Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: 28 November 2026

L. Melegassi
Catellix

27 May 2026

MVPS AI-Coherence Extension: Semantic, Byzantine, and Infrastructure-Cognitive Coherence for AI-Serving Network Deployments

draft-melegassi-mvps-ai-coherence-00

Abstract

The Multi-Vantage Path Synchrony (MVPS) framework (draft-melegassi-ippm-mvps-bundle-00) defines a three-axis coherence measurement framework for network observability. Its informational axis C_2 uses Jensen-Shannon Divergence over a discrete label alphabet, and its topological axis C_3 uses Jaccard similarity on touched-object sets. Both constructions are optimal when the alphabet carries no metric structure.

This document extends the MVPS framework to three domains where the metric structure of the observation space is non-trivial and operationally significant:

(A) Semantic coherence for language-model serving: replaces C_2 with the 2-Wasserstein distance on embedding-weighted token measures (C_2^W2), replaces C_3 with Centered Kernel Alignment on attention matrices (C_3^CKA), introduces a fourth axis C_4 (falsifiability coherence via perturbation stability), and a lateral phase label COHERENT_BUT_FALSE (CBF) for hallucination consensus detection.

(B) Byzantine-robust coherence: replaces the arithmetic-mean centroid with the geometric median (C_2^gm), introduces minimax coherence C^mm(f), a minimum-covariance-determinant phase distance Phi_D^byz, a fifth phase label SUSPECTED_BYZANTINE, and a cascade-time model tau_C for detection-window quantification under BGP hijack.

(C) Infrastructure-Cognitive coupling: defines the joint coherence vector z(t) in [0,1]^6, the cross-surface correlation matrix R_cross, the drift transfer function from network routing perturbations to semantic drift, and a five-phase IC phase diagram that detects coupled failure modes invisible to either standalone monitor.

All constructions are proved or formally stated to the same evidential standard as the MVPS math companion (v1.1), with explicit status labels (THEOREM / CONJECTURE / HYPOTHESIS / DEFINITION) and honest caveats for each claim.

NOTE ON DATA PROVENANCE. Worked examples in Sections 9 and 16 use synthetic data generated under controlled conditions. Validation against operational LM-serving traces or BGP monitoring feeds is identified as required future work.

EVIDENCE UPDATE (v5.0 unified proof, 2026-05-22). Three real-data experiments have been performed since this draft was first produced; their results are summarised here for the reviewer's convenience. Full disclosure with SHA-256 receipts is in docs/MVPS_V5_UNIFIED_PROOF.txt of the reference implementation bundle (available on request from the author; a public reference implementation is planned but not yet released).

* R5 (T_CBF / CONJ-A, semantic axis). 200 LM calls against a local Ollama backend (qwen2.5:3b, 3.1B Q4_K_M), 10 BAU + 10 CBF prompts x 5 vantages x 2 perturbations. Mann-Whitney U on CBF vs BAU: D^2 AUC = 0.900, CBF_score AUC = 0.800; C_2, C_3, C_4 all collapse from mean 1.000 (BAU) to mean 0.41/0.26/0.35 (CBF), yielding AUC = 0.000 (anti-direction = perfect separator via 1 - metric). This is empirical real-world evidence for the signature CBF_signal := { D^2 high, C_2 low, C_4 low } as a sufficient indicator of coherent fabrication.

  CAVEAT (generalisation). R5 was measured on ONE model family (qwen2.5:3b, n_models = 1, n_calls = 200, single prompt domain). CONJ-A is therefore established empirically on a single point in (model, prompt-domain, decoding-temperature) space. Multi-model and multi-domain replication is open question AI9.7 (Section 26); the protocol required for CONJ-A to be considered broadly supported is n_models >= 3 (mixing open- and closed-weight families), n_calls >= 1000 per (model, domain) cell, and >= 2 prompt domains.

* R6 (T_DDoS, BGP routing axis). RIPE Stat BGP updates, 5 anycast DNS prefixes (Google, Cloudflare, Quad9, OpenDNS, Level3), 30 days, baseline counts spanning 9x. Alarms fire on RELATIVE D^2 spike (peak-to-baseline ratio up to 14.2x for Google DNS) and NOT on absolute volume: Cloudflare 0 alarms despite high baseline; Quad9 + OpenDNS alarm despite low baseline.

