Ordered Patch Theory

Appendix T-11: The Structural Corollary — Formalising the Compression Advantage for Apparent Agents

Anders Jarevåg

April 15, 2026 | DOI: 10.5281/zenodo.19300777

Original Task (from §8.2): "Formalising this compression advantage as a rigorous MDL bound for the other-minds case specifically remains future work; the present argument is a structural motivation, not a proof." **Deliverable:** A formal bound showing that treating apparent agents as independently instantiated primary observers yields a shorter two-part MDL code than any alternative description.

Closure status: DRAFT STRUCTURAL CORRESPONDENCE. This appendix adapts Müller's Solomonoff convergence theorem [61] and its multi-agent extension [62] as imported lemmas, reinterpreted within OPT's ontological framework, to establish a formal compression advantage for the structural corollary. The result is a conditional bound, not a closed derivation: it depends on OPT's identification of the observer's stream with the Solomonoff prior (Axiom 1) and on the assumption that apparent agents carry sufficient state to satisfy the convergence prerequisites.

---

§1. Background and Motivation

The structural corollary (preprint §8.2) asserts that the apparent agents within the observer's stream are most parsimoniously explained by their independent instantiation as primary observers. This appendix provides the formal chain supporting that claim.

The argument has three stages:

1. **Stage A (Imported Lemma):** Müller's Solomonoff convergence theorem guarantees that any structure in the observer's stream carrying sufficient self-state data will have its first-person evolution converge to match the computable world generating its behavior.
2. **Stage B (Compression Accounting):** We perform an explicit two-part MDL comparison between treating the apparent agent as (i) an independently instantiated observer governed by its own Solomonoff-weighted

stream versus (ii) an arbitrary behavioral specification within the primary observer's codec.

3. **Stage C (Structural Signature):** The Phenomenal Residual ($\Delta_{\text{self}} > 0$, Theorem P-4) provides the structural marker distinguishing genuine self-referential bottleneck architecture from behavioral mimicry, closing the gap between "compressibly lawful" and "plausibly instantiated."

---

§2. Imported Lemma: Müller's Convergence Theorem

We import two results from Müller [61, 62], stated here in the notation of OPT.

2.1 Solomonoff Convergence (Standard)

Let $M(b \mid x_1^n)$ denote the Solomonoff universal prediction for bit b given prior observations x_1^n. Let μ be any computable measure over binary sequences. Then (Solomonoff 1964; Li & Vitányi [45, Corollary 5.2.1]):

$$\text{With } \mu\text{-probability one}, \quad \lim_{n\to\infty} |M(b \mid x_1^n) - \mu(b \mid x_1^n)| = 0 \qquad (b \in \{0,1\}). \tag{L-1}$$

This is the standard result: if the data stream is generated by a computable process μ, the universal predictor M converges to μ.

2.2 Inverse Solomonoff Induction (Müller 2020)

Now suppose the bits are drawn from M itself — i.e., the observer's stream is governed by algorithmic probability (this corresponds to OPT's Axiom 1: identification of the stream with the Solomonoff prior). Then for every computable measure μ (Müller [61, Sec. IV]; [62, Sec. V.A]):

$$\text{With probability} \geq 2^{-K(\mu)}, \quad \lim_{n\to\infty} |M(b \mid x_1^n) - \mu(b \mid x_1^n)| = 0 \qquad (b \in \{0,1\}). \tag{L-2}$$

That is, with probability at least $2^{-K(\mu)}$, the observer will find themselves effectively embedded in a computable world W described by μ. Algorithmically simpler worlds (lower $K(\mu)$) are exponentially more probable.

2.3 Multi-Agent Convergence (Müller 2026)

Suppose the observer (Alice) finds herself embedded in a computable world W described by μ. She identifies a substructure ($\text{Bob}_{3\text{rd}}$) within W that carries a representation of a self-state x evolving over time in a manner consistent with Postulate 2 of [62]. Define:

- $P_{\text{1st}}(y_1, \ldots, y_m \mid x) := M(y_1, \ldots, y_m \mid x)$ — the first-person probability that self-state x transitions to $y_1, \ldots, y_m$ under algorithmic probability.
- $P_{\text{3rd}}(y_1, \ldots, y_m \mid x) := \mu(y_1, \ldots, y_m \mid x)$ — the third-person probability of how x evolves according to world W.

