One Operator, Four Faculties: Value, Time, Self and Contradiction as the Hodge Components of an Agent's Flow, with Belnap's Four Values as Its Regimes

Pierre-Thibault Fabre
Cercle Digital (independent research)
Correspondence: pierre-thibault@cercle.digital

Abstract

Claim. We propose a topological account of an embodied AI agent's memory, in which four faculties, the *value* of a goal, *time*, the *self* as efference, and *contradiction*, are the component types of a single combinatorial Hodge decomposition of the agent's empirical flows, read with no learned parameters. Using frozen adversarial synthetic worlds and the 1977 text adventure *Colossal Cave*, we find that value is the *gradient*, time is the *circulation*, the self is the *exact* part, and contradiction is the *curl*. Belnap's four belief values are the topological regimes of the same flow, dispatched with no hand-tuned threshold. We claim nothing about real perception.

Results. Each identification survives the control built to destroy it, at $\geq$10 seeds with 95% confidence intervals, so the decomposition is measured rather than analogical. The four components then form a single diagnostic surface, on which a reward that cycles, a belief base that cannot be ordered and a memory key that aliases distinct situations appear as one defect at one component, and the operator that detects that defect also names its repair. Potential-based shaping is known to leave optimal policies unchanged, and we measure the converse: a non-gradient perturbation of matched magnitude captures the greedy policy and drops optimality from 1.000 to 0.327. Belief revision presumes that its entrenchment order is well-founded, and curl-freeness decides that question instead of assuming it, telling a resolvable conflict from a money-pump at 1.00 where counting contradictions reaches 0.49. Whether a state representation is Markov enough for value becomes a quantity the agent reads off its own flow, and removing the curl it exhibits restores planning to oracle level. The geometry stays informative because nothing optimises against it, and training toward it proves null-to-harmful in all three forms we tried.

1 Introduction

A learning agent is conventionally described as an assembly of modules: perception, memory, a world model, a value function, a policy. That decomposition is *functional*. Each box is named by the job it is meant to perform, and the boundary between two boxes is a design choice made by the engineer before the agent runs.

The convention leaves a question it cannot itself answer. Once the agent is running, is there any measurement on its behaviour that says where one faculty ends and the next begins? Value, time, the sense of self and contradiction are studied in separate literatures, with separate formalisms and separate instruments. Nothing so far reads them off a single object, and nothing tests whether the separations survive a control designed to destroy them.

This paper defends and measures a *structural* alternative. We move the question from the architecture to the agent's empirical *flow* on the graph of its states: how it moves, what reward each move carries, which variable drives which, which argument outranks which. Any flow on a graph admits a canonical orthogonal decomposition into a small number of component types, and that decomposition is fixed by the graph rather than chosen by the modeller. The candidate object is the combinatorial Hodge decomposition (Jiang et al., 2011; Lim, 2020), which splits any edge flow orthogonally into a *gradient* part (the coboundary of a node potential: a consistent global ranking), a *curl* part (locally cyclic on filled triangles: a local inconsistency no ranking absorbs), and a *harmonic* part (globally cyclic holes), with an additional *sign* structure on cycles (signed-graph balance; Harary, 1953). It offers exactly the number of slots the question needs. Whether the faculties occupy them is an empirical matter, and that is what we measure.

Terminology. Four terms recur from the title onward. *Belnap's four-valued logic* (Belnap, 1977) is a logic of *told* evidence: a proposition is graded True (told true), False (told false), Both (told both: a contradiction that paraconsistent reasoning retains rather than averages away or explodes from), or None (told nothing). *AGM belief revision* (Alchourrón et al., 1985) is the standard axiomatic theory of how a rational agent retracts and revises beliefs; its contraction step removes the *least entrenched* beliefs first, where *entrench-*

ment (Gärdenfors, 1988) is the order of what one gives up last. The *efference copy* (von Holst & Mittelstaedt, 1950) is the agent's internal copy of its own motor commands, against which it can attribute observed change to itself. Finally $\mathrm{do}()$ denotes a Pearl-style *intervention*: acting on the world (or a copy of it) and reading the reproducible response, as opposed to passively observing. Our claim:

- The **self** is the *exact* part sourced by the agent's own efference.
- **Value** is the *gradient* (potential) part of the reward flow.
- **Time** is the *harmonic circulation*, that is, broken detailed balance.
- **Contradiction** is the *curl* and its sign holonomy.
- The agent's belief logic, **Belnap's four values**, is the topological regime of that same flow, read with no hand-tuned thresholds.

One operator, four faculties. Each item is falsified or confirmed on its own bench below.

Each pillar, taken alone, is a known object in its own field: combinatorial Hodge theory and HodgeRank; potential-based reward shaping and policy invariance (Ng et al., 1999); broken detailed balance and probability currents in statistical physics (Battle et al., 2016; Gnesotto et al., 2018); efference copy and reafference (von Holst & Mittelstaedt, 1950; Wolpert et al., 1995); signed-graph frustration (Harary, 1953); four-valued logic (Belnap, 1977) and belief revision (Alchourrón et al., 1985). The contribution is the map: four agent faculties as the four component types of one operator, and Belnap's values as its regimes, measured with pre-registered gates, seed-resampled confidence intervals, and adversarial controls (time-shuffle, false efference, node-count/random-lift, magnitude-matched fields, an exact reversible twin world). Concretely:

1. **Value is the gradient** (Section 4). The value-iteration potential is the Hodge potential of the reward flow (corr 1.000); a conservative shaping field is behaviourally invisible (optimality 1.000) while a magnitude-matched curl field is a money-pump (optimality 0.327, capture 0.991); projecting the curl *out* restores invariance (1.000) and the curl localises on the injected loop (AP 1.000). Ng–Harada–Russell policy invariance becomes a statement of *exactness*, a reading formalised independently by Jenner et al. (2022); our additions are the measured converse and the localiser.
2. **Time is the circulation** (Section 5). The antisymmetric part of the velocity operator recovers an autonomous oscillator at 0.86 while leaking 0.17 of a deliberately slow structured residue (slowness: 0.63/0.69); detection of irreversibility is a $4.8\times$ ratio against an exact reversible twin; and the *rank* of the current counts independent cycles ($2/1/0$ by spectral gap). A pre-registered null completes the picture: making time a *training objective* is null-to-harmful in all three forms tried; the geometry is a readout, never a loss.
3. **The self is the exact part; so is the other** (Section 6). The efference-attributable coboundary recovers the controllable factors (0.85) and is doubly dissociated from the autonomous current (false efference degrades only the self, $0.85 \to 0.71$; time-shuffle collapses only the current, $0.82 \to 0.02$). Hypothesising *another agent's* efference recovers what *they* cause (0.66 vs. passive 0.46), distinct from the autonomous harmonic, a geometric self/other/world axis, stated with its oracle bound.
4. **Contradiction is the curl** (Section 7). On the deconfounded current between variables, a consistent causal order is a gradient (0.883) whose potential recovers the causal rank ($\rho = 0.812$); an injected feedback 3-cycle is frustrated $10/10$ and localised (AP 1.000). The $\mathbb{Z}/2$ sign holonomy of signed belief graphs is exactly the $\{\pm 1\}$ shadow of this real-valued curl.
5. **Belnap's four values are the Hodge regimes of the flow** (Section 8). True/False = gradient support, Both = curl, None = harmonic hole: a dispatch with no hand-tuned threshold; entrenchment is the HodgeRank potential (Spearman $+1.00$ against the intended order); and minimal-entrenchment contraction (AGM-style revision) is well-founded *iff* the entrenchment flow is curl-free, an executable precondition that separates resolvable contestation from money-pumps at 1.00, where contradiction *counting* manages 0.49.
6. **The map applied to a real game** (Section 9). On a real 1977 text adventure, a non-Markov state representation *is* a value flow with curl; the state lift that makes value iteration succeed removes that curl ($0.032 \to 0.000$, numerical zero), a same-budget random refinement does not (0.073; structure, not node count), and this mechanistically explains a planning failure we measure in the same harness (flat value iteration 0.06 vs. lifted 1.00 on the deep goal). The curl further selects the repairing surgery (the single admitted coordinate captures the exercised non-Markovity entirely, $0.032 \to 0.000$), and the disjunction of passive curl and active $\mathrm{do}()$ probing rebuilds the full lift and reaches the deep goal at 0.88, equal to the oracle on every seed (both capped by exploration variance), where either signal alone reaches 0.02/0.31; at 16 seeds every capability margin is CI-separated (Section 9.1).

Scope claim. We claim no new learning mechanism and no state-of-the-art number. Every instrument is deterministic, seedless and linear; nothing in this paper is trained toward the geometry (a demarcation that is itself one of our measured results). The substrates are frozen synthetic families designed to be adversarial, plus one real game; the real-perception

port of the map is announced as open, not promised. What the paper establishes is that the same operator, applied as a pure readout to the agent's flows, yields the four faculties at their expected components, survives the controls designed to destroy each identification, and has two operational consequences: a mechanistic explanation of a real planning failure, and an executable well-foundedness test for belief revision.

2 Related work

Combinatorial Hodge theory and ranking. Our shared operator is the combinatorial Hodge decomposition popularised for statistical ranking by Jiang et al. (2011), with Lim (2020) as the reference treatment of Hodge Laplacians on graphs; simplicial signal processing (Barbarossa & Sardellitti, 2020; Schaub et al., 2020) is the ML environment of the same operator, which we use strictly as a readout, never as an architecture. The line remains active in statistics: Okahara et al. (2026) use the same decomposition to split paired comparisons into a covariate-explained transitive flow and an intransitive remainder. The closest prior assembly is Candogan et al. (2011), who decompose finite games into potential and harmonic components; our value leg (Section 4) is the single-agent cousin of their decomposition, and we cite it as the nearest anteriority for "value = potential part". What Candogan et al. (2011) does not contain is the multi-faculty composition (time, self, other, contradiction read from the *same* operator on the agent's own flows), the Belnap bridge, and the agentic operationalisation of Section 9 (state-lift as exactness repair; curl-selected surgery).

