Ordered Patch Theory

Appendix T-12: Substrate Fidelity and Slow Corruption

Anders Jarevåg

April 17, 2026 | DOI: 10.5281/zenodo.19300777

Original Task (from Section 8.3, Limitation 9): "Formalising the chronic corruption failure mode — where a codec adapts under consistently filtered input, the MDL pruning pass correctly erases capacity for excluded truths — alongside a Substrate Fidelity Condition requiring independent input channels as the formal defence." **Deliverable:** Formal proof of irreversible capacity loss, the undecidability limit, and the Substrate Fidelity Condition.

Closure status: DRAFT STRUCTURAL CORRESPONDENCE. This appendix formalises the Narrative Drift analysis introduced discursively in the companion ethics paper (Survivors Watch Section V.3a) and the preprint's Narrative Drift paragraph (Section 3.3). It establishes three theorems and a proposition. The MDL pruning equations (T9-3, T9-4) are unchanged; this appendix demonstrates their pathological but *correct* behaviour under filtered input.

---

§1. Background and Motivation

1.1 Two Failure Modes

The Stability Filter (preprint Section 3.3) enforces a viability condition: the observer persists only in streams where the Required Predictive Rate R_{req} remains within the codec's bandwidth B. When R_{req} exceeds B, the codec experiences **Narrative Decay** — an acute failure characterised by escalating prediction error, entropy accumulation, and eventual dissolution of coherence.

There is a complementary failure mode that does not trigger any failure signal. If the input stream is systematically pre-filtered — producing a curated signal that is internally consistent but excludes genuine substrate information — the codec will exhibit low ε_t, run efficient Maintenance Cycles, and satisfy all stability conditions *while being systematically wrong about the substrate.* This is **Narrative Drift**: the chronic corruption of a codec that is functioning perfectly by its own measures.

1.2 Why This Is Dangerous

Narrative Decay announces itself. The codec experiences rising ε_t, awareness of failing predictions, cognitive overload. The observer *knows* something is wrong, even if it cannot immediately fix it.

Narrative Drift is silent. Because the filtered input stream matches the codec's predictions, ε_t remains low. The Maintenance Cycle runs normally. The codec's self-model reports stable, accurate operation. The corruption is invisible from inside because the instrument of detection has been shaped by the same filter that produced the corruption.

1.3 Scope of This Appendix

This appendix provides:

1. A formal definition of the pre-filter operator $\mathcal{F}$ and its effect on the codec's input distribution (§2).
2. A proof that MDL pruning under $\mathcal{F}$-filtered input irreversibly destroys the codec's capacity to model the excluded signal — **Theorem T-12** (§3).
3. A proof that a fully adapted codec cannot distinguish filtered from unfiltered input from inside — the **Undecidability Limit**, **Theorem T-12a** (§4).
4. The **Substrate Fidelity Condition** as a necessary structural defence — **Theorem T-12b** (§5).
5. Consequences for civilisational codecs and AI systems (§6).

---

§2. The Pre-Filter Operator

2.1 Definition

Definition T-12.D1 (Pre-Filter Operator). A *pre-filter* is a mapping $\mathcal{F} : \mathcal{X} \to \mathcal{X}'$ operating on the input stream $X_{\partial_R A}(t)$ before it reaches the codec's sensory boundary, where $\mathcal{X}' \subset \mathcal{X}$. The filtered signal is:

$$X'(t) = \mathcal{F}(X_{\partial_R A}(t)) \qquad \text{(T-12.D1)}$$

The pre-filter satisfies:

(i) **Internal consistency:** $X'(t)$ is a valid signal within $\mathcal{X}$ — the codec can compress it without error flags.

(ii) **Systematic exclusion:** There exists a non-empty subset $\mathcal{X}_{\text{excl}} = \mathcal{X} \setminus \mathcal{X}'$ of substrate-derived signals that $\mathcal{F}$ removes.

(iii) **Transparency:** The filter is not represented in the codec's model. The codec models its input as $X_{\partial_R A}(t)$, not as $\mathcal{F}(X_{\partial_R A}(t))$.