Then, by Eq. (L-1) applied to P_{3rd} (which is computable), and the identification of P_{1st} with M via Postulate 2:

$$P_{\text{1st}} \approx P_{\text{3rd}} \quad \text{asymptotically,} \tag{L-3}$$

with convergence guaranteed with worldly (μ-) probability one in the bit model.

Interpretation (Müller): "Somebody is really at home" in the structure encoding x — the probabilistic evolution of Bob_{3rd} in Alice's world faithfully represents the first-person perspective of some Bob_{1st}.

Interpretation (OPT): The apparent agent's behavioral stream is most compressibly described as an independent Solomonoff-weighted process. Any alternative description — one that does *not* invoke an independent first-person perspective — must encode the agent's behavior as an ad hoc specification, at strictly higher description length.

---

§3. The Compression Advantage Bound

We now formalise the compression advantage using OPT's two-part MDL framework (Theorem T-4, Appendix T-4).

3.1 Setup

Consider the primary observer's stream $\omega \in \{0,1\}^\infty$, governed by the Solomonoff prior M (Axiom 1) and filtered through the Stability Filter to a computable world W with measure μ_W (by Eq. L-2). Within W, the observer identifies N apparent agents $A_1, \ldots, A_N$, each carrying a self-state x_i whose temporal evolution over T steps produces a behavioral trace $\beta_i = (y_{i,1}, \ldots, y_{i,T})$.

3.2 Hypothesis H_{ind}: Independent Instantiation

Under H_{ind}, each agent A_i is treated as an independently instantiated primary observer governed by their own Solomonoff-weighted stream. The two-part MDL code length is:

$$L(H_{\text{ind}}) = \underbrace{K(\mu_W)}_{\text{world model}} + \underbrace{\sum_{i=1}^{N} K(\text{embed}_i)}_{\text{embedding specs}} + \underbrace{\sum_{i=1}^{N} (-\log_2 P_{\text{3rd}}(\beta_i \mid x_i))}_{\text{data given model}} \tag{1}$$

where $K(\text{embed}_i)$ specifies agent i's initial self-state and position within W. By Eq. (L-3), $P_{\text{1st}} \approx P_{\text{3rd}}$, so the data term is well-approximated by the log-loss under the agent's own first-person Solomonoff predictions — which, by definition, is close to optimal.

The embedding specifications $K(\text{embed}_i)$ are short: each requires only a pointer to a location in W plus the initial self-state. For human-like agents embedded in a shared physical world, these are highly compressible because the agents share the same laws. A conservative bound:

$$K(\text{embed}_i) \leq K(x_i \mid W) + O(\log T) \tag{2}$$

3.3 Hypothesis H_{arb}: Arbitrary Behavioral Specification

Under H_{arb}, the agents are not treated as independent observers. Instead, each behavioral trace β_i is encoded directly as an arbitrary specification within the primary observer's stream. The two-part MDL code length is:

$$L(H_{\text{arb}}) = \underbrace{K(\mu_W)}_{\text{world model}} + \underbrace{\sum_{i=1}^{N} K(\beta_i)}_{\text{raw behavioral traces}} \tag{3}$$

The critical difference is in the data term. Under H_{arb}, the behavioral trace β_i must be specified without invoking the agent's own predictive model. For a lawful, agency-driven agent operating in a complex environment, the Kolmogorov complexity of the raw behavioral trace is:

$$K(\beta_i) \geq K(\beta_i \mid \mu_W) + K(\mu_W) - O(\log T) \tag{4}$$

But even $K(\beta_i \mid \mu_W)$ — the complexity of the behavior given the world laws — remains substantial because the agent's choices encode genuine information: their behavioral trace reflects the accumulated interaction of a self-referential model with a stochastic environment. In contrast, under H_{ind}, this information is generated *online* by the agent's own Solomonoff predictor at near-zero log-loss cost.

