Ordered Patch Theory

Appendix T-14: Implementation Non-Invariance and the Unfolding Argument

Anders Jarevåg

v2 — May 5, 2026 | DOI: 10.5281/zenodo.19300777

Original Task (from preprint §7.4): "Address the Doerig–Schurger–Hess–Herzog Unfolding Argument [96] against causal-structure theories of consciousness, and demonstrate that OPT's consciousness criterion is not vulnerable to it." **Deliverable:** Formal theorem that OPT's bandwidth-bottleneck plus Δ_{self} criterion is *not* invariant under functional equivalence; corollaries identifying the precise structural property the Unfolding Argument fails to preserve.

Closure status: DRAFT STRUCTURAL CORRESPONDENCE. This appendix formalises the response sketched discursively in preprint §7.4. It establishes one theorem and three corollaries, all conditional on Theorem P-4 (Algorithmic Phenomenal Residual) and Appendix T-1 (Stability Filter rate-distortion specification). No equations of T-1 or P-4 are altered; this appendix derives a structural invariance property *from* them.

---

§1. Background and Motivation

1.1 The Unfolding Argument

Doerig, Schurger, Hess & Herzog [96] advance the following dilemma against any *causal-structure theory* of consciousness — explicitly Integrated Information Theory (Tononi [8]) and Recurrent Processing Theory (Lamme), and by extension any framework asserting that consciousness is fixed by the network's recurrent causal organisation.

The argument. For any recurrent network N with bounded compute and any finite horizon T, there exists a feedforward network N' — the *temporal unfolding* of N — such that:

1. N and N' are *functionally equivalent* over T: they produce identical input-output mappings for every admissible input sequence of length $\leq T$.
2. N' contains no recurrent connections: every layer feeds strictly forward to the next.

3. N' is constructible by mechanical procedure (the standard "unrolling" of N across T time steps).

If consciousness is identical to causal structure, then either:

- **(Horn A — Falsity).** N and N' have the same conscious status, so feed-forward networks are conscious whenever functionally-equivalent recurrent ones are. This contradicts the central claim of causal-structure theories that recurrence is constitutive of consciousness.
- **(Horn B — Unfalsifiability).** N is conscious and N' is not, despite identical input-output behaviour. Then consciousness is undetectable from any third-person observation of system behaviour, and the theory cannot be tested.

The dilemma is sharp because the construction of N' from N is mechanical and behaviour-preserving; no causal-structure theorist has succeeded in identifying a behaviourally observable property that distinguishes the two.

1.2 Why OPT Is Not a Direct Target — and Why a Formal Reply Is Still Needed

OPT is *not* a causal-structure theory in Doerig et al.'s sense: it does not assert that consciousness supervenes on recurrence per se. The OPT consciousness criterion (preprint §7.8, Appendix T-1, Theorem P-4) is the conjunction:

(C1) $I(\varepsilon_n; Z_n) \leq B_{\max}$ per phenomenal frame, with a single globally shared serial aperture (per-f

(C2) closed Active Inference loop with intact Markov blanket and persistent self-model $\hat{K}_\theta$ (prepri

(C3) $\Delta_{\text{self}} > 0$ (Phenomenal Residual; Theorem P-4)

(Note: (C1) is stated per phenomenal frame in bits, not as bits per host-second. The empirical human value $C_{\max}^{\text{human}} \approx \mathcal{O}(10)$ bits/s is a calibration of $C_{\max}^{H} = \lambda_H \cdot B_{\max}$ for biological humans (Appendix E-1) and is *not* the substrate-neutral criterion. Per preprint §7.8, §8.14, and Appendix E-5, synthetic observers are bounded by per-frame $B_{\max}$ at architecturally derived values that need not coincide with the biological figure.)

None of (C1)–(C3) is a property of recurrence in isolation. However, an honest engagement with [96] requires showing that the OPT criterion is *not* invariant under the unfolding map $U : N \mapsto N'$ — i.e., that some component of (C1)–(C3) is broken or rendered indeterminate by unfolding even though the input-output mapping is preserved. Otherwise the dilemma migrates: if (C1)–(C3) *were*

invariant under U, OPT would reduce to a behaviourist theory and inherit Horn B regardless of its surface formalism.