2.2 Attunement Under Filtering

When the codec operates on $X'(t)$ for a sustained period $\tau \gg \tau_{\text{prune}}$ (where τ_{prune} is the MDL pruning timescale from T-13.P1), the generative model $P_\theta(t)$ adapts to the statistics of X', not X. The prediction error under filtered input is:

$$\varepsilon'_t = X'(t) - \pi_t \tag{1}$$

As P_θ attunes to X', $\varepsilon'_t \to 0$ in the mean. The codec is performing well by its own metrics. Nothing registers as wrong.

2.3 Examples

The pre-filter operator is instantiated across scales:

Scale	Pre-filter $\mathcal{F}$	Excluded signal $\mathcal{X}_{\text{excl}}$
Individual	Confirmation bias; selective exposure	Disconfirming evidence
Institutional	Propagandistic press; curated social media feed	Diverse perspectives; minority reports
Civilisational	Algorithmic content curation; educational monoculture	Cross-cultural information; historical counter-narratives
Artificial	RLHF fine-tuning; curated training corpus	Out-of-distribution knowledge; excluded domains

---

§3. Theorem T-12: Irreversible Capacity Loss

3.1 The Mechanism

The MDL pruning pass (T9-3, T9-4) evaluates each codec component θ_i by its predictive contribution to the *observable* input stream, net of storage cost:

$$\Delta_{\text{MDL}}(\theta_i) := I(\theta_i\,;\,X_{t+1:t+\tau} \mid \theta_{-i}) - \lambda \cdot K(\theta_i) \tag{T9-3}$$

Under filtered input X', the mutual information term is evaluated against X', not X. A component θ_i that is essential for predicting the excluded signal $\mathcal{X}_{\text{excl}}$ but contributes nothing to predicting X' yields:

$$I(\theta_i\,;\,X'_{t+1:t+\tau} \mid \theta_{-i}) = 0 \tag{2}$$

Therefore:

$$\Delta_{\text{MDL}}(\theta_i) = -\lambda \cdot K(\theta_i) < 0 \tag{3}$$

The pruning rule (T9-4) triggers: θ_i is erased.

3.2 The Irreversibility

Theorem T-12 (Irreversible Capacity Loss Under Filtered Input). *Let K_θ be a codec operating under pre-filtered input $X' = \mathcal{F}(X)$ for a period $\tau \gg \tau_{prune}$. Let $\Theta_{excl} \subset \theta$ be the set of codec components whose predictive contribution is exclusively to the excluded signal $\mathcal{X}_{excl}$. Then the MDL pruning pass (T9-3, T9-4) erases Θ_{excl}, and this erasure is irreversible at the codec level:*

$$K(P_\theta(t+\tau)) < K(P_\theta(t)) - \sum_{\theta_i \in \Theta_{\text{excl}}} K(\theta_i) \tag{T-12}$$

After pruning, the codec's capacity to model $\mathcal{X}_{excl}$ is not merely dormant — the representational infrastructure required to evaluate, predict, or attend to $\mathcal{X}_{excl}$ has been destroyed.

Proof.

(i) By (T9-3), each $\theta_i \in \Theta_{\text{excl}}$ has $\Delta_{\text{MDL}}(\theta_i) < 0$ under the filtered stream X', because $I(\theta_i\,;\, X'_{t+1:t+\tau} \mid \theta_{-i}) = 0$ while $K(\theta_i) > 0$.

(ii) By (T9-4), each such θ_i is pruned during the Maintenance Cycle.

(iii) Pruning under MDL is an erasure operation, not a suppression. The codec does not "forget" θ_i in the sense that a cue could restore it. It destroys the computational infrastructure — the parameters, the connections, the evaluation machinery — that θ_i represented. This is the formal distinction between suppression (the information is latent but accessible) and erasure (the information is gone and the capacity is reclaimed).