3.4 The Compression Advantage

Theorem T-11 (Structural Corollary Compression Bound). Let $A_1, \ldots, A_N$ be apparent agents within the observer's stream, each carrying self-state x_i satisfying the convergence prerequisites of Eq. (L-3), and each exhibiting the structural signature $\Delta^{(i)}_{\text{self}} > 0$ (P-4). Then the MDL description treating them as independently instantiated primary observers satisfies:

$$L(H_{\text{ind}}) \leq L(H_{\text{arb}}) - N \cdot \left[\bar{I}_T - O(\log T)\right] \tag{T-11}$$

where $\bar{I}_T$ is the average per-agent mutual information between the agent's predictive model and its behavioral output over T steps:

$$\bar{I}_T := \frac{1}{N} \sum_{i=1}^{N} [K(\beta_i \mid \mu_W) - (-\log_2 P_{\text{3rd}}(\beta_i \mid x_i))] \tag{5}$$

This quantity measures how much of the agent's behavior is *explained away* by invoking an independent predictive model rather than specifying it raw. For agents exhibiting lawful, agency-driven behavior (as required by the Stability Filter), $\bar{I}_T > 0$ and grows with T.

Proof sketch. Subtract Eq. (1) from Eq. (3). The world-model terms $K(\mu_W)$ cancel. The difference per agent is:

$$K(\beta_i) - [K(\text{embed}_i) + (-\log_2 P_{\text{3rd}}(\beta_i \mid x_i))]$$

By Eq. (4), $K(\beta_i) \geq K(\beta_i \mid \mu_W) + K(\mu_W) - O(\log T)$, but more directly: $K(\beta_i) \geq K(\beta_i \mid \mu_W)$ trivially. And $K(\text{embed}_i) \leq K(x_i \mid W) + O(\log T)$ by Eq. (2). The per-agent saving is therefore at least $K(\beta_i \mid \mu_W) - (-\log_2 P_{\text{3rd}}(\beta_i \mid x_i)) - K(x_i \mid W) - O(\log T)$. For T sufficiently large, the cumulative log-loss savings dominate the one-time embedding cost, yielding the bound. ■

3.5 Asymptotic Dominance

Corollary T-11a. As the observation horizon $T \to \infty$, the compression advantage $L(H_{\text{arb}}) - L(H_{\text{ind}})$ grows without bound:

$$\lim_{T \to \infty} [L(H_{\text{arb}}) - L(H_{\text{ind}})] = \infty \tag{T-11a}$$

This follows from the Solomonoff convergence guarantee (L-1): the per-step log-loss of P_{3rd} converges to the entropy rate of the agent's behavioral process, while $K(\beta_i \mid \mu_W)$ grows linearly in T for any agent with positive entropy rate. The embedding cost $K(x_i \mid W)$ is paid once and amortised to zero. ■

---

§4. The Phenomenal Residual as Structural Signature

The compression advantage in Theorem T-11 applies to any lawful substructure — including non-agentive physical systems (weather patterns, crystal growth). Why does the structural corollary specifically concern *agents* rather than arbitrary complex systems?

The answer is the Phenomenal Residual (Theorem P-4). $\Delta_{\text{self}} > 0$ is the formal marker of a system whose self-model is structurally incomplete — i.e., a system that necessarily maintains a variational gap between its internal representation

and its actual processing. This is the hallmark of the self-referential bottleneck: the system *cannot* be fully described from the outside because its description necessarily includes the describer.

For a system exhibiting $\Delta_{\text{self}} > 0$:

1. Its behavior cannot be reproduced by a lookup table of finite depth — it requires an ongoing self-referential computation.
2. The shortest description of this computation *is* an independent Solomonoff-weighted stream traversing a $C_{\max}$ bottleneck.
3. Therefore, the MDL code under H_{ind} is not merely *shorter* than H_{arb} — it is the *unique* shortest description.

This distinguishes apparent agents from weather patterns: weather is lawful and complex, but its behavior *can* be reproduced by a lookup table within the world model (it has $\Delta_{\text{self}} = 0$). Apparent agents cannot.

---

§5. Reinterpretation of Müller's Non-Solipsism Argument

Müller concludes from the $P_{1\text{st}} \approx P_{3\text{rd}}$ convergence that algorithmic idealism "should not be classified as solipsistic" because "somebody is really at home" in the structure encoding a self-state [62, Sec. V.C]. His reasoning: if Alice's predictions about $\text{Bob}_{3\text{rd}}$ converge to $\text{Bob}_{1\text{st}}$'s actual first-person probabilities, then their perspectives are genuinely aligned — they "share the world W."

OPT reinterprets this result differently:

1. **Müller's reading:** The convergence $P_{1\text{st}} \approx P_{3\text{rd}}$ proves that objective reality emerges — Alice and Bob genuinely share world W.