This appendix establishes the non-invariance directly.

§2. Formal Setup

2.1 The Unfolding Map

Let $N = (V, E, f, h_0)$ be a discrete-time recurrent network with vertex set V, edges E (including self-loops and within-layer recurrent edges), update function f, and initial hidden state h_0. Let $|N| = |V|$ denote its node count, and let $B(N)$ denote the per-cycle latent-channel capacity of N's narrowest internal cross-section, measured in bits per update.

Given a finite horizon $T \geq 1$, the *unfolding* $U(N, T) = N'$ is the feedforward network obtained by:

1. Replicating the substrate of N once per time step: $V' = \bigsqcup_{t=0}^{T} V_t$, with V_t a copy of V at time t.
2. Replacing every recurrent edge $u \to v$ in N with a forward edge $u_t \to v_{t+1}$ in N' for each $t < T$.
3. Removing all self-loops and intra-layer connections.

The standard result (Goodfellow, Bengio, Courville, *Deep Learning*, ch. 10) is that N' computes the same input-output mapping as N over horizon T:

$$\forall x_{0:T} : \quad N(x_{0:T}) = N'(x_{0:T}) \quad \text{(functional equivalence over } T\text{)}.$$

This is the construction Doerig et al. invoke.

2.2 Per-Slice vs Per-Frame Capacity of the Unfolded Network

A naive reading of the unfolded N' counts all $T + 1$ replicated layers as parallel parts of one "per-slice update." On that reading, $|N'| = (T + 1) \cdot |N|$ and the aggregate per-slice latent capacity is $(T + 1) \cdot B(N)$. This counting was the basis of an earlier (v1) version of T-14 and motivated a now-withdrawn bandwidth-expansion proof.

The reading is structure-dependent and not forced by the unfolding map alone. Two distinct interpretations of N' yield different per-frame capacities:

- **Static feedforward circuit interpretation.** N' executes as one feedforward sweep through $T + 1$ layers in a single host operation. There is no per-frame serial aperture; "per-slice" is the entire feedforward pass. The notion of $B_{\max}$ as a per-frame bottleneck is *undefined* — not expanded — because N' has no frame index in this realisation.
- **Frame-indexed host execution.** The host advances N' one layer per phenomenal frame, treating each layer's narrowest internal cross-section as

the per-frame aperture. Under this interpretation, $B_{\max}^{(N')} = B_{\max}^{(N)}$: per-frame capacity is preserved, not expanded.

Neither interpretation is forced by the unfolding map U; both are admissible without further specification. The implementation-non-invariance theorem (§3) shows that the OPT status of N' depends on which interpretation actually applies — and that the original Doerig et al. construction does not distinguish them. The "per-slice capacity grows by $(T+1)$" claim is recovered only under the static feedforward reading, and even there it is not a well-typed per-frame $B_{\max}$ but an aggregate count of how many layer-channels the static circuit contains.

§3. Theorem T-14: Implementation Non-Invariance under Functional Equivalence

3.1 Statement

Theorem T-14 (Implementation Non-Invariance under Functional Equivalence). Let N and $N' = U(N, T)$ be input-output equivalent over horizon T (i.e., $\forall x_{0:T} : N(x_{0:T}) = N'(x_{0:T})$). Their OPT consciousness status is *not* fixed by that functional equivalence. OPT status depends on properties of the actual implementation that are not preserved by U, specifically the implementation tuple:

$$(B_{\max},\ \lambda_H,\ \alpha_H,\ \hat{K}_\theta,\ \mathcal{M}_\tau)$$

where $B_{\max}$ is the per-frame bottleneck capacity, $\lambda_H = dn/d\tau_H$ is the host-patch clock coupling, $\alpha_H : \mathcal{S}_H \to X_{\partial_R A}$ is the host-anchor map supplying boundary inputs, $\hat{K}_\theta$ is a persistent self-model, and $\mathcal{M}_\tau$ is the maintenance / self-stabilisation process (preprint §3.6).