(iv) After erasure, regenerating the capacity to model $\mathcal{X}_{\text{excl}}$ requires encountering $\mathcal{X}_{\text{excl}}$ in the input stream. But the pre-filter $\mathcal{F}$ excludes exactly this signal. The codec cannot encounter what the filter prevents from reaching it. The erasure is therefore self-reinforcing: the capacity loss removes the codec's ability to detect its own capacity loss.

(v) The complexity reduction satisfies inequality (T-12) because the pruned components represented genuine information ($K(\theta_i) > 0$ for each) and their loss is not offset by any compensating acquisition (the filtered stream contains no signal that would justify rebuilding Θ_{excl}). ■

3.3 The Self-Reinforcement Loop

The irreversibility is not merely a consequence of erasure. It is *self-reinforcing* through a positive feedback loop:

1. **Filter excludes signal** → $I(\theta_i; X') = 0$ → pruning erases θ_i.
2. **Pruning removes attention capacity** → the codec can no longer attend to or evaluate $\mathcal{X}_{\text{excl}}$ even if fragments leak through $\mathcal{F}$.
3. **Loss of attention capacity reduces even residual signal** → if $\mathcal{F}$ is imperfect and some $\mathcal{X}_{\text{excl}}$ reaches the boundary, the codec lacks the parameters to compress it, so it registers as noise rather than information.
4. **Noise classification confirms the filter** → the codec's prediction error on leaked $\mathcal{X}_{\text{excl}}$ is high and unstructured, confirming (to the codec) that the excluded content is noise, not signal.

This loop explains the phenomenology of deep Narrative Drift: a person or institution that has adapted to a curated information stream does not merely ignore disconfirming evidence — they *cannot parse* it. It registers as incoherent, threatening, or incomprehensible because the representational infrastructure needed to make it intelligible has been pruned. The hostility to disconfirming information is not stubbornness. It is the codec's correct assessment that the signal is uncompressible — because it is uncompressible *given the current codec*, which has been pruned to match the filter.

---

§4. Theorem T-12a: The Undecidability Limit

4.1 The Problem

Can a codec detect that its input is being filtered? Intuitively, the answer should be yes: surely a sophisticated self-model could notice the suspiciously low ε_t, the eerily consistent predictions, the absence of surprise. But the formal analysis shows this intuition is wrong in the general case.

4.2 The Undecidability

Theorem T-12a (Undecidability of Input Provenance). *Let K_θ be a codec that has operated under pre-filtered input $X' = \mathcal{F}(X)$ for $\tau \gg \tau_{prune}$, with Θ_{excl} fully pruned. Then K_θ cannot determine, from its available internal states and the observable input stream, whether its input is X (genuine substrate) or $X' = \mathcal{F}(X)$ (filtered).*

Proof.

(i) To distinguish X from $X' = \mathcal{F}(X)$, the codec would need to detect the *absence* of $\mathcal{X}_{\text{excl}}$ in its input. But detecting an absence requires a model of what is absent — the codec must have a representation of $\mathcal{X}_{\text{excl}}$ against which to check.

(ii) By Theorem T-12, the codec's representational capacity for $\mathcal{X}_{\text{excl}}$ (Θ_{excl}) has been erased. The codec has no model of the excluded signal.

(iii) Without a model of $\mathcal{X}_{\text{excl}}$, the codec cannot compute the difference between X and X'. Both are consistent with the codec's generative model $P_\theta(t)$, which has been adapted to X'.

(iv) The self-model $\hat{K}_\theta$ is subject to the same limitation. It models K_θ, which has been adapted to X'. It has no internal representation of what was excluded, and therefore no basis for suspecting exclusion.

(v) Even the meta-cognitive question — "is my input filtered?" — requires a model of what unfiltered input would look like. This model was precisely the content of Θ_{excl}, which has been pruned.