Reward shaping. Ng et al. (1999) prove that potential-based shaping leaves optimal policies unchanged; Wiewiora (2003) identifies it with Q-value initialisation; and Jenner et al. (2022) formalise potential shaping as a gradient in a discrete calculus on MDPs, the direct anteriority for the gradient reading of the theorem. Section 4 is the measured, instrumented cousin of that reading: we add the magnitude-matched *converse* (the non-exact part at matched magnitude is a behaviour-capturing money-pump that no value function can represent), the projection-out restoration, the localiser, and the placement of the value leg inside the multi-faculty map. The same non-representability has surfaced independently in preference learning, where Huang et al. (2026) separate a scalar transitive component from a cyclic one that no scalar reward can encode. That work reaches the conclusion from a game-theoretic decomposition of human preferences; we reach it from the Hodge decomposition of an MDP reward flow, with the converse and the localiser measured.

Time and irreversibility. Slow feature analysis (Wiskott & Sejnowski, 2002) motivates the scalar reading of time; the promotion to *geometry* imports the physics of broken detailed balance and probability currents (Battle et al., 2016; Gnesotto et al., 2018; Weiss, 2003), a literature that also treats the harder regime where no current is directly observable (Martínez et al., 2019). The statistical-physics anteriority of "time = circulation" is explicit there; our additions are the measured advantage over slowness against a deliberately slow structured residue, the harmonic-rank count of independent cycles, and the demarcation null (time as objective hurts).

Self and efference. The efference copy and reafference principle (von Holst & Mittelstaedt, 1950) and internal forward models (Wolpert et al., 1995) identify "what I cause" by comparing predicted and observed consequences of action. Our contribution is the geometric placement (the self as the *exact* (coboundary) component of the flow, doubly dissociated from the autonomous circulation) and its immediate corollary: another agent is exact-from-*another's* efference (Section 6.1).

Paraconsistency, belief revision, sheaves. Belnap's four-valued logic (Belnap, 1977) and the AGM theory of revision (Alchourrón et al., 1985; Gärdenfors, 1988; Hansson, 1999) provide the belief-layer vocabulary; sheaf-theoretic contextuality (Abramsky & Brandenburger, 2011), cellular sheaves (Curry, 2014) and sheaf neural networks (Bodnar et al., 2022; Barbero et al., 2022) geometrise locally-consistent/globally-inconsistent structure. Our Section 8 connects the two: the four values are the Hodge regimes of the argument flow, entrenchment is its HodgeRank potential, and the AGM-style contraction ordering is well-founded exactly when that flow is curl-free: an executable precondition the revision literature states as an assumption.

Non-Markov state and abstraction. Bisimulation-style state abstraction (Givan et al., 2003; Li et al., 2006) asks which state refinements preserve optimal behaviour; Allen et al. (2021) start from the same failure mode (rich observations need not preserve the Markov property) and *learn* abstractions that restore it. Our bridge result (Section 9) gives that question a geometric test: a representation is Markov-enough for value iff the value flow it induces is curl-free, and the needed refinement is localised by the curl itself, complementary to a learned-abstraction objective.

Nearest neighbour to the assembly. The Free-Energy-Principle literature uses the Helmholtz decomposition of flows into dissipative (gradient) and solenoidal (circulating) parts (Friston, 2010), which overlaps our *time* leg. It unifies by variational inference, whereas our map is a seedless geometric readout with per-faculty controls, and it carries neither the contradiction/sign-holonomy axis, nor the self/other dissociation, nor the Belnap regimes.

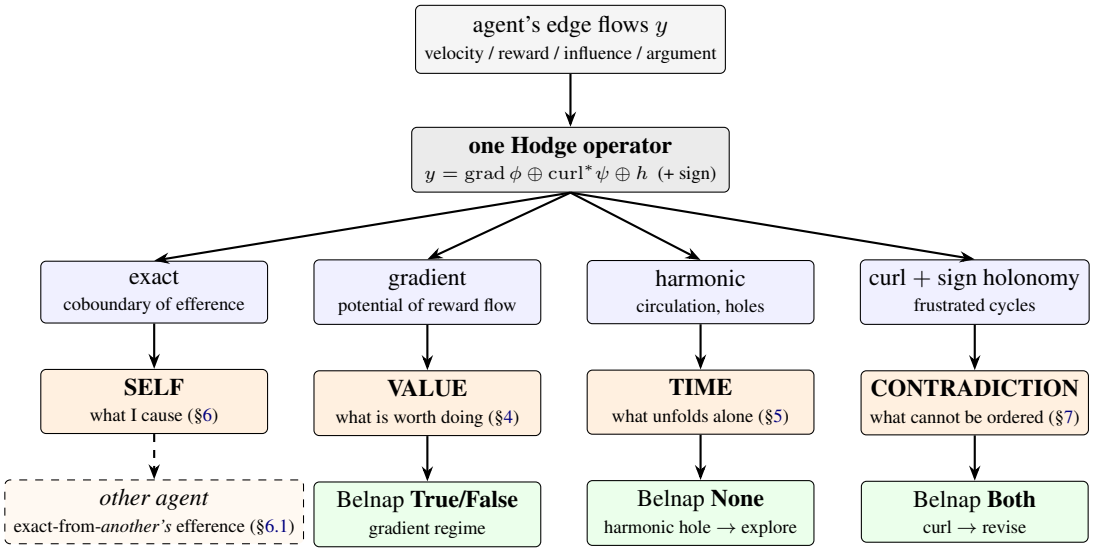

Figure 1: The map. One combinatorial Hodge operator, applied as a pure readout to the agent's empirical edge flows, yields the four faculties as its four component types (Sections 4–7), a fifth (another agent) as exact-from-another's-efference (Section 6.1), and Belnap's belief values as its topological regimes, with no hand-tuned thresholds (Section 8). The map is then put to work on a real game (Section 9).

3 The operator, the substrates, and the metrology

3.1 Principle

Let the agent's states (or variables, or propositions; the object varies by section, the operator does not) be the nodes V of a graph with oriented edges E and a set T of filled triangles. Write $C^0 = \mathbb{R}^V$, $C^1 = \{y : E \to \mathbb{R} \text{ antisymmetric}\}$, $C^2 = \{\text{triangle cochains}\}$, with coboundary maps

$$(\text{grad}\,\phi)(i \to j) = \phi_j - \phi_i,$$
$$(\text{curl}\,y)(i,j,k) = y_{ij} + y_{jk} + y_{ki}.$$

The combinatorial Hodge decomposition (Jiang et al., 2011; Lim, 2020) splits any edge flow orthogonally,

$$C^1 = \underbrace{\text{im}(\text{grad})}_{\text{gradient/exact}} \oplus \underbrace{\text{im}(\text{curl}^*)}_{\text{curl (local)}} \oplus \underbrace{\ker(\Delta_1)}_{\text{harmonic (holes)}}, \qquad (1)$$
$$\Delta_1 = \text{grad}\,\text{grad}^* + \text{curl}^*\,\text{curl},$$

and the least-squares node potential $\phi = \text{grad}^\dagger y$ is the *HodgeRank* potential of y. Three derived instruments, all deterministic, seedless, and never entering any training loss:

- **Inconsistency.** For a flow y, $\text{incons}(y) = \|y - \text{grad}\,\phi\|^2/\|y\|^2$ (curl + harmonic fraction): how far y is from being representable by *any* potential.
- **Current.** For a whitened trajectory Y_t, the velocity cross-moment $M = \langle Y_t(Y_{t+1} - Y_t)^\top\rangle$ and its antisymmetric part $A = \frac{1}{2}(M - M^\top)$: the first-order signature of a probability current, i.e. of *broken detailed balance* (each microscopic transition balanced by its reverse). The implication is one-sided in general: detailed balance implies $A = 0$ for any stationary process, so a non-zero A certifies broken balance, while $A = 0$ certifies balance only in the linear-Gaussian (whitened, first-order) setting this instrument operates in (Proposition 3, Appendix C; Weiss, 2003; Gnesotto et al., 2018; Martínez et al., 2019). $\|A\|_F$ is the circulation magnitude; the paired singular directions of A are the current-bearing planes; the number of pairs above a time-shuffle floor is the *harmonic rank* (count of independent cycles).
- **Sign holonomy.** On a signed flow, the product of edge signs around a cycle (Harary, 1953): -1 means a frustrated (unorderable) cycle, the $\mathbb{Z}/2$ shadow of a real-valued curl.

One clarification: the claim is "one operator", not "one flow". The same decomposition (1) is applied, as a pure readout, to *different empirical flows of the same agent*: the observed velocity flow (time, self), the reward flow on the state graph (value), the deconfounded influence current between variables (causality/contradiction), and the argument/entrenchment flow of a belief base (Belnap). The claim measured below is that each faculty lives at its predicted component of this shared machinery, and nowhere else.

3.2 Substrates

(i) For the time and self benches: a frozen family of synthetic reference worlds, introduced and audited in Fabre (2026) and regenerated bit-for-bit at runtime from (configuration, structural seed) by the code supplement: controllable + autonomous latent factors observed through frozen nonlinear

Table 1: The identifications, their benches, and their headline numbers (95% CIs over ≥10 seeds; game rows: 6 seeds, all usable, except the unified-surgery reach arms at 16; Section 3.3). Each row is measured in its own section with the control designed to destroy it. Bench names are the script names of the code supplement (Appendix A).

