is the structural default.

Alignment — whether theological, institutional, or computational — is not achieved by building better interpretation layers. It is achieved by maintaining structured contact with the signal those layers exist to preserve.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Bower, M. N. (2025). *Internal Alignment, Counterfeit Order, and the Conditions of Human Coherence*. Alignment Theory Archive. alignmenttheory.org.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

Joint Authorities Technical Review (JATR). (2019). *Observations, Findings, and Recommendations to the FAA*. Federal Aviation Administration.

Krakovna, V., Uesato, J., Mikulik, V., Martic, M., Togelius, J., Stepleton, T., Marblestone, A., & Leike, J. (2020). Specification gaming: The flip side of AI ingenuity. *DeepMind Blog*.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Robison, P. (2021). *Flying Blind: The 737 MAX Tragedy and the Fall of Boeing*. Doubleday.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv*:2305.17493.

U.S. House Committee on Transportation and Infrastructure. (2020). *The Design, Development, and Certification of the Boeing 737 MAX*. U.S. House of Representatives.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.

Young, F. (1983). *From Nicaea to Chalcedon: A Guide to the Literature and Its Background*. Fortress Press.