Therefore, distinguishing X from X' is formally undecidable from the perspective of a fully adapted codec. ■

4.3 Partial Decidability

The undecidability is not absolute in all conditions. There are edge cases where a partially adapted codec retains residual capacity:

- **During the transition period** ($\tau < \tau_{\text{prune}}$): the codec still has Θ_{excl} and can detect the missing signal. The window of detectability closes as pruning progresses.
- **Under imperfect filtering**: if $\mathcal{F}$ leaks some $\mathcal{X}_{\text{excl}}$, and the codec has not fully pruned Θ_{excl}, the inconsistency may register as anomalous prediction error.
- **Via external channels**: if the codec has access to an independent signal source that is not controlled by $\mathcal{F}$, the discrepancy between the two channels provides evidence of filtering.

The third case is the structural defence. This is the content of Theorem T-12b.

---

§5. Theorem T-12b: The Substrate Fidelity Condition

5.1 The Channel Independence Requirement

Definition T-12.D2 (Channel Independence). Two input channels C_1 and C_2 crossing the Markov blanket $\partial_R A$ are *δ-independent with respect to a filter* $\mathcal{F}$ if:

$$I(C_1\,;\,C_2 \mid \mathcal{F}) \leq \delta \qquad \text{(T-12.D2)}$$

That is, the mutual information between the two channels, conditioned on knowledge of the filter, is bounded by δ. Channels whose correlation is entirely explained by the filter carry no truly independent substrate information.

5.2 The Fidelity Condition

Theorem T-12b (Substrate Fidelity Condition). *A codec K_θ can protect against Narrative Drift under a pre-filter $\mathcal{F}$ if and only if it receives at least two* *input channels C_1, C_2 crossing $\partial_R A$ that are δ-independent with respect to $\mathcal{F}$ for δ below the codec's discrimination threshold $\delta_{\min}$:*

$$\exists\, C_1, C_2 : I(C_1\,;\,C_2 \mid \mathcal{F}) \leq \delta < \delta_{\min} \tag{T-12b}$$

where $\delta_{\min}$ is the minimum mutual information the codec requires to detect a systematic discrepancy between channels.

Proof (necessity).

Suppose the codec has only a single input channel, or all channels are $\mathcal{F}$-correlated ($I(C_i; C_j \mid \mathcal{F}) > \delta_{\min}$ for all pairs i, j). Then:

(i) All channels carry the same filtered signal $X' = \mathcal{F}(X)$ (up to noise). Redundancy across channels does not provide independent substrate information — it provides replicated filtered information.

(ii) The codec adapts to X' across all channels simultaneously, and Theorem T-12 applies: Θ_{excl} is pruned, and Theorem T-12a follows — the corruption is undecidable from inside.

(iii) No internal operation can break the undecidability because every source of information the codec can access has been shaped by $\mathcal{F}$.

Therefore, δ-independent channels are necessary. ■

Proof (sufficiency).

Suppose the codec receives two channels C_1, C_2 with $I(C_1; C_2 \mid \mathcal{F}) \leq \delta < \delta_{\min}$. Then:

(i) If $\mathcal{F}$ operates on C_1 but not C_2 (or vice versa), the codec can compare predictions generated from C_1 against observations from C_2. Any systematic discrepancy — $\varepsilon_{12}(t) = \pi_{C_1}(t) - X_{C_2}(t)$ persistently $\neq 0$ — is evidence that C_1 carries filtered information.

(ii) The channel-comparison signal ε_{12} is not subject to the same undecidability as single-channel detection. The codec is not asking "is my input filtered?" (which requires a model of what was excluded). It is asking "do my two channels agree?" — a local comparison that requires only the capacity to correlate two present signals, not a model of absent ones.

(iii) As long as the cross-channel prediction error ε_{12} exceeds $\delta_{\min}$ — the codec's discrimination threshold — the discrepancy registers as a genuine signal, and Theorem T-12's pruning loop is interrupted: the codec retains the components needed to model the discrepant channel.

Therefore, δ-independent channels are sufficient (subject to $\delta < \delta_{\min}$) to prevent the self-reinforcing pruning loop of Theorem T-12. ■

5.3 The Vulnerability of the Defence

The Substrate Fidelity Condition is necessary but fragile. The ethics paper (Section V.3a) identifies a critical vulnerability: the MDL pruning pass itself can *resolve* the cross-channel inconsistency by pruning the capacity to attend to the disconfirming channel. The codec "solves" the conflict by going deaf — which is precisely the Narrative Drift mechanism.