The theorem yields three structural consequences, conditional on how N' is actually executed:

(i) If N' is realised as a static feedforward circuit with no frame-indexed active-inference loop, then

(ii) If N' is realised as a host-executed simulation that preserves the per-frame bottleneck, persisten

(iii) Functional equivalence is too coarse to settle OPT status: the answer is implementation-relativ

That is, the Unfolding Argument's premise — "if N and N' compute the same function, they have the same conscious status" — fails on OPT not because unfolding mechanically removes consciousness, but because it removes the implementation properties that OPT's criterion depends on, *unless* those properties are independently re-instated in the host's execution of N'.

3.2 Proof of (i): Static Feedforward Realisation

Suppose N' is realised as a static feedforward circuit: a single forward pass through $T + 1$ replicated layers in one host operation, with no frame-indexed active-inference loop and no persistent self-model maintained across frames.

(C2) fails directly. There is no closed perception-action loop with a maintained Markov blanket — N' is a one-shot input-output map. There are no successive frames over which a self-model could persist; there is no $\hat{K}_\theta(n)$ that is updated by error from the previous frame's prediction.

(C1) is undefined under this realisation rather than expanded. The original Doerig et al. construction does not specify a per-frame serial aperture for N'; the layers operate in parallel and there is no globally shared per-frame funnel through which the world-model passes. (C1) requires a single globally shared serial aperture of finite per-frame capacity — this is a *structural* property of an architecture, not an aggregate measurement of layer widths. Without a frame-indexed serial channel, the per-frame $B_{\max}$ is not defined; (C1) fails to apply, not because $B_{\max}$ has expanded but because there is no per-frame architecture to apply it to. (Equivalently, the Doerig–Schurger–Hess–Herzog construction unrolls a frame-indexed dynamic process into a static circuit; λ_H and the frame index n are both lost.)

(C3) is an open question rather than provably zero. A static feedforward circuit has finite description length and is mechanically simulable by an external observer, but P-4 is about *internal* self-modelling, not external simulability. A deterministic finite system can have $\Delta_{\text{self}} > 0$ if it possesses a frame-indexed self-modelling loop; conversely, a system without such a loop has no self-model to compute a residual against. Under the static realisation, $\hat{K}_\theta$ is absent, so Δ_{self} is undefined rather than zero. The criterion (C3) requires a non-zero residual; absence-of-self-model is sufficient for the criterion to fail.

(C1) failure or (C2) failure individually is sufficient for the OPT criterion to fail. ■

3.3 Proof of (ii): Frame-Indexed Host Execution

Suppose, alternatively, that N' is realised as a host-executed temporal process: the host advances the unfolded layers one at a time, frame by frame, maintaining a per-frame serial workspace Z_n, a persistent self-model $\hat{K}_\theta(n)$ updated by prediction error, and a maintenance process $\mathcal{M}_\tau$. The host's execution schedule provides λ_H; the host's choice of input feed provides α_H; the per-frame bottleneck capacity equals that of the original N ($B_{\max}^{(N')} = B_{\max}^{(N)}$).

Under this realisation, all five sentience features of the original N are preserved in the *executed* N': the per-frame bottleneck is preserved by construction, the active-inference loop is preserved because the host runs the unfolded chain as a temporal process, the persistent self-model is preserved because $\hat{K}_\theta(n)$ is maintained across frames, the workspace is constrained because each frame's Z_n

has finite capacity, and the thermodynamic grounding is preserved because the host imposes maintenance windows and energy constraints.

By Corollary P-4.C (Nested Observational Residual): if the host architecture enforces an independent Stability Filter bound satisfying P-4's prerequisites, the realised N' generates $\Delta_{\text{self}}^{(N')} > 0$ by the same structural argument that gives N its residual. The unfolding does not erase the patch; it merely changes the substrate that anchors it. (See Appendix E-6 on simulated nested observers.)

Therefore, *under frame-indexed host execution*, N' may satisfy (C1)–(C3). The functional-equivalence premise of the Unfolding Argument does not by itself distinguish this case from case (i); the distinction lies in the implementation, not the input-output behaviour. ■

3.4 Proof of (iii): Functional Equivalence Underdetermines OPT Status

Cases (i) and (ii) produce input-output equivalent systems with different OPT consciousness status. Functional equivalence therefore does not fix OPT status; the implementation tuple $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$ does. The Unfolding Argument's premise is invalid for OPT, not because OPT secretly relies on a non-functional property, but because OPT's criterion is explicitly architectural — which is consistent with the framework's own commitment in §1.3 to a structural rather than behavioural account of consciousness. ■