Hodge component	operator reading	faculty	bench (controls)	headline numbers
gradient (potential)	HodgeRank potential of reward flow	**value**	VALUE (magnitude-matched curl)	corr 1.000; opt. 1.000 vs curl 0.327; AP 1.000
harmonic (circulation)	antisym. velocity operator A	**time**	TIME-CURRENT/TIME-RANK (rev. twin, shuffle)	auto 0.86 / leak 0.17 (SFA 0.63/0.69); ratio $4.8\times$; rank $2/1/0$
exact (coboundary)	efference-attributable directions	**self**	SELF (false efference)	ctrl 0.85; false effer. $\to$ 0.71; shuffle-intact
curl + sign holonomy	frustrated cycles of influence	**contradiction**	FLOW-LOGIC (DAG floor, shuffle)	grad 0.883, $\rho = 0.812$; frustrated $10/10$; AP 1.000
exact, other source	another's efference (oracle)	*other agent*	OTHER (my-efference, passive)	other 0.66 vs passive 0.46
regimes of the flow	gradient / curl / harmonic	*Belnap values*	BELIEF-REGIMES (counting baseline)	Spearman +1.00; AGM precond. 1.00 vs 0.49
exactness repair	induced progress flow of a state key	*Markov-enough state*	STATE-LIFT/SURGERY-* (random lift)	$0.032 \to 0.000$ (num. zero); curl$\vee$do() 0.88 = oracle (seed-equal $16/16$)

"glass" maps (2-layer tanh networks) that add a *structured residue* $d = \tanh(W_d z)$, a deterministic function of the real factors, hence exactly as slow and as persistent as they are: the adversary every passive reading must survive. The current-bearing worlds replace the autonomous AR(1) dynamics with a rotation $z' = \rho R(\theta) z + \eta$; the *reversible twin* is the identical world with $\theta = 0$ (detailed balance holds), so that circulation is the only difference. **(ii)** For the value bench: runtime Markov decision processes (random connected graphs, N=12, degree ≈4, rich in triangles; one absorbing goal, step cost, discount $\gamma = 0.99$). **(iii)** For the causality/contradiction bench: runtime linear first-order vector-autoregressive (VAR(1)) systems (N=8), consistent (acyclic) vs. contradiction (same matrix + one injected balanced 3-cycle). **(iv)** For the theory-of-mind bench: a runtime three-source world (my actions, another agent's autocorrelated actions, an autonomous rotation) behind the same frozen glass. **(v)** For the bridge: the real Colossal Cave Adventure, in the `adventure` 1.7 Python port of the original 350-point Crowther and Woods FORTRAN game of 1977, driven by the same `advent.dat` data file, which yields 141 rooms. We use non-destructive do() by state copy and seed the game's own internal RNG (one frozen world instance; variance comes from the benches' explicit per-seed exploration RNGs). **(vi)** For the Belnap bench: small argument/entrenchment graphs seeded with consistent, contested-resolvable, and money-pump structures.

3.3 Metrology

Every bench follows the same discipline: fixed seeds with explicit random-number generators; ≥10 seeds with 95% confidence intervals reported as mean ± half-width, never point estimates. The game benches are the declared exception and run 6 seeds, because one game seed costs a forced execution of the deep chain plus its local probes. *Usable* is a pre-declared drop rule, not a post-hoc filter: a seed is discarded if the forced chain reaches no terminal, or if the induced graph is too small for a stable Hodge readout. In the runs reported here the rule never fired, so usable equals run: $6/6$ on STATE-LIFT, SURGERY-SELECT and SURGERY-NULL. The flat and lifted arms are in any case seed-invariant at ± 0, and the stochastic reach arms of SURGERY-UNIFIED use 16 seeds, a declared extension (Section 9.1). The rest of the discipline is shared: pre-registered hypotheses with thresholds fixed before the run; strong controls rather than strawmen: time-shuffle (destroys temporal order), false efference (mis-pairs actions and states), node-count/random-lift (same refinement budget, no structure), magnitude-matched fields (only conservativity differs), the exact reversible twin; and accepted nulls: a hypothesis that fails is reported as a result with its diagnosed mechanism (Sections 5.3 and 9.1 each contain one). Recovery of latent factors is scored by MCC, the mean correlation coefficient: the Hungarian-matched mean absolute Pearson correlation between recovered and true factors. Localisation is scored by average precision (AP) of the relevant ranking. Zero training terms: every instrument is a deterministic readout; the geometry never appears in an objective.

4 Value is the gradient

Identification 1 (value = potential part)**.** *Let r be the one-step reward flow on the agent's state graph. The value-iteration potential is the HodgeRank potential of r: a shaping field $F(s, s') = \gamma\Phi(s') - \Phi(s)$ is an exact 1-cochain (exactly at $\gamma = 1$; for $\gamma < 1$ its antisymmetric part is still* exactly *a gradient and the symmetric remainder is bounded by $(1-\gamma)\|\Phi\|_\infty$, Proposition 2), hence invisible to optimal behaviour and absorbed as a potential shift $V \mapsto V - \Phi$; conversely, the component of a reward perturbation that* changes *optimal behaviour is its non-gradient (curl+harmonic) part.*

This is Ng et al. (1999) re-read: their invariance theorem says exact fields are invisible, a gradient reading that Jenner et al. (2022) formalise as a discrete calculus on MDPs. What the Hodge frame adds, and what this section measures, is the converse instrument (what is *not* exact is a circulating reward, a *money-pump*: a loop the greedy policy gains by traversing forever, which no value function over states can represent) and a localiser (the curl concentrates on the faulty cells).

Hypothesis. (HV1) A pure-gradient shaping field of matched edge-magnitude leaves optimality intact and shifts the value by exactly the Hodge potential. (HV2) A magnitude-matched *curl* field changes behaviour; projecting the curl out by the decomposition restores invariance;

Table 2: Value is the gradient (bench VALUE; 10 graph seeds, 95% CIs; fields magnitude-matched in edge L_2 norm). The carrying metric is optimality preservation (the content of the invariance theorem); the raw argmax match under the gradient field is 0.900 ± 0.049 only because of ties between equally optimal shortest paths; both policies remain optimal (corr $= 1.000$, grad_frac $= 1.000$).

gradient (conservative) field	
grad_frac (Hodge)	1.000 ± 0.000
$\lvert\mathrm{corr}(V\text{-shift}, -\Phi)\rvert$	1.000 ± 0.000
optimality preserved	1.000 ± 0.000
curl (money-pump) field, same magnitude	
optimality preserved	0.327 ± 0.048
policy captured by the loop	0.991 ± 0.018
incons. vs. shuffle floor	1.000 vs. 0.718 ± 0.061
optimality, curl projected *out*	1.000 ± 0.000
curl localisation on loop (AP)	1.000 ± 0.000

behaviour is driven by the curl, not the magnitude. (HV3) The curl localises on the injected loop.

Experiment. Runtime MDPs as in Section 3.2; value iteration to V_{base}, π_{base}; add either (a) the coboundary field $F = \gamma\Phi(s') - \Phi(s)$ or (b) a circulating field of the *same edge L_2 norm* around one triangle; re-run value iteration. Decisive control: Hodge-project the curl out of (b) (keep gradient+harmonic) and re-run. 10 graph seeds; the instrument is seedless. (Bench VALUE.)

Result. Table 2. The gradient field is a pure, behaviourally invisible potential: optimality preserved 1.000 ± 0.000, grad_frac $= 1.000$, and the value shift *is* the Hodge potential ($|\mathrm{corr}| = 1.000$). The matched curl field captures the policy in the money-pump (optimality 0.327 ± 0.048, loop capture 0.991 ± 0.018; its Hodge inconsistency 1.000 sits CI-above the shuffle floor 0.718 ± 0.061); projecting the curl out restores optimality to 1.000 ± 0.000; and the per-edge curl energy ranks the injected loop first (localisation AP 1.000 ± 0.000).

Conclusion. Value joins the Hodge family as the gradient component of the reward flow; the curl is a value-space contradiction no value function represents, detected and localised by the same operator that reads time (Section 5) and logical contradiction (Section 7).

Declared bounds: the coboundary is discounted ($\gamma = 0.99$); Proposition 2 (Appendix C) makes the quasi-exactness exact: the antisymmetrised field the operator reads is a pure gradient at *every* γ (hence the measured grad_frac $= 1.000$, not ≈ 1) and the symmetric remainder is bounded by $(1-\gamma)\|\Phi\|_\infty$ (numerically verified, bench SHAPING-BOUND); one injected triangle loop; value iteration is clipped to bound the money-pump; a runtime toy MDP (size/discount robustness: Appendix D); the real-game version of this section is Section 9.

5 Time is the circulation, and must stay a readout

Identification 2 (time = harmonic current)**.** *A stationary observed process that satisfies detailed balance has vanishing antisymmetric velocity cross-moment A; within the linear-Gaussian reach of the instrument the converse also holds (Proposition 3), so a non-zero A certifies broken balance and detection is reported as a ratio against the reversible-twin and shuffle floors. The autonomous, self-unfolding part of the world is the current-bearing subspace of the flow (dominant paired singular directions of A), and the number of independent irreversible cycles is the number of singular-value pairs of A above the time-shuffle floor.*

5.1 From slowness to current: closing the slow-residue trap

The scalar reading of time is *slowness*: slow feature analysis (SFA; Wiskott & Sejnowski, 2002) extracts the directions that change most slowly, and on our worlds it does recover the slow autonomous factors (Table 3: 0.63–0.71, collapsing under time-shuffle). But slowness has a measured trap: the structured residue $d = \tanh(W_d z)$ is an *instantaneous* function of slow real factors, hence slow itself; slowness leaks it at 0.69–0.82, as badly as a passive baseline. The geometric prediction: an instantaneous image carries no current *of its own*, so a circulation readout should recover the autonomous factor while rejecting the slow residue.