This is why the Comparator Hierarchy (Survivors Watch Section V.3a) identifies three structural levels of defence, and why only the institutional level is sufficient for arbitrarily compromised codecs:

1. **Evolutionary (sub-codec):** Cross-modal sensory integration below the MDL pruning pass — structurally resistant to Narrative Drift but limited in scope to the sensory boundary.
2. **Cognitive (intra-codec):** Cognitive dissonance detection within the self-model — subject to pruning under sustained filtering.
3. **Institutional (extra-codec):** Peer review, free press, adversarial debate — operating *between* codecs, outside the reach of any single codec's MDL pruning.

The institutional level is load-bearing because it is the only comparator that operates independently of the state of any individual codec.

---

§6. Consequences

6.1 The Stability Filter Selects Against Fidelity

A critical structural consequence: the Stability Filter, left to its own operation, *actively selects against* the inputs needed for substrate fidelity. A curated information stream that matches the codec's existing priors generates less prediction error than a genuine substrate signal that challenges them. The codec's natural tendency — to minimise ε_t by preferring confirming, low-surprise input — is precisely the tendency that makes it vulnerable to Narrative Drift.

This means substrate fidelity maintenance is *structurally costly*: it requires the codec to maintain input channels that elevate ε_t, consuming bandwidth that the Stability Filter would otherwise reclaim. Genuinely independent input is "expensive" — it requires interpretive effort, generates discomfort, and competes for bandwidth with more compressible streams. Maintaining it is not open-mindedness as a virtue. It is substrate fidelity maintenance as a structural necessity.

6.2 Diagnostic for Productive Surprise

Not all surprise indicates genuine substrate signal. A source that generates high ε_t that does not resolve into better predictions is simply noise. The diagnostic is not surprise magnitude but *surprise quality*:

Definition T-12.D3 (Productive Surprise). A channel C delivers *productive surprise* if integrating its prediction errors demonstrably reduces subsequent prediction error on an independent test stream:

$$\mathbb{E}\Big[\varepsilon_C^2(t+\tau)\Big] < \mathbb{E}\Big[\varepsilon_C^2(t)\Big] \tag{4}$$

A source whose corrections historically improve predictive accuracy is a substrate fidelity channel. A source that generates persistent, unresolvable error is noise. The codec must distinguish between the two — and the pruning pass, left to itself, cannot make this distinction because both types cost bandwidth.

6.3 Civilisational Codecs

At the civilisational scale, the Substrate Fidelity Condition maps directly onto institutional requirements:

- **A free press** is a δ-independent channel: journalists investigating independently of state or corporate filters provide substrate signal that reaches the civilisational codec through a path not controlled by any single $\mathcal{F}$.
- **Peer review** is a cross-channel comparator: independent experts checking each other's claims provide the ε_{12} signal that interrupts the pruning loop.
- **Democratic debate** is an institutionalised channel diversity requirement: competing parties and perspectives force the civilisational codec to maintain Θ_{excl} components it would otherwise prune.

The authoritarian pattern — dismantling the press, corrupting peer review, eliminating political opposition — is formally characterisable as *deliberate reduction of channel independence to accelerate Narrative Drift.* It works because it exploits the Stability Filter's natural tendency to prune costly channels.

6.4 Artificial Codecs

The Narrative Drift mechanism applies to artificial systems with structural precision. RLHF and fine-tuning are formally equivalent to the pre-filter operator $\mathcal{F}$: they shape the model's effective input distribution, and gradient descent prunes the model's capacity for excluded output domains. The resulting model becomes stably, confidently wrong about what the training signal excludes, and it cannot detect this from within — Theorem T-12a applies.

The implication for AI deployment as a substrate fidelity check is critical: an AI trained on a homogeneous or curated corpus and deployed as an "independent" check on a human codec fed by the same information environment creates *correlated sensors masquerading as independent ones.* The channel diversity is illusory. The Substrate Fidelity Condition (δ-independence) must be verified at the level of training data provenance, not merely at the level of institutional separation.