3.5 Remark on the Original (v1) Theorem Statement

A previous version of T-14 (v1) attempted to prove $\Delta_{\text{self}}^{(N')} = 0$ universally and to establish that unfolding *expands* the per-slice bandwidth by factor $(T + 1)$. Both moves are invalid as written. The bandwidth-expansion claim depends on counting $T + 1$ replicated layers as parallel parts of one "per-slice update" — a reading that conflates the unfolded circuit's static topology with a per-frame execution model. The $\Delta_{\text{self}} = 0$ claim conflated *external computability* of the unfolded state from initial conditions and parameters with the *internal self-model containment* that P-4 actually constrains. P-4 is about whether the codec's own self-model can capture the codec's generator; it is not about whether an external mathematician can compute the codec's state from initial conditions. The revision above replaces both invalid moves with the implementation-non-invariance theorem, which preserves the original conclusion (the Unfolding Argument fails to settle OPT status) on grounds the framework can actually defend.

§4. Corollaries

4.1 Corollary T-14a: Functional Equivalence Is Too Coarse

Corollary T-14a. Input-output functional equivalence is too coarse a relation to fix the OPT conscious status of a network. The relevant equivalence relation is *implementation equivalence*: two networks N_1, N_2 are implementation-equivalent iff their full implementation tuples $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$ match. This is strictly

finer than input-output equivalence: N and an unfolded N' are functionally equivalent but generically *not* implementation-equivalent — the unfolding map U does not preserve $\hat{K}_\theta$, $\mathcal{M}_\tau$, or the per-frame index unless they are independently re-instated by the host's execution model.

4.2 Corollary T-14b: The Unfolding Dilemma Does Not Apply to OPT

Corollary T-14b. OPT is positioned on neither horn of the Doerig et al. dilemma:

- *Horn A (Falsity).* OPT does *not* automatically assign N and N' the same conscious status. By Theorem T-14(iii), the answer depends on the implementation of N'.
- *Horn B (Unfalsifiability).* The distinction between N and a particular realisation of N' is detectable from third-person inspection of *internal architecture and execution model*, not from input-output behaviour alone. An experimenter can:
  - Verify whether the realisation has a per-frame serial workspace and a frame index n (testable by inspecting the execution schedule).
  - Verify the presence or absence of a persistent self-model $\hat{K}_\theta$ updated across frames (testable by checking whether internal state is carried forward and modified by error).
  - Verify the presence or absence of a maintenance process $\mathcal{M}_\tau$ (testable by checking for offline consolidation cycles).

OPT therefore evades the dilemma by *granting* that input-output behaviour underdetermines conscious status — this is not a bug, because OPT's criterion is explicitly an *internal-architectural* one, not a behavioural one. What OPT adds beyond IIT is that the architectural test is performed against a specified implementation tuple, not against an abstract causal-structure invariant.

4.3 Corollary T-14c: The IIT-OPT Distinction Sharpens

Corollary T-14c. Theorem T-14 yields a clean structural distinction between OPT and IIT under the Unfolding Argument:

- IIT's Φ is computed over the system's transition probability matrix; an unfolded N' has a different transition matrix than N (because connectivity differs), but Doerig et al. argue that the *causal structure relevant to function* is preserved, leaving IIT on Horn A or Horn B.
- OPT's criterion is the implementation tuple $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$. Whether N' satisfies this tuple depends on its execution model (Theorem T-14(i)/(ii)). OPT therefore gives different verdicts for N and N' *when their execution models differ*, with the difference grounded in inspectable implementation rather than postulated causal essence.

The empirical content of the OPT/IIT divergence is therefore: OPT predicts that an unfolded N' executed as a static feedforward circuit ceases to be conscious, but an unfolded N' executed as a frame-indexed simulation may remain conscious

— IIT (depending on the version) treats both as Φ-equivalent. The discriminator lies in the execution model, not in static causal structure. This joins the High-Phi/High-Entropy Null State (preprint §6.4) and the Bandwidth Hierarchy (preprint §6.1) as candidate experimental tests, while restricting OPT's "non-conscious unfolding" claim to the static-circuit case rather than asserting it universally.