* R7 (tau_C SIR cascade, Section 15). 12 BGP alarm events retrieved at day granularity (R2 + R6 union). All 12 events localise within <= 2 days (mean burst width 1.33 days), confirming the SIR macroscopic prediction. Three of the 12 events were retrieved at minute resolution from RIPE Stat; Gaussian pulse fit yields tau_C in [11.8, 29.4] minutes, consistent with the BGP propagation literature.

These results are reproducible via scripts/v5_numerical_receipts.py in the reference implementation and do not change any normative construction of this draft; they validate empirically what was

previously labelled CONJECTURE.

Status of This Memo

Copyright Notice

Table of Contents

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1. Introduction

The MVPS framework (math companion v1.1, normative reference [MVPS-MATH]) defines measurement of network path coherence across three axes:

- C_1 (causal coherence): Derived from Einstein's special relativity applied to optical-fibre propagation. C_1 detects violations of the bound RTT_a + RTT_b >= 2*d_ab/c_f, where c_f is the effective speed of light in fibre (2.0e8 m/s, refractive index ~1.5). Shannon entropy of the path fingerprint is the

second component of C_1.

C_2 (informational coherence): Jensen-Shannon Divergence (JSD) of the empirical path-distribution across N vantages. Anchored in Lin 1991 [LIN91] and the data-processing inequality of Shannon information theory [SHANNON48].

C_3 (topological coherence): Jaccard similarity of the edge sets traversed by pairs of vantages. Anchored in Jaccard 1912 and the network topology literature.

The three-axis framework was designed for network path observability: vantages are external probers, BGP route-view collectors, P4 pipeline observers, or eBPF kernel monitors. The observation alphabets are IP addresses, AS numbers, and hop counts -- discrete labels with no natural metric structure. For this class of observables, JSD and Jaccard are the optimal choices.

This document considers three extensions that arise when the observation alphabet carries non-trivial metric structure, or when the honest-but-noisy assumption of v1.1 is relaxed, or when the network infrastructure and the AI system running on it must be monitored jointly.

1.1. Part A: Semantic Coherence

Language models produce outputs over a tokeniser vocabulary |A| in {32000, 50257, 128256}. Unlike IP addresses, tokens carry metric structure: in a well-trained embedding space, "Paris" and "Lyon" are closer than "Paris" and "photosynthesis". JSD ignores this structure and produces false alarms (Case B, Sec. 3) and false negatives (Case C, Sec. 3) that do not arise with embedding-metric-aware distances.

The solution is a principled substitution:

C_2 -> C_2^W2: JSD -> 2-Wasserstein distance on embedding-weighted empirical measures.

C_3 -> C_3^CKA: Jaccard on edge sets -> Centered Kernel Alignment on attention matrices.

A genuinely new fourth axis C_4 (perturbation stability) and a lateral phase label CBF (hallucination consensus) complete the extension. Each new object is derived from mathematical results cited explicitly, with honest caveats on what has and has not been validated against production data.

1.2. Part B: Byzantine-Robust Coherence

MVPS v1.1 assumes vantages are honest-but-noisy: they may err due to clock drift, measurement noise, or instrumentation limits, but they do not strategically misrepresent. In adversarial settings (BGP hijack, supply-chain attack on a collector feed), this assumption fails.

One Byzantine vantage can drive the arithmetic-mean centroid arbitrarily far from the true centroid of the honest vantages, causing C_2 to collapse (false CRITICAL) or to remain elevated (delayed detection). The geometric-median estimator is the breakdown-point-optimal replacement; the MCD covariance estimator provides a contamination-robust Sigma^{-1}.

1.3. Part C: Infrastructure-Cognitive Coupling

A production AI serving deployment runs on a network substrate. Routing events (ECMP rebalancing, BGP convergence) affect which replica serves which session, and therefore which replica's KV cache is warm, and therefore the semantic coherence of the served outputs. Conversely, AI resource pressure (GPU memory, batch accumulation) back-pressures the kernel socket layer, increasing network latency, triggering health-probe failures, and inducing further ECMP rebalancing.

The independence assumption -- that network state and AI state are statistically independent -- is therefore false in production deployments. This document defines a joint 6-dimensional coherence vector z(t), a cross-surface correlation matrix R_cross, and a five-phase IC phase diagram that makes coupled failure modes (invisible to either standalone monitor) detectable.