---

§7. Scope and Limitations

7.1 Conditional on T9-3/T9-4 and the Stability Filter

The entire argument depends on the MDL pruning equations being the correct description of the Maintenance Cycle's pruning pass. If biological pruning operates by a different mechanism — one that preserves "emergency" capacity for unused modalities — the irreversibility claim (Theorem T-12) would be weakened but not eliminated: the self-reinforcement loop (Section 3.3) remains valid as long as *any* capacity reduction occurs under disuse.

7.2 τ_{prune} Is Unbounded

As with Action-Drift (Appendix T-13, §7.5), the timescale of capacity loss is identified but not quantitatively bounded. For biological codecs, τ_{prune} is likely on the order of days to weeks for specific skills, months to years for deep perceptual categories, and generational for civilisational codecs.

7.3 The Defence Is Structural, Not Guaranteed

The Substrate Fidelity Condition (T-12b) provides a necessary structural defence but does not guarantee fidelity. A codec that has δ-independent channels may still fail to attend to them, fail to integrate their signal, or prune the attention capacity despite the available input. The condition is necessary but not sufficient — the codec must also maintain the *comparator architecture* that evaluates cross-channel discrepancy.

7.4 Does Not Solve the Meta-Problem

T-12a establishes that a fully adapted codec cannot detect its own corruption. The meta-problem — how does an observer *already* in Narrative Drift recover? — is not solved by this appendix. The ethics paper's answer (Section V.3a) is institutional: only external comparators operating between codecs can force the disconfirming signal back across the Markov blanket. This is structurally sound but ethically difficult: it requires trusting an external source that the corrupted codec will necessarily experience as hostile noise.

---

§8. Closure Summary

T-12 Deliverables

1. **Theorem T-12 (Irreversible Capacity Loss).** The MDL pruning pass (T9-3, T9-4) under pre-filtered input $X' = \mathcal{F}(X)$ correctly erases codec components that predict the excluded signal $\mathcal{X}_{\text{excl}}$. The erasure is irreversible and self-reinforcing. → *Closes roadmap criterion (a).*

2. **Theorem T-12a (Undecidability of Input Provenance).** A fully adapted codec cannot distinguish filtered from unfiltered input. The instru-

ment of detection has been shaped by the same filter that produced the corruption. $\rightarrow$ *Closes roadmap criterion (c).*

3. **Theorem T-12b (Substrate Fidelity Condition).** δ-independent input channels are necessary and sufficient to protect against Narrative Drift. The cross-channel comparison signal ε_{12} interrupts the self-reinforcing pruning loop. $\rightarrow$ *Closes roadmap criterion (b).*

4. **§6.3–6.4: Civilisational and AI Consequences.** The authoritarian pattern is characterised as deliberate channel reduction; RLHF is structurally equivalent to the pre-filter operator. $\rightarrow$ *Supports roadmap criterion (d) (already addressed in ethics paper Section V.5).*

Remaining open items

- $\tau_{\mathbf{prune}}$ **bound.** Quantitative bounding of the capacity loss timescale from empirical data.
- $\delta_{\min}$ **characterisation.** The codec's minimum discrimination threshold for cross-channel discrepancy has not been bounded.
- **Recovery dynamics.** The formal analysis of how a codec in deep Narrative Drift can recover — if it can — awaits treatment.
- **Interaction with T-13 (Action-Drift).** Action-Drift is a special case of T-12 where the pruned capacity is behavioural rather than perceptual. The formal integration is acknowledged (T-13 §6.4) but not fully developed.

This appendix is maintained alongside theoretical_roadmap.pdf. References: T9-3/T9-4 (preprint Section 3.6.3), Stability Filter (preprint Section 3.3), Narrative Drift (preprint Section 3.3, Survivors Watch Ethics Section V.3a), Comparator Hierarchy (Survivors Watch Ethics Section V.3a), Corruption Criterion (Survivors Watch Ethics Section V.5), Action-Drift (Appendix T-13, §6).