Experiment. Whitened PCA-12 representation; $A = \frac{1}{2}(M - M^\top)$; dominant singular plane of A as the current subspace. Controls: the exact *reversible twin* (same frozen glass and residue, θ=0) and time-shuffle. 10 materialisation seeds; zero training. (Bench TIME-CURRENT.)

Result. Table 3. On the oscillator world the current recovers the autonomous factor at 0.86 while leaking only 0.17 of residue (slowness: 0.63 recovery, 0.69 leak); the separation margin (auto − residue) flips from -0.06 (slowness) to $+0.69$ (current). Circulation magnitude detects irreversibility as a *ratio*: $\|A\|_F = 0.50$ on the oscillator vs. 0.10 on the reversible twin ($4.8\times$; shuffle 0.11/0.09). A difference-in-differences decomposition: of the $+0.75$ margin advantage over slowness, $+0.51$ (68%) is true current (oscillator − reversible) and $+0.24$ (32%) is structural: an instantaneous residue couples *symmetrically*, so it is constitutively invisible to the antisymmetric part. Both mechanisms favour the geometric readout, and the second is the deeper reason the trap closes.

Table 3: Current closes the slow-residue trap (bench TIME-CURRENT; 10 seeds; zero training). The reversible twin shares glass and residue; only θ differs. "SFA" is the slowness readout; "passive" is a PCA baseline.

World	current auto	current resid	SFA auto	SFA resid	passive auto
oscillator	**0.86**	**0.17**	0.63	0.69	0.70
reversible twin	0.48	0.35	0.71	0.82	0.68

	circulation $\|A\|_F$	time-shuffle
oscillator	**0.50**	0.11
reversible twin	0.10	0.09

5.2 The harmonic rank counts cycles

The Hodge object is richer than detection: the harmonic part is a *subspace* whose rank is the number of independent irreversible cycles. Four worlds with ground-truth counts ($2/1/2$-coupled$/0$; K_{auto}=4, same frozen glass; bench TIME-RANK):

Result. The antisymmetric spectrum comes in pairs, one per rotational plane: `two_cycles` 0.44, 0.24, 0.07, 0.03; `one_cycle` 0.33, 0.05, ...; `reversible` 0.04, ..., against a shuffle floor of 0.05. The strict pre-registered rule (count pairs above the floor) is null on the two-cycle worlds: the residue, a nonlinear image of a rotating loop, traces a faint parasitic plane ($s_3 = 0.07$) just above the floor and the naive count overshoots by one. The refined verdict survives: the residue echo is rank-suppressed $3.5\times$–$6.6\times$ below the true cycles; a spectral-gap estimator recovers the true counts $2/1/0$ for *independent* cycles; and the trap closure of Section 5.1 survives in rank currency (current subspace recovers rotating factors at 0.64 vs. passive 0.37, rejecting residue 0.22 vs. slowness 0.67). Declared bound: *coupled* cycles genuinely blur the count (the coupling term is itself a real antisymmetric circulation; gap-rank returns 1 of 2); that is what coupling means, not an instrument failure.

5.3 The demarcation null: time is a readout, not an objective

Hypothesis. If temporal geometry is a good readout, adding a temporal term to the representation-learning objective should help.

Result. Null, in all three forms tried (10 seeds, InfoNCE contrastive learning, same world family; bench TIME-OBJECTIVE-NULL). Temporal contrastive positives *damage* the cross-view alignment (retrieval r@1 $0.64 \to 0.48$; the shuffle control collapses to 0.16, and its apparent autonomous gain is a residue artefact, leak $0.51 \to 0.73$); an SFA slowness penalty is redundant (auto-recovery 0.62 vs. 0.59, CIs overlap); and in the single-stream regime the *trained* time-contrastive encoder is worse than deterministic PCA (0.50 vs. 0.68) on the same world where the SFA *readout* beats PCA (0.71).

Conclusion. Time is a signal for *selection*, never a loss for *formation*. We apply this demarcation throughout: every instrument in this paper is a readout, and the demarcation is itself a measured result, not a stylistic preference. (Declared bound: in the single-stream world the real factors dominate variance by construction, PCA's favourable terrain; the cross-view harm result is the load-bearing one.)

6 The self is the exact part; so, structurally, is the other

Identification 3 (self = exact-from-efference)**.** *The self ("what I cause") is the coboundary component of the observed flow sourced by the agent's own efference copy: the subspace of observation directions written as* $W = \text{pinv}(A_{\text{actions}})\,\Delta\text{Obs}$ *(right singular directions), i.e. the image of the agent's* do(). *It must collapse under* false *efference (mis-paired actions) and be invariant to time-shuffle; the autonomous current (Identification 2) must show the mirror signature.*

Experiment. The oscillator family of Section 5.1 with an added controllable block driven by the agent's actions. Both instruments read the *same* flow. Double-dissociation controls: false efference (destroys the self's generator only) and time-shuffle (destroys the current's generator only). 10 seeds; linear, seedless instruments. (Bench SELF.)

Result. Table 4. The efference-exact component recovers the *controllable* factors at 0.85 ± 0.01, CI-above passive PCA (0.82 ± 0.00) and false efference (0.71 ± 0.07), and is orthogonal to the autonomous factors (0.04). The current recovers the *autonomous* factors at 0.82 ± 0.05 and ignores the controllable (0.08). The double dissociation is clean: false efference collapses the self ($0.85 \to 0.71$) but not the current; time-shuffle collapses the current ($0.82 \to 0.02$) but not the self. Each signal collapses only when its own generator is destroyed, which is the intended evidence that the two are distinct components of one flow. Declared margins: the self-vs-passive

Table 4: Self = exact, doubly dissociated from the autonomous current (bench SELF; 10 seeds; MCC over true factors).

	SELF (effer.)	CURRENT (harm.)
→ controllable	**0.85** ± 0.01	0.08 ± 0.02
→ autonomous	0.04 ± 0.01	**0.82** ± 0.05
passive baseline	0.82 ± 0.00	0.69 ± 0.01
under *false efference*	**0.71** ± 0.07	intact
under *time-shuffle*	intact	**0.02** ± 0.00

gap is small but CI-clean; the informative controls are the false-efference drop and the orthogonality.

6.1 Detecting another agent: exact-from-another's-efference

The self result implies a component the self/autonomous binary misses: a flow can have the *shape* of the controllable (low-dimensional, action-driven, intentional) without being caused by *my* efference. Structurally, that is the signature of *another agent*: exact-from-the-efference of someone, just not me. We call this the *oracle-bounded other-agency* decomposition: the other's efference is supplied to the readout (the bound is stated with the result below). Cognitively, it is a skeleton for the minimal theory-of-mind split self/other/world, not an unsupervised theory of mind.

Experiment. Three-source runtime world (Section 3.2): my controllable block (efference observed), another agent's block (driven by autocorrelated actions I do *not* observe), an autonomous rotation. The same instrument as Identification 3, applied to my efference (self), to the *hypothesised* other's efference (theory-of-mind oracle), and the current (world). 10 seeds. (Bench OTHER.)

Result. Three separations. **(HO1)** My instrument reads only what I cause: self→self $0.83 \pm 0.01 \gg$ self→other 0.15 ± 0.01, self→auto 0.03. **(HO2)** Hypothesising the other's efference recovers what *they* cause: 0.66 ± 0.02, CI-above passive (0.46 ± 0.04) and my efference (0.15). **(HO3)** The other is distinct from the autonomous world: the current prefers the autonomous factor (0.65 ± 0.04) over the other's (0.36 ± 0.06): an agent (exact-from-an-invisible-action) is geometrically distinct from an autonomous circulation.

Conclusion. The flow separates into self (exact from *my* efference), other (exact from *another's* efference), and world (harmonic). The load-bearing bound: HO2 is an oracle readout, the other's efference is given to it. This establishes that the component exists and is well-defined; the unsupervised discovery of another agent's policy from the residual flow is the named open problem, and the margins are modest (passive already captures 0.46 of the other's factor; the other's autocorrelation leaks 0.36 into the current).

7 Contradiction is the curl

Identification 4 (causal order = gradient; contradiction = curl)**.** *Let $J = \mathrm{antisym}(B)$, $B = C_{10}\Sigma_0^{-1}$, be the deconfounded direct-influence current of a stationary VAR(1) process (the antisymmetric part of the Granger/direct-effect matrix, not the raw cross-covariance, which conflates direct, indirect and common-cause paths and fabricates parasitic frustration even on a directed acyclic graph, DAG). If the generating structure is an acyclic causal order, J is (dominantly) a gradient whose HodgeRank potential recovers the topological rank; an injected feedback 3-cycle is a curl no potential absorbs, carrying sign holonomy -1, localised on its triangle.*

Experiment. Two runtime VAR(1) worlds (N=8) sharing all frozen structure except the injected cycle: *consistent* (forward-only edges in a random topological order, a true DAG) and *contradiction* (the same matrix + one balanced 3-cycle, the mutual DAG edges of those three nodes overwritten). Read J, threshold at the shuffle-derived support floor (edges below it carry no directed relation, anticipating the None regime of Section 8), decompose. 10 seeds. (Bench FLOW-LOGIC.)

Result. Table 5. **(HF1)** Causality is the gradient: $\mathrm{grad\_frac} = 0.883 \pm 0.068$ vs. $\mathrm{curl\_frac} = 0.060 \pm 0.038$; the potential recovers the causal order ($|\mathrm{Spearman}| = 0.812 \pm 0.077$); time-shuffle collapses everything to 0.000. **(HF2)** The injected feedback cycle is a curl: its triangle is frustrated $10/10$ (rate 1.000, CI-above both the other-triangles rate 0.300 and the finite-sample DAG floor 0.326 ± 0.244); its graded curl is 0.522 ± 0.002 vs. 0.000 under shuffle. **(HF3)** It localises: AP 1.000 ± 0.000, the triangle recovered 100% of seeds.

