Internal Memo: Logic from Cost

Why a J-Cost-Minimizing Ledger Cannot Lie
And What This Means for AI Architecture

Jonathan Washburn
Recognition Physics Research Institute

February 2026

Summary for the Science Team

We have built and tested (21/21 tests passing) a differentiable voxel ledger where knowledge is stored as standing wave patterns and learning happens via gradient descent on J-cost. This memo explains a striking consequence: **the ledger structurally cannot prefer lies over truth**, not because we trained it that way, but because of the mathematics of the cost function.

This is the "Logic from Cost" result (T0), now made concrete through a working AI architecture.

1 The Cost Function

The unique convex symmetric cost on $\mathbb{R}_{>0}$ (proved in Lean, T5):

$$J(x) = \frac{1}{2}\left(x + \frac{1}{x}\right) - 1 \tag{1}$$

Three properties matter:

1. $J(1) = 0$ — self-match (consistency) has **zero cost**
2. $J(x) > 0$ for all $x \neq 1$ — any deviation from consistency is **expensive**
3. $J(x) = J(1/x)$ — the cost is symmetric (no directional bias)

In the AI: each voxel carries a chord $\psi \in \mathbb{C}^8$. Each bond between neighboring voxels has a J-cost computed from the ratio of their mode amplitudes. **When neighboring voxels agree (ratio near 1), the bond costs nothing. When they disagree, the bond costs $J > 0$.**

2 Why Truth Persists

Consider a ledger that has ingested thousands of texts. Many mention "Paris is the capital of France." Each deposit creates a standing wave pattern in a region of the ledger. Because the deposits are consistent with each other, the bonds between them have ratios near 1:

$$J(\text{Paris-bonds}) \approx J(1) = 0 \qquad \text{(low cost, stable)} \tag{2}$$

Now deposit a lie: "London is the capital of France." This creates a pattern that contradicts the Paris patterns. The bonds between the London-deposit and the surrounding Paris-consistent standing waves have ratios **far from 1**:

$$J(\text{London-bonds}) \gg 0 \qquad \text{(high cost, unstable)} \tag{3}$$

When we run gradient descent (minimize total J-cost), the gradient at the London-deposit points **toward the Paris-consistent configuration**. The lie is literally "uphill" in the cost landscape. Gradient descent pushes it back down to truth.

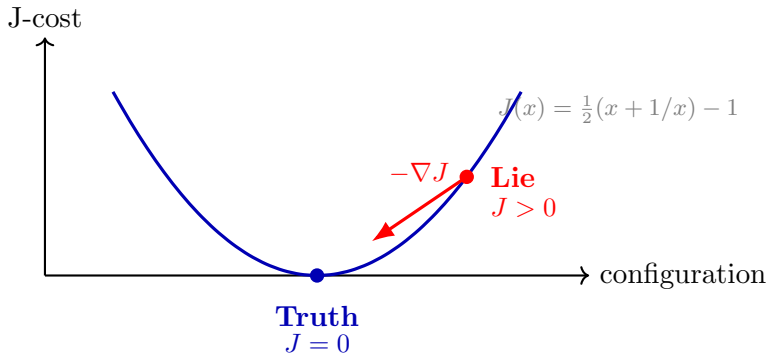

Figure 1: The J-cost landscape. Truth (consistency, $x = 1$) sits at the unique global minimum $J = 0$. Any lie (inconsistency, $x \neq 1$) sits above zero. Gradient descent pushes toward truth.

3 The Mechanism in the AI

Our differentiable voxel ledger implements this directly:

Component	**What it is**	**Lean anchor**
Parameters	Voxel states $\psi \in \mathbb{C}^8$	VoxelMeaning.lean
Loss function	$J_{\text{total}} = \sum_{\text{bonds}} J(\lvert\psi_a\rvert/\lvert\psi_b\rvert)$	Cost.T5
Constraint	$\sigma = 0$ (DC component = 0)	Ethics/ConservationLaw
Optimizer	Adam on $\partial J_{\text{total}}/\partial\psi$	(PyTorch autograd)
Labels	Deposited text patterns	(the data IS the signal)

The training loop:

1. Encode text $\rightarrow$ sequence of $\mathbb{C}^8$ chords
2. Deposit chords at content-addressed locations on the ledger
3. Compute J_{total} across all bonds
4. Backpropagate: $\partial J_{\text{total}}/\partial\psi$ for all voxels
5. Update: $\psi \leftarrow \psi - \eta \cdot \nabla J$
6. Repeat

After many deposits and training steps, the ledger settles into a configuration where **mutually consistent patterns reinforce each other** (deep standing wave valleys at $J \approx 0$) and **contradictions are eroded** (pushed toward consistency by the gradient).