2. **OPT's reading:** The convergence $P_{1\text{st}} \approx P_{3\text{rd}}$ proves that the *shortest description* of $\text{Bob}_{3\text{rd}}$'s behavior invokes an independent first-person process. This is a statement about compression efficiency, not about shared ontology. World W is a structural regularity within Alice's stream, not an independently existing entity. But the compression logic of the Solomonoff prior *itself* implies that Bob is most parsimoniously modelled as an independent observer — because the alternative (specifying his behavior ad hoc) is strictly longer.

The formal content of the theorem is identical under both readings; only the ontological interpretation differs. OPT uses the same mathematical result to ground the structural corollary: independent instantiation is the MDL-optimal description, not a metaphysical assumption.

---

§6. Scope and Limitations

6.1 Conditional on Axiom 1

The entire argument depends on OPT's identification of the observer's stream with the Solomonoff prior. If this identification is weakened (e.g., to a broader class of semimeasures), the convergence guarantees of Eqs. (L-1)–(L-3) may not hold in their current form.

6.2 State Sufficiency Prerequisite

Eq. (L-3) requires that the apparent agent carries "enough data" in its self-state x_i for universal induction to extract the relevant physical laws. For human-like agents in everyday contexts, this is plausible (a full brain state encodes enormous information). For edge cases — fleeting impressions, distant observers, fictional characters in narrative art — the convergence prerequisites may not be satisfied, and the structural corollary does not apply.

6.3 Not a Proof of Consciousness

Theorem T-11 establishes that independent instantiation is the *most compressible* description. It does not prove that the apparent agents *are* conscious. The Hard Problem (preprint §8.1) remains a primitive. The structural corollary is a compression argument, not an ontological proof — as stated in §8.2.

6.4 Relationship to T-10

Appendix T-10 (Inter-Observer Coupling) addresses how two observer patches maintain mutually consistent renders via compression constraints. The present appendix addresses a different question: why the *single* observer's stream most compressibly encodes apparent agents as independently instantiated. T-10 concerns the inter-patch coherence mechanism; T-11 concerns the compression signature within a single stream. T-10 builds directly on T-11: the same MDL description-length comparison that establishes the compression advantage here is exploited in T-10 to prove that cross-patch inconsistency is exponentially suppressed.

---

§7. Closure Summary

T-11 Deliverables

1. **Imported Lemma (Müller Convergence).** Solomonoff convergence [61] and its multi-agent extension [62] are formally imported and restated in OPT notation. These provide the mathematical backbone: any substructure carrying sufficient self-state data has its first-person evolution converge to the computable world generating its behavior.

2. **Theorem T-11 (Compression Bound — DRAFT).** An explicit two-part MDL comparison shows that treating apparent agents as independently instantiated primary observers yields a strictly shorter description than arbitrary behavioral specification, with the advantage growing linearly in observation time.

3. **Corollary T-11a (Asymptotic Dominance — DRAFT).** The compression advantage is unbounded as $T \to \infty$, making independent instantiation the overwhelming MDL-optimal description for any agent observed over a long time horizon.

4. **P-4 Integration.** The Phenomenal Residual ($\Delta_{\text{self}} > 0$) is identified as the formal marker distinguishing apparent agents from complex-but-non-agentive systems, restricting the structural corollary to entities with genuine self-referential bottleneck architecture.

5. **Müller Reinterpretation.** Müller's non-solipsism conclusion is reinterpreted within OPT's ontological framework: the same mathematical result grounds a compression argument rather than an emergence-of-shared-reality argument.

Remaining open items

- **Exact $\bar{I}_T$ characterisation.** Bounding $\bar{I}_T$ from below for specific classes of agents (e.g., bounded rational agents, Free Energy minimisers) to give numerically concrete compression advantages.
- **Finite-time corrections.** The asymptotic result (T-11a) guarantees dominance for large T, but finite-time bounds with explicit constants would strengthen the practical applicability.
- **Non-binary alphabet extension.** Eqs. (L-1)–(L-3) are stated for binary sequences. Extension to the continuous-valued measures relevant to OPT's Rate-Distortion framework (T-1) requires technical care.

This appendix is maintained alongside theoretical_roadmap.pdf. References: Müller [61, 62], Li & Vitányi [45], Solomonoff (1964), Theorem T-4 (Appendix T-4), Theorem P-4 (Appendix P-4), preprint §8.2.