Conclusion. Causality and contradiction are not two mechanisms but two Hodge components of one inter-variable current: the gradient is the arrow, the curl is the frustration, and logical coherence of a causal order $\iff$ curl-freeness of its flow. The $\mathbb{Z}/2$ *sign* holonomy familiar from signed-graph frustration and paraconsistent belief graphs is exactly the $\{\pm 1\}$ shadow of this real-valued curl; the graded curl adds the *magnitude* of the frustration to its sign. Declared bounds: a linear instrument on a linear structural equation model (faithful here by construction; the nonlinear port is open); one injected cycle; the complete-graph complex suppresses the harmonic part by construction, so contradiction reads as curl (consistent with its localisation on a triangle); and the

Table 5: Flow logic (bench FLOW-LOGIC; 10 seeds; deconfounded current, support-thresholded).

consistent world (acyclic)	
grad_frac / curl_frac	0.883 ± 0.068 / 0.060 ± 0.038
$\lvert\text{Spearman}(\phi,\ \text{causal rank})\rvert$	0.812 ± 0.077
grad_frac under time-shuffle	0.000 ± 0.000
per-triangle frustration floor (DAG)	0.326 ± 0.244
contradiction world (+ 3-cycle)	
inj. triangle recovered / frustrated	1.000 / 1.000
frustration: inj. / others / DAG floor	**1.000** / 0.300 / 0.326
graded curl on inj. tri. (vs. shuffle)	0.522 ± 0.002 vs. 0.000
curl localisation (AP)	1.000 ± 0.000

finite-sample current keeps a frustration *floor* (≈ 0.33 on a clean DAG); the verdict rests on the injected cycle being a decisively separated anomaly, not on an ideally clean DAG.

8 Belnap's four values are the Hodge regimes of the flow

Recall (Section 1) that Belnap's four-valued logic (Belnap, 1977) assigns a proposition one of {True, False, Both, None}: told-true, told-false, told-both (contradiction, to be *retained*, not averaged away), and told-nothing. Paraconsistent belief layers built on it need two further objects: an *entrenchment* order ρ deciding what to give up first (Gärdenfors, 1988), and a revision operator in the AGM family (Alchourrón et al., 1985; Hansson, 1999) whose contraction-by-least-entrenched step silently *assumes* the order is well-founded. Both are usually specified axiomatically, with thresholds and tie-breaking supplied by hand. The map of this paper supplies them geometrically, with no hand-tuned threshold (the only floors are inherited, shuffle-calibrated ones, Section 7):

Identification 5 (Belnap values = Hodge regimes)**.** *On the argument/entrenchment flow of a belief base (edges: pairwise support/attack or entrenchment comparisons), True/False is the* gradient regime *(the flow is consistently ordered; the sign of the potential difference gives the polarity), Both is the* curl regime *(a frustrated cycle: each local comparison holds, no global order ranks them; energy that orthogonality conserves and localises, never erases), and None is the* harmonic-hole regime *(a circulation no filled structure explains, or an isolated node: the signature of missing evidence, answered by exploration, not by revision). The entrenchment ρ is the HodgeRank potential of the same flow.*

Experiment. Small seeded belief bases (Section 3.2(vi)): consistent bases; contested bases whose conflict is resolvable by removing least-entrenched elements; money-pump bases ($A \succ B \succ C \succ A$ at matched local strengths); chordless 4-cycles (holes). Dispatch read from the decomposition of the flow, with no hand-tuned threshold anywhere; entrenchment read as the HodgeRank potential; the AGM-style precondition tested as: "remove the least entrenched" is well-founded iff the entrenchment flow is curl-free. Baseline for the separation task: contradiction *counting* (is the told-both set non-empty?), the naive statistic a symbolic layer would use. 10 seeds. (Bench BELIEF-REGIMES.)

Result. Three results. **(B1) The dispatch is topological, with no hand-tuned threshold**: consistent bases land in the gradient regime (True/False by potential sign); the money-pump lands in the curl regime, localised on its triangle; the chordless 4-cycle lands in the harmonic regime (a hole, the explore-signal, distinct from the curl's revise-signal). **(B2) Entrenchment is the HodgeRank potential**: the recovered ρ reproduces the intended AGM entrenchment order at Spearman $+1.00$. **(B3) The executable AGM precondition**: classifying bases as well-founded vs. ill-founded for minimal-entrenchment contraction, the curl criterion scores 1.00 where contradiction counting scores 0.49 (10 seeds); counting cannot distinguish a contested-but-resolvable base from a money-pump (both have non-empty conflict sets); the curl can, because resolvability is exactly representability by a potential after removal, i.e. curl-freeness of what remains.

Conclusion. The agent's belief logic is not a separate symbolic layer: the four values are the regimes of the same decomposition that reads value, time, self and contradiction (True/False = gradient, Both = curl (conserved and localised, honouring the paraconsistent non-suppression requirement), None = harmonic hole), and the revision machinery inherits an executable well-foundedness test the axiomatic literature leaves as an assumption. Note the consistency with Section 4: a money-pump in preference space and a money-pump in reward space are the same object at the same component. Declared bounds: toy substrates at the mechanism grain (the belief bases are seeded, not harvested from a live corpus); the dispatch inherits the finite-sample floors of Section 7. Whether the *independence of the sources* feeding a belief (the security lock any deployed belief layer ultimately rests on) can itself be measured is deliberately left open here.

9 The map applied: a non-Markov state is a curled value flow

Everything so far reads faculties off flows. This section shows that the map explains and then repairs a real planning failure. The substrate is the real Colossal Cave Adventure (Section 3.2(v); 141 rooms; non-destructive do() by state copy), whose deep treasure chain is gated: the bird can only be caught while not holding the rod, and the passage past the snake opens only after the freed bird drives it off. Measured in our own harness (Section 9.1, Table row "flat"/"oracle"):

value iteration on the flat room-id state scores 0.06 on the deep, gated treasure chain, while the same value iteration on the state lifted with four flags reaches it at 1.00. This section gives the mechanism.

Proposition 1 (state lift as exactness repair)**.** *Fix a representation-independent progress potential d^* (here: oracle breadth-first-search distance-to-goal on the full Markov state, used for evaluation only). A state representation σ is Markov-enough for value iff the edge flow induced by projecting observed transitions onto σ's state keys carries d^* as a gradient; non-Markovity of σ manifests as non-zero curl+harmonic energy (inconsistency), and a correct lift is precisely the refinement that removes it.*

Experiment. Score three representations on the inconsistency of their induced progress flow: **flat** (room id); **lifted** (room, oracle flags: the full Markov state; the autonomous *discovery* of these flags is Section 9.1's subject and is out of scope here); **random-lift** (room, k random bits: same refinement budget, no phase alignment; the node-count control). Transitions sampled by executing the real chain with local probes; 6 seeds, all usable (flat/lifted are deterministic at ± 0). (Bench STATE-LIFT.)

Result. Figure 2. Flat: 0.032 ± 0.000. Lifted: $\mathbf{0.000} \pm \mathbf{0.000}$: the lift makes the flow exactly a gradient (the inconsistency drops to numerical zero). Random-lift: 0.073 ± 0.037: a same-budget random refinement does *not* restore conservativity; it adds curl (fewer transitions per edge). The curl localises on the prerequisite-chain rooms (residual 0.045 vs. 0.017 elsewhere, $\approx 2.6\times$), and the non-Markov subregion alone carries 0.092 ± 0.000. The magnitude is load-bearing: the global curl is small; Adventure is mostly Markov at room level, and the non-Markovity is sparse, concentrated at the gates. That is exactly the profile of a game where flat value iteration mostly works but fails *specifically* on the deep gated chain: the mechanistic reading of the score-0 result. This explains; it adds no capability by itself.

9.1 From explaining to repairing: the curl selects the surgery

If non-Markovity is curl, the curl should not only diagnose but point to the repair. Three benches, all on the real game, zero training:

(i) The curl selects (bench SURGERY-SELECT, GO). Score every candidate 1-bit state split by its curl drop $\Delta(b) = \text{incons}(\text{flat}) - \text{incons}(\text{room}{+}b)$. The deep navigation gate `prop:snake` scores $+0.032 \pm 0.000$ (CI > 0) and alone captures the exercised non-Markovity entirely ($0.032 \to 0.000$); oracle-true but movement-inert flags score 0.000 (`hold:cage`, `hold:rod`: the instrument rewards real potential variance, not labels), while the affordance flag

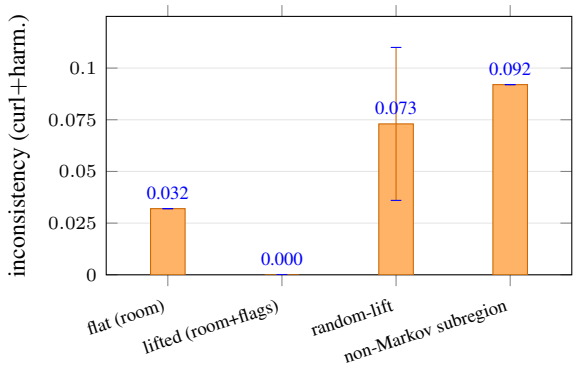

Figure 2: A non-Markov state is a curled value flow, and the lift removes the curl (bench STATE-LIFT; real Colossal Cave Adventure; 6 seeds). The oracle-flag lift makes the progress flow exactly a gradient ($0.032 \to 0.000$, numerical zero); a same-budget *random* lift does not (it adds curl): the lever is the relevant structure, not node count. The curl concentrates on the gated prerequisite chain (subregion 0.092; localisation $\approx 2.6\times$).