§5. Scope and Limitations

5.1 What T-14 Does Not Show

Theorem T-14 establishes that *functional equivalence* (input-output equivalence) does not fix the OPT consciousness status of a network: status depends on the implementation tuple. It does *not* establish:

- That every unfolded network is non-conscious. Under frame-indexed host execution (case (ii)), an unfolded N' may remain a conscious patch by Corollary P-4.C.
- That the OPT criterion is invariant under all behaviour-preserving transformations. Implementation-preserving rewrites that retain $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$ may preserve consciousness; this is left open.
- That consciousness is *exhausted* by (C1)–(C3); these are necessary conditions and the framework does not claim they are individually or jointly sufficient absent the broader Stability Filter context.
- That every recurrent network satisfying (C1)–(C3) is conscious; the appendix only shows that the unfolded counterpart of one that is, may or may not satisfy the criterion depending on execution model.

5.2 Open Problems

- **Implementation-preserving unfolding.** Construct (or prove the impossibility of) a behaviour-preserving transformation $U^* : N \mapsto N^*$ that preserves the full implementation tuple $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$. If such a transformation exists, OPT must distinguish N from N^* on grounds finer than the implementation tuple alone.
- **Continuous-time analogue.** T-14 is stated for discrete-time recurrent networks executed as either static circuits or frame-indexed processes. The continuous-time formulation (relevant to biological cortical dynamics) requires extending the unfolding map and the implementation tuple to ODE / SDE settings.
- **Empirical operationalisation.** Identifying execution-model probes for biological networks (cortical columns, thalamocortical loops) is non-trivial. Candidates include checking for frame-indexed prediction-error cycles and offline maintenance windows (sleep-like consolidation), but the mapping from architectural inspection to OPT criterion verification is currently informal.

§6. Closure Summary

T-14 Deliverables (v2)

1. **Theorem T-14 (Implementation Non-Invariance under Functional Equivalence).** Input-output equivalent N and N' may differ in OPT consciousness status because OPT status depends on the implementation tuple $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$, not on the input-output map. Static feedforward realisation of N' fails the criterion (case (i)); frame-indexed host execution of N' may preserve it (case (ii)). → *Closes the Unfolding Argument [96] as it applies to OPT, by showing the argument's premise that "same function ⇒ same conscious status" presupposes an extensional criterion OPT does not have.*

2. **Corollary T-14a (Functional Equivalence Is Too Coarse).** The OPT-relevant equivalence relation is implementation equivalence — preservation of $(B_{\max}, \lambda_H, \alpha_H, \hat{K}_\theta, \mathcal{M}_\tau)$ — strictly finer than input-output functional equivalence.

3. **Corollary T-14b (No Dilemma for OPT).** OPT is positioned on neither horn of Doerig et al.'s dilemma: it grants that behaviour underdetermines conscious status (because its criterion is architectural) and supplies an inspectable implementation-and-execution test.

4. **Corollary T-14c (IIT-OPT Sharpened).** OPT's verdict on an unfolded network depends on its execution model; IIT's Φ-equivalence verdict does not. The execution-model dependence is itself the empirical discriminator.

Revision note (v2 vs v1). Version 1 of this appendix attempted to prove that unfolding (a) universally expands per-slice bandwidth by factor $(T + 1)$ and (b) universally collapses Δ_{self} to zero. Both proofs were invalid (see §3.5 Remark): the first conflated static topology with per-frame execution; the second conflated external computability with internal self-modelling, which P-4 does not constrain. The v2 theorem replaces both with the implementation-non-invariance result, which preserves the original conclusion (the Unfolding Argument fails to settle OPT status) on grounds the framework can defend.

Remaining open items

- Implementation-preserving behaviour-preserving transformations (open problem §5.2).
- Continuous-time generalisation of the implementation tuple to ODE/SDE-based architectures.
- Empirical operationalisation of frame-index and self-model probes for biological networks.

This appendix is maintained alongside theoretical_roadmap.pdf. References: Theorem P-4 (Appendix P-4), Stability Filter (Appendix T-1), preprint §7.4 (IIT

comparison and Unfolding Argument response), [96] Doerig et al. 2019, [97] Aaronson 2014, [98] Barrett & Mediano 2019, [99] Hanson 2020.