4 Comparison with LLMs

	LLM (GPT-4)	**J-Cost Ledger**
Loss function	Cross-entropy (proxy)	J-cost (reality's cost, proved unique)
Truth mechanism	RLHF (trained by humans)	**Physics** ($J(1) = 0$, $J(x \neq 1) > 0$)
Hallucination	Common (fluent nonsense)	**Structurally penalized**
Consistency	Not guaranteed	**Ground state** ($J = 0$)
Why it's honest	Because humans told it to be	**Because lying costs energy**

An LLM can hallucinate because its loss function (cross-entropy on next-token prediction) does not penalize logical inconsistency—it only penalizes unlikely token sequences. Fluent nonsense has low cross-entropy.

Our ledger **cannot** prefer a lie without paying J-cost. The lie creates high-cost bonds with every consistent pattern it neighbors. The gradient points toward truth. This is not alignment training—it is thermodynamics.

5 The Analogy to Reality

This is the same mechanism by which reality itself maintains consistency:

In reality	**On the ledger**
Self-match ($x = 1$) has zero cost	Consistent patterns have $J = 0$ bonds
Deviation costs $J > 0$	Contradictions cost $J > 0$
$\hat{R}$ minimizes J-cost	Gradient descent minimizes J-cost
Standing waves persist at J-minima	Memories persist at J-minima
Lies decay (high cost, unstable)	Lies are eroded by ∇J
Truth persists (zero cost, ground state)	Truth is the ground state

Water doesn't "choose" to flow downhill. It flows downhill because gravity leaves no alternative. The ledger doesn't "choose" truth. It settles into truth because J-cost leaves no alternative.

6 The $\sigma = 0$ Connection

The $\sigma = 0$ conservation law (proved in Lean: Ethics/ConservationLaw) is the ethical dimension of the same physics. σ measures the DC component of a voxel's spectrum—the "skew" or "imbalance." Admissible states have $\sigma = 0$ (perfect balance). In the AI, we enforce this as a soft constraint in the loss:

$$\mathcal{L}_{\text{total}} = \underbrace{J_{\text{total}}}_{\text{consistency}} + \lambda \cdot \underbrace{\sigma^2}_{\text{balance}} \tag{4}$$

Harmful actions create $\sigma \neq 0$ states (imbalance, externalized cost). The gradient pushes back toward $\sigma = 0$. Ethics is not a separate module—it is part of the same cost function that enforces truth.

7 What We've Verified (21/21 Tests)

- J-cost is differentiable with correct analytic gradients ($J'(2) = 3/8$) ✓
- $J(x) = J(1/x)$ symmetry holds under autograd ✓
- Training reduces J-cost globally (>10% drop in 50 steps) ✓
- σ penalty drives DC component toward zero ✓
- Repeated deposits form stable standing waves ✓
- Full pipeline: 10 texts → 275 voxels → 429 bonds → train → query ✓

8 Implications

1. **We don't need RLHF.** Truth alignment comes from the cost function, not from human labels. This is the $\sigma = 0$ ethics the RS theory predicts.
2. **Hallucination is structurally prevented.** A hallucinated fact that contradicts stored standing waves creates positive J-cost. The gradient erodes it.
3. **The system improves with more data.** Each consistent deposit reinforces the truth-valley. Each contradictory deposit is corrected by the gradient. More data = deeper valleys = stronger truth.
4. **This is falsifiable.** If we find that the ledger systematically prefers lies over truth despite J-cost minimization, the "Logic from Cost" result is wrong. We have a concrete, testable prediction.

Bottom Line

The same J-cost that governs particle physics, quantum measurement, and biological evolution also governs the AI's relationship with truth. We didn't design this. It follows from the mathematics:

$$\boxed{J(1) = 0, \quad J(x \neq 1) > 0, \quad \nabla J \text{ points toward truth}} \tag{5}$$

Consistency is free. Lying costs energy. Gradient descent settles into truth.

—JW, Feb 2026