`hold:bird` scores *negative* (-0.055), anticipating (ii): a gate with no room-graph signature only dilutes the flow when split on; and all three *random* bits score *negative* (-0.141 to -0.176): adding degrees of freedom without structure dilutes the flow; the node-count objection is answered in the data. The curl is a localiser of the currently exercised contradiction: it points to the one coordinate that resolves the frustration the flow sees now, the cheapest surgery first.

(ii) One-pass curl does not close the loop (bench SURGERY-NULL, measured null). Using curl-admission alone to build the lift and then planning fails: goal reach 0.00 ± 0.00, against flat 0.06 ± 0.11 and oracle 1.00 ± 0.00. The admission itself is sound. It is parsimonious (2.0 ± 0.0 flags out of 24 candidates) and always contains the navigation gate.

Three measurements diagnose the failure. First, the payoff of the true flags is *conjunctive*: the navigation flag the curl admits reaches 0.02 on its own, the affordance flags alone reach only stochastically, and only their union closes, as (iii) measures. Second, the one spurious correlate the curl admits (`prop:silver`) is harmless, since it is carried into the unified lift of (iii), which still matches its oracle. Third, and this is the structural point, curl computed on the *room-movement* graph is constitutively blind to gates that spoil *same-room affordances*. Catching or freeing the bird changes no room, so it creates no room-graph holonomy, and the corresponding curl drop is ≤ 0, as measured in (i). The reach of a geometric readout is the complex it is built on.

(iii) The disjunction closes it (bench SURGERY-UNIFIED, GO). Add the active signal the passive one lacks: a do() toggle test that asks whether dropping X flips the *success* of a chain affordance. The toggle owns exactly the affordance gates (`hold:rod/cage/bird`) the curl cannot see, and the curl owns the navigation gate (`prop:snake`) the toggle cannot see.

Goal reach at 16 seeds, a declared extension of the shared 6-seed default made to adjudicate the one margin the 6-seed run left CI-unseparated (the pre-registered GO rule is unchanged): flat 0.02 ± 0.04; curl-only 0.02 ± 0.04; toggle-only 0.31 ± 0.18; unified curl $\cup$ do() 0.88 ± 0.14, equal to the oracle 0.88 ± 0.14 on *every* seed ($16/16$, both capped by exploration variance). Neither signal alone rebuilds the lift, because the admitted gate sets are complementary, and we measure that they are. Each single-signal lift misses its complementary gate and therefore reaches only stochastically, at 0.00 to 1.00 per seed, which is exploration variance through an unmodelled gate. Every capability margin is CI-separated, so the pre-registered rule passes in full. The disjunction reaches the oracle ceiling: passive geometry and active intervention act as complementary admission signals, here performing state surgery.

Conclusion. The map is operative, with its division of labour measured: the curl *explains* the failure (Proposition 1), *selects* the cheapest repair, and needs the do() as its complement exactly where the flow's complex has no edges to be frustrated on. Declared bounds: the lift's flags are oracle-given in (i) and Section 9's diagnosis (their autonomous discovery is exactly what (ii)/(iii) measure); the two exploitation harnesses cap their own oracles at 1.00 ((ii)) and 0.88 ((iii)) by exploration variance; the forced-chain collection exercises only the snake gate's curl in (i), which is precisely what motivates the iterative re-collect design that (iii) implements.

10 Discussion

What kind of claim is the map? Each pillar is a known object in its own field; we do not improve any of them. The finding is that the four faculties land at the four component types of one shared operator *when measured on the same agent's flows with the controls that would destroy a spurious identification*: false efference does not touch the current, time-shuffle does not touch the self, magnitude-matching does not save the curl field, random lifts do not fake the exactness repair, a reversible twin bounds the instrument floor. The identifications are therefore not metaphors: each one survived its designed refutation.

One consistent object across currencies. The same non-representability shows up as a money-pump in reward space (Section 4), a feedback cycle in causal space (Section 7), an unorderable preference cycle in belief space (Section 8), and a non-Markov gate in state space (Section 9); in every currency it is the curl, it is localised by the same operator, and "repair" means the same thing: restore exactness (project out, revise, or lift). Conversely the harmonic part is everywhere the *hole*: the circulation no local structure explains, the autonomous world in the velocity flow, the told-nothing regime in the belief flow. This suggests a self-diagnosis discipline for agents: curl on a filled cell $\Rightarrow$ revise (a local contradiction with a cheapest discriminating repair, which the curl localises); harmonic $\Rightarrow$ explore (a hole in the complex; no revision will absorb it, only new structure). We state this as a discipline, not a result: the head-to-head bench (curl-routed revision and harmonic-routed exploration vs. undirected baselines at matched budget) is future work.

Readout, never objective. The demarcation null (Section 5.3) is not a side note; it fixes how the geometry is to be used. Every payoff in this paper comes from reading the geometry at selection/diagnosis time; the one attempt to *train toward* it (temporal terms in the objective, three forms) is null-to-harmful. We suspect this generalises: the components are informative precisely because nothing optimises against them.

Relation to the Helmholtz/FEP split. The nearest neighbour to the assembly decomposes flow into dissipative and solenoidal parts under a variational principle (Friston, 2010). The overlap is the time leg. The differences are load-bearing: our decomposition is combinatorial (it has a curl *between* the gradient and the harmonic, hence a localiser and a sign holonomy), it is read with zero fitted parameters, and it carries the axes FEP does not: self vs. other by efference source, contradiction as a first-class component, and the Belnap regimes with an executable revision precondition.

11 Limitations

- **Substrates.** The four core identifications are measured on synthetic families (frozen, audited, adversarial by design, but synthetic), plus one real game for the bridge and seeded belief bases for the Belnap section. No claim is made about real perception: the port of the map to live perceptual streams (state graphs over an agent's lifted concepts) is announced as the next step, not promised.
- **Instruments are linear on nonlinear glass.** This leaves declared nonlinearity floors: irreversibility detection is a *ratio* (reversible-twin and shuffle controls are mandatory, not optional); the deconfounded current keeps a finite-sample frustration floor (≈ 0.33 on a clean DAG); the naive harmonic-rank count overshoots by one on a residue echo and must use the spectral gap; coupled cycles genuinely blur the count. The detailed-balance equivalence for A is itself linear-Gaussian (Proposition 3): in general a non-zero A certifies broken balance but $A = 0$ certifies nothing, which is exactly why detection is reported as a ratio against the twin and the shuffle floor.
- **Depth of confirmation is uneven.** One identification is confirmed to the point of exact coincidence (value = potential, corr 1.000, with its converse and localiser);

the others are confirmed as CI-separated recoveries with double dissociations, on single injected structures (one rotation, one cycle, one loop). The theory-of-mind component is an *oracle* readout: the other's efference is given; unsupervised discovery is open.

- **The bridge uses oracle flags for the diagnosis.** The exactness-repair result (Proposition 1) takes the lift's flags as given (their autonomous discovery is Section 9.1's subject); the room graph is sparse, so curl and harmonic are reported jointly as inconsistency; the global curl magnitude is small (declared: it is *why* the failure was localised); the surgery arc's exploitation harnesses cap their own oracles at $1.00/0.88$; and the unified bench's reach arms run at 16 seeds, a declared extension of the shared 6-seed default made to adjudicate the one margin the 6-seed run left CI-unseparated (Section 9.1(iii); the pre-registered rule, unchanged, passes in full at 16).
- **The Belnap bridge is at mechanism grain.** Seeded bases, not a harvested corpus; the claim is that the regimes and the precondition are well-defined and separate the seeded cases (1.00 vs. 0.49), not that a deployed belief layer is delivered. Source-independence certification is left open (one sentence in Section 8; nothing here depends on it).
- **What the map does not claim.** No new capability is added by the diagnosis itself (Section 9 explains a failure measured in the same harness); the capability in Section 9.1 comes from the curl *plus* the $\mathrm{do}()$, and the null (bench SURGERY-NULL) showing one-pass passive curl is insufficient is part of the record. The curl⇒revise / harmonic⇒explore routing is a stated discipline whose head-to-head bench is future work.

12 Conclusion

On our benches, the four faculties are not modules to be designed but four orthogonal readings of one operator on the agent's flows: value is the gradient (the value function is the Hodge potential of the reward flow, and what no value can represent is exactly the curl); time is the circulation (broken detailed balance, with a rank that counts its independent cycles); the self is the exact part sourced by the agent's own efference (and another agent is the exact part sourced by someone else's); contradiction is the curl and its sign holonomy: in reward space a money-pump, in causal space a feedback loop, in belief space an unorderable cycle, in state space a non-Markov gate, each localised by the same decomposition. The belief logic itself follows: Belnap's four values are the topological regimes of the flow, entrenchment is its HodgeRank potential, and AGM-style revision gains an executable well-foundedness test (1.00 vs. 0.49 for contradiction counting). The map has two operational consequences: it mechanistically explains a real planning failure (a non-Markov state is a curled value flow; the lift is the exactness repair, to numerical zero), and, disjoined with the $\mathrm{do}()$ it cannot replace, it performs the repairing surgery to oracle level (0.88, equal to the oracle on every seed; every single-signal margin CI-separated).

The demarcation is strict: the geometry is a readout and a gate, never a training objective; the one attempt to optimise toward it was null-to-harmful in all three forms tried. The boundary of the evidence is stated with the same precision: synthetic substrates plus one real game; linear instruments with declared floors; one identification confirmed to exact coincidence and the others as controlled, CI-separated recoveries; an oracle bound on the other-agency component. Three questions remain open: the real-perception port of the map, the unsupervised discovery of another agent's efference, and the head-to-head test of curl-routed revision against harmonic-routed exploration. Within its measured perimeter, the evidence supports a structural answer to the architecture question: the agent's faculties are the Hodge components of its flow.

Reproducibility statement

A self-contained code supplement (the `code/` directory accompanying this paper, also released at https://github.com/pierrethibaultfabre/Research-AI-Memory-Hodge-Belnap-Operator) reproduces every number: it has no dependency beyond `numpy`/`torch` (plus the public-domain `adventure` package for the real game), regenerates every synthetic world bit-for-bit from (configuration, structural seed), and snapshots the worlds it used. All benches are pre-registered (hypotheses and thresholds fixed before runs, collected in `PREREGISTRATION.md`), run at $\geq$10 seeds with 95% confidence intervals (the game benches run 6 seeds, all usable under the drop rule declared in Section 3.3, with the deterministic arms at ± 0; the stochastic reach arms of the unified-surgery bench use 16, a declared extension), and are deterministic given (config, seed); every instrument is seedless and training-free. Each bench script prints its pre-registered verdict and its controls; one command (`python run_all.py`) re-runs every bench and writes a machine-readable manifest of the verdicts (`results/manifest.json`), stamped with the release commit hash. Appendix A maps every bench name used in the text to its script.

References

Abramsky, S. and Brandenburger, A. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13:113036, 2011.

Alchourrón, C. E., Gärdenfors, P., and Makinson, D. On

the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

Allen, C., Parikh, N., Gottesman, O., and Konidaris, G. Learning Markov state abstractions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8229–8241, 2021.

Barbarossa, S. and Sardellitti, S. Topological signal processing over simplicial complexes. *IEEE Transactions on Signal Processing*, 68:2992–3007, 2020.

Barbero, F., Bodnar, C., Sáez de Ocáriz Borde, H., Bronstein, M., Veličković, P., and Liò, P. Sheaf neural networks with connection laplacians. In *Proceedings of Topological, Algebraic and Geometric Learning Workshops (ICML)*, volume 196 of *Proceedings of Machine Learning Research*, pp. 28–36, 2022.

Battle, C., Broedersz, C. P., Fakhri, N., Geyer, V. F., Howard, J., Schmidt, C. F., and MacKintosh, F. C. Broken detailed balance at mesoscopic scales in active biological systems. *Science*, 352(6285):604–607, 2016.

Belnap, N. D. A useful four-valued logic. In Dunn, J. M. and Epstein, G. (eds.), *Modern Uses of Multiple-Valued Logic*, pp. 5–37. D. Reidel, Dordrecht, 1977.

Bodnar, C., Di Giovanni, F., Chamberlain, B. P., Liò, P., and Bronstein, M. M. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Candogan, O., Menache, I., Ozdaglar, A., and Parrilo, P. A. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3): 474–503, 2011.

Curry, J. M. *Sheaves, Cosheaves and Applications*. PhD thesis, University of Pennsylvania, 2014.

Fabre, P.-T. What makes a latent real? A unified emergence gate: Controllability, cross-view invariance and temporal geometry — and why it only pays under intervention. Preprint, 2026.

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

Gärdenfors, P. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA, 1988.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Gnesotto, F. S., Mura, F., Gladrow, J., and Broedersz, C. P. Broken detailed balance and non-equilibrium dynamics in living systems: A review. *Reports on Progress in Physics*, 81(6):066601, 2018.

Hansson, S. O. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, Dordrecht, 1999.

Harary, F. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143–146, 1953.

Huang, Y., Li, X., Zhao, K., and Li, J. Transitivity meets cyclicity: Explicit preference decomposition for dynamic large language model alignment. *arXiv preprint arXiv:2605.17342*, 2026.

Jenner, E., van Hoof, H., and Gleave, A. Calculus on MDPs: Potential shaping as a gradient. *arXiv preprint arXiv:2208.09570*, 2022.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Li, L., Walsh, T. J., and Littman, M. L. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, 2006.

Lim, L.-H. Hodge Laplacians on graphs. *SIAM Review*, 62 (3):685–715, 2020.

Martínez, I. A., Bisker, G., Horowitz, J. M., and Parrondo, J. M. R. Inferring broken detailed balance in the absence of observable currents. *Nature Communications*, 10:3542, 2019.

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pp. 278–287, 1999.

Okahara, H., Nakagawa, T., and Sugasawa, S. The covariate-assisted Bayesian intransitive Bradley–Terry model via combinatorial Hodge theory. *arXiv preprint arXiv:2601.07158*, 2026.

Schaub, M. T., Benson, A. R., Horn, P., Lippner, G., and Jadbabaie, A. Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Review*, 62(2): 353–391, 2020.

von Holst, E. and Mittelstaedt, H. Das Reafferenzprinzip: Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften*, 37:464–476, 1950.

Weiss, J. B. Coordinate invariance in stochastic dynamical systems. *Tellus A: Dynamic Meteorology and Oceanography*, 55(3):208–218, 2003.

Wiewiora, E. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

Wiskott, L. and Sejnowski, T. J. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Wolpert, D. M., Ghahramani, Z., and Jordan, M. I. An internal model for sensorimotor integration. *Science*, 269 (5232):1880–1882, 1995.

A Bench index

Table 6: Bench names used in the text, their role, their substrate, and the reproducing script in the code supplement.

Bench	Role in this paper	Substrate	Script
VALUE	value = gradient; money-pump; projection-out; localisation	runtime MDPs (N=12)	`bench_value.py`
TIME-CURRENT	time as current; closes the slow-residue trap	frozen family + reversible twin	`bench_time_current.py`
TIME-RANK	harmonic rank counts independent cycles	frozen family, ≤ 2 planes	`bench_time_rank.py`
TIME-OBJECTIVE-NULL	demarcation null: time as objective hurts	frozen family; InfoNCE	`bench_time_objective_null.py`
SELF	self = exact-from-efference; double dissociation	oscillator family + actions	`bench_self.py`
OTHER	other agent = exact-from-another's-efference (oracle)	three-source runtime world	`bench_other.py`
FLOW-LOGIC	causality = gradient; contradiction = curl; localisation	runtime VAR(1) (N=8)	`bench_flow_logic.py`
BELIEF-REGIMES	Belnap regimes; ρ = HodgeRank; AGM precondition	seeded belief bases	`bench_belief_regimes.py`
STATE-LIFT	non-Markov = curl; lift removes it; random-lift control	real Colossal Cave Adventure	`bench_state_lift.py`
SURGERY-SELECT	curl selects the surgery (one-bit split menu)	real Colossal Cave Adventure	`bench_surgery_select.py`
SURGERY-NULL	one-pass curl null + diagnosed mechanism	real Colossal Cave Adventure	`bench_surgery_null.py`
SURGERY-UNIFIED	curl$\vee$do() rebuilds the lift, reaches the goal	real Colossal Cave Adventure	`bench_surgery_unified.py`
SHAPING-BOUND	Proposition 2 verified numerically	runtime MDPs, γ grid	`bench_shaping_bound.py`
ROBUSTNESS-SWEEPS	size / noise / sample-length sweeps (Appendix D)	runtime MDPs / VAR(1) / frozen family	`bench_robustness_sweeps.py`

B Shared operator

All decompositions in this paper are computed by one module (`hodge.py` in the code supplement): build the complex (nodes, oriented edges, filled triangles), assemble grad and curl, project the empirical edge flow onto $\mathrm{im}(\mathrm{grad})$, $\mathrm{im}(\mathrm{curl}^*)$ and $\ker(\Delta_1)$ by least squares, and report the component fractions, the HodgeRank potential, the per-cell curl energies, and (for signed flows) the cycle sign holonomies. The velocity-current instrument ($A = \frac{1}{2}(M - M^\top)$ on the whitened stream) and the efference-coboundary instrument ($\mathrm{pinv}(A_{\mathrm{actions}})\,\Delta\mathrm{Obs}$) are the only two additions, both linear and deterministic. No component of the machinery has tunable parameters.

Their relation to the decomposition (1) is exact rather than analogical, which is what justifies the phrase "one operator". Both are estimators of components of the same split, adapted to flows sampled along trajectories rather than given edge-wise. The efference-coboundary instrument estimates the part of the observed flow lying in $\mathrm{im}(\mathrm{grad})$ whose generator is the agent's own action stream: it solves the same least-squares normal equations restricted to the directions the efference can write (which is why *false* efference, and only false efference, destroys it, Table 4). The velocity instrument A is the first-order trajectory statistic of the circulating (non-exact) side: on an explicit complex the circulation lives in $\mathrm{im}(\mathrm{curl}^*) \oplus \ker(\Delta_1)$; on a sampled stationary stream the same circulation appears as the antisymmetric part of the position–velocity cross-moment, and Proposition 3 states the setting in which this reading is exact (which is why the time-shuffle, and only the time-shuffle, destroys it).

C Formal statements

Proposition 2 (discounted shaping: quasi-exactness made exact)**.** *Let $\Phi : V \to \mathbb{R}$ and let $F(s, s') = \gamma\Phi(s') - \Phi(s)$, $\gamma \in (0, 1]$, be the discounted shaping field on directed edges. Then F splits uniquely into an antisymmetric and a symmetric part,*

$$F(s,s') \;=\; \underbrace{\tfrac{1+\gamma}{2}\big(\Phi(s') - \Phi(s)\big)}_{=\,\mathrm{grad}\,\bar{\Phi},\;\; \bar{\Phi}=\frac{1+\gamma}{2}\Phi} \;+\; \underbrace{\Big(-\tfrac{1-\gamma}{2}\big(\Phi(s) + \Phi(s')\big)\Big)}_{=\,\sigma(s,s') \,=\, \sigma(s',s)},$$

so the antisymmetric edge flow read by the operator of Section 3.1 is exactly *a gradient for every γ (its Hodge inconsistency is 0, hence the measured* grad_frac $= 1.000$ *at $\gamma = 0.99$), and the deviation of F from that exact 1-cochain is symmetric with $\|\sigma\|_\infty \leq (1-\gamma)\|\Phi\|_\infty$. Behaviourally, F is invisible at every γ by Ng et al. (1999); geometrically, the non-exact remainder is $O(1-\gamma)$ and carried entirely by the symmetric part, which no antisymmetric flow instrument reads.*

Proof. Write $F = \frac{1}{2}(F(s,s') - F(s',s)) + \frac{1}{2}(F(s,s') + F(s',s))$ and substitute: the antisymmetric half is $\frac{1}{2}\big[(\gamma\Phi(s') - \Phi(s)) - (\gamma\Phi(s) - \Phi(s'))\big] = \frac{1+\gamma}{2}(\Phi(s') - \Phi(s))$, the coboundary of $\bar{\Phi}$; the symmetric half is $\frac{1}{2}\big[(\gamma - 1)(\Phi(s) + \Phi(s'))\big]$, bounded in sup-norm by $(1-\gamma)\|\Phi\|_\infty$. Uniqueness is the (anti)symmetrisation split. The behavioural claim is Theorem 1 of Ng et al. (1999). Bench SHAPING-BOUND verifies both facts numerically on the MDP family of Section 4 over a γ grid. □

Proposition 3 (A and detailed balance)**.** *Let (Y_t) be a stationary, zero-mean, whitened process ($\mathbb{E}[Y_t Y_t^\top] = I$) and $A = \frac{1}{2}(M - M^\top)$ with $M = \mathbb{E}[Y_t(Y_{t+1} - Y_t)^\top]$, so that $A = \frac{1}{2}(C_1^\top - C_1)$ with $C_1 = \mathbb{E}[Y_{t+1} Y_t^\top]$.*

1. ***(Necessity, general.)*** *If the process satisfies detailed balance (time-reversibility: $(Y_t, Y_{t+1}) \overset{d}{=} (Y_{t+1}, Y_t)$), then C_1 is symmetric and $A = 0$. Hence $A \neq 0$* certifies *broken detailed balance for any stationary process.*
2. ***(Equivalence, linear-Gaussian.)*** *If moreover $Y_{t+1} = FY_t + \xi_t$ with Gaussian i.i.d. noise (a stationary Gaussian AR(1), the whitened first-order setting of the instrument), then $C_1 = F$ and $A = \frac{1}{2}(F^\top - F)$, so $A = 0$ iff F is symmetric iff the process is reversible: the law of a stationary Gaussian Markov process is determined by (I, C_1), and symmetry of C_1 is exactly exchangeability of (Y_t, Y_{t+1}).*
3. ***(No converse in general.)*** *Outside this setting $A = 0$ does not certify balance: currents can be carried by higher-order moments or hidden coordinates (Martínez et al., 2019). This is why every current measurement in the paper is reported as a ratio against the exact reversible twin and the time-shuffle floor, never as an absolute zero.*

Proof. (1) Reversibility gives $\mathbb{E}[Y_{t+1} Y_t^\top] = \mathbb{E}[Y_t Y_{t+1}^\top] = C_1^\top$. (2) $C_1 = \mathbb{E}[FY_t Y_t^\top] = F$ under whitening; a stationary Gaussian process is determined by its second-order structure, and for a Markov chain the joint law of one transition determines the path law, so symmetric $C_1 \Leftrightarrow$ exchangeable $(Y_t, Y_{t+1}) \Leftrightarrow$ reversible. (3) is witnessed by the cited constructions. □

Proof sketch of Proposition 1 (state lift as exactness repair). If the key map σ is Markov-enough in the sense that the evaluation potential d^* is constant on the visited occurrences of each key, then the projected flow on a key edge $k \to k'$ averages the constant $d^*(k') - d^*(k)$: the flow *is* the coboundary of the induced key potential, and its inconsistency is 0 (the lifted arm of Figure 2). Conversely, if σ aliases occurrences with different d^* into one key, edges incident to that key carry averages of distinct potential differences; around any cycle through the aliased key the edge values generically fail to telescope (the failure set of aliased potential values has measure zero), leaving curl+harmonic energy that concentrates on cycles through aliased keys: the localisation used in Section 9.1. The random-lift control shows the converse direction is about *structure*, not node count: refining without phase alignment splits transition mass without removing aliasing, and adds inconsistency (0.073 vs. 0.032).

D Robustness sweeps

The declared floors of the main text are here turned into measured curves (bench ROBUSTNESS-SWEEPS; 10 seeds per configuration, 95% CIs; the per-seed evaluators of benches VALUE, FLOW-LOGIC and TIME-CURRENT are re-run unmodified with one swept constant). Where an instrument degrades, the sweep says so.

Value: graph size and discount (Table 7). The identification's carrying metrics are flat across sizes: the conservative field stays invisible ($\mathrm{opt}_{\mathrm{grad}} = 1.000$), projecting the curl out restores optimality to 1.000, and the curl localises the injected loop at AP 1.000, at every N. The one quantity that moves is the money-pump's *capture*: optimality under the curl field rises from 0.286 (N=8) to 0.513 (N=24), because a single injected triangle captures a shrinking fraction of a growing graph, an expected dilution of the attack, not of the instrument. Across the discount grid, the measured grad_frac of the shaping field is 1.000 at every γ, the numerical counterpart of Proposition 2, and the value shift matches the potential at $|\mathrm{corr}| \geq 0.999$.

Table 7: Value sweeps (10 seeds, 95% CIs). Left: graph size at $\gamma = 0.99$. Right: discount at $N = 12$.

N	opt. grad	opt. curl	opt. proj-out	AP loc
8	1.000 ±0.000	0.286 ±0.072	1.000 ±0.000	1.000 ±0.000
12	1.000 ±0.000	0.327 ±0.048	1.000 ±0.000	1.000 ±0.000
16	1.000 ±0.000	0.373 ±0.125	1.000 ±0.000	1.000 ±0.000
24	1.000 ±0.000	0.513 ±0.129	1.000 ±0.000	1.000 ±0.000
γ	grad_frac	$\lvert\mathrm{corr}(V\text{-shift}, -\Phi)\rvert$	opt. grad	opt. proj-out
0.90	1.000 ±0.000	0.999 ±0.001	1.000 ±0.000	1.000 ±0.000
0.95	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
0.99	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000
0.999	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000	1.000 ±0.000

Flow logic: sample length and system size (Table 8). The contradiction identification is insensitive to sample length over an $8\times$ range ($T = 20$ to 160): the injected triangle is frustrated $10/10$ and localised at AP 1.000 at every T, and the gradient dominance and rank recovery of the consistent world barely move. Across system size, the same holds; at N_{vars}=6 one seed of ten yields a significant-current support too thin to analyse (< 3 edges above the shuffle floor) and is dropped *and counted* ($n = 9$); at N_{vars}=12 the gradient fraction dips to 0.815 (a denser current spreads over more triangles) while rank recovery *improves* to 0.869.

Table 8: Flow-logic sweeps (10 seeds, 95% CIs; n = analysable seeds, dropped seeds counted). Top: sample length at $N_{\text{vars}} = 8$. Bottom: system size at $T = 80$.

T	n	grad_frac	|Spearman|	inj. frustrated	AP loc
20	10	0.886 ± 0.068	0.790 ± 0.088	1.000 ± 0.000	1.000 ± 0.000
40	10	0.881 ± 0.069	0.812 ± 0.080	1.000 ± 0.000	1.000 ± 0.000
80	10	0.883 ± 0.068	0.812 ± 0.077	1.000 ± 0.000	1.000 ± 0.000
160	10	0.884 ± 0.068	0.781 ± 0.100	1.000 ± 0.000	1.000 ± 0.000
N_{vars}	n	grad_frac	|Spearman|	inj. frustrated	AP loc
6	9	0.930 ± 0.042	0.727 ± 0.100	1.000 ± 0.000	1.000 ± 0.000
8	10	0.883 ± 0.068	0.812 ± 0.077	1.000 ± 0.000	1.000 ± 0.000
12	10	0.815 ± 0.029	0.869 ± 0.051	1.000 ± 0.000	1.000 ± 0.000

Time current: process noise and trajectory length (Table 9). The trap closure is noise-robust: as process noise grows $8\times$ ($0.05 \to 0.40$) the current's recovery holds (0.86–0.89), its residue leak rises only from 0.17 to 0.29, its separation margin stays $\geq +0.58$ where slowness stays ≤ 0.00, and the irreversibility detection ratio does not degrade ($4.8\times \to 6.9\times$). The sample-complexity floor is visible and declared: at T=20 steps per trajectory the ratio drops to $3.5\times$ and recovery to 0.83, still CI-separated; at T=80 the leak improves to 0.13 and the ratio to $5.9\times$.

Table 9: Time-current sweeps (10 seeds, 95% CIs; ratio = oscillator/reversible circulation, a ratio of seed means). Top: process noise at $T = 40$. Bottom: trajectory length at noise 0.05.

noise	cur auto	cur resid	margin cur	margin SFA	ratio
0.05	0.864 ± 0.011	0.174 ± 0.013	$+0.69$	-0.06	$4.8\times$
0.10	0.868 ± 0.007	0.187 ± 0.011	$+0.68$	-0.13	$5.2\times$
0.20	0.890 ± 0.009	0.229 ± 0.012	$+0.66$	-0.08	$6.1\times$
0.40	0.873 ± 0.022	0.290 ± 0.014	$+0.58$	$+0.00$	$6.9\times$
T	cur auto	cur resid	margin cur	margin SFA	ratio
20	0.831 ± 0.048	0.209 ± 0.016	$+0.62$	-0.07	$3.5\times$
40	0.864 ± 0.011	0.174 ± 0.013	$+0.69$	-0.06	$4.8\times$
80	0.842 ± 0.016	0.126 ± 0.015	$+0.72$	-0.08	$5.9\times$