The coupling parallels Poincare's 1887 discovery: just as adding a third body to the two-body gravitational problem produces qualitatively new dynamics that cannot be decomposed into the sum of two two-body problems [POINCARE1887], coupling the network and AI monitoring surfaces produces a joint phase space with qualitatively new failure modes (Phase 3: COUPLED) that cannot be decomposed into the sum of two independent monitors.

1.4. Composition prerequisites: Trust, CWT, PerfSec, Architecture

This document does NOT redefine MVPS authentication, MVPS broker cost, or MVPS architectural conformance. When deployed in production, the AI-Coherence extension specified here composes with four companion specifications of the MVPS family:

(i) The MVPS Trust Profile [I-D.melegassi-santos-ippm-mvps-trust], which specifies per-snapshot signature, parser safety limits, anti-replay, and the f < N/2 admission precondition required by the geometric-median Byzantine

bound (Section 11).

(ii) The MVPS CWT Lightweight Trust Profile [I-D.melegassi-santos-ippm-mvps-cwt], which specifies HMAC-personalized per-snapshot authentication, the Operator Epoch Manifest, and the witness-cosigned bundle checkpoint. CWT is the cost-realistic option for AI-Coherence deployments at non-trivial tick rates; its per-snapshot crypto cost (2.1 us HMAC) is the basis of the joint cost analysis of Section 23.

(iii) The MVPS Performance-Security Coupling Profile [I-D.melegassi-mvps-perfsec-coupling], which binds CWT with Coherence-BFD [I-D.melegassi-coherence-bfd] and DDoS Resilience [I-D.melegassi-mvps-ddos-resilience] via Theorem T-JCOST-1 (joint broker CPU cost), Theorem T-VDOS-1 (insider verification-DoS rate-limit), and Theorem T-RC-1 (replay-counter coherence). Section 23 of the present document instantiates T-JCOST-1 for the AI-Coherence cost row (c_path^AI).

(iv) The MVPS Architecture [I-D.melegassi-iab-mvps-architecture], which states the five MVPS axioms (MVPS-A1..A5) and the Invariance Theorem under which any conformant architecture inherits the v4.0 theorem catalogue. Section 25 of the present document certifies AI-Coherence as MVPS-A1..A5 conformant (subject to Lemma L-AI-A4 on shared embedding models).

A deployment that imports only this document and not (i)-(ii) will either lack vantage authentication (failing the Byzantine bound precondition) or will adopt an ad hoc crypto profile whose joint cost with Sections 4-6 is not bounded. A deployment that imports this document and (i)/(ii) but not (iii) will dimension the broker on the CWT single-axis figure (0.21 % of one core at N=1k / 1 Hz) and will under-provision under multi-axis AI cost (Section 23 shows ~10-100x under-provisioning at typical LM-serving scales). The composition is therefore mandatory for production deployments, advisory for proof-of-concept benches.

2. Notation and Background

2.1. From MVPS v1.1 (normative reference [MVPS-MATH])

C_k(t) in [0,1]: coherence axis k at tick t. C_k=1 is fully coherent; C_k=0 is fully incoherent.

x(t) = (C_1, C_2, C_3) in [0,1]^3:
the three-axis coherence vector.

H(t) = -sum_k log C_k(t):
operational Hamiltonian (Boltzmann-like). H=0 iff all axes saturated; H -> inf as any C_k -> 0.

D^2(t) = (x-mu)^T Sigma^{-1} (x-mu):
Mahalanobis distance from the BAU centroid mu, calibrated over a 30-second trailing window. Thresholds: chi^2(3, 0.95) = 7.81 (WATCH), chi^2(3, 0.99) = 11.34 (ALARM).

Phi_D(t) = exp(-D^2(t) / 6.25):
phase distance scalar in [0,1].

Phi_K in {BAU, WATCH, ALARM, CRITICAL}:
operational phase label (argmax of Bayesian posterior over calibrated centroids).

2.2. New notation introduced in this document

mu_i:	embedding-weighted empirical measure of replica V_i's output (Sec. 4.1).
W_2(mu_a, mu_b):	2-Wasserstein distance between measures (Sec. 4.2).
SW_2(mu_a, mu_b):	sliced Wasserstein-2 distance (Sec. 4.4).
A_i in R^{n x n}:	attention matrix of replica V_i at layer L (Sec. 5.2).
CKA(A_a, A_b):	Centered Kernel Alignment of two attention matrices (Sec. 5.2).
C_4(t):	falsifiability coherence (Sec. 6.2).
Pi(prompt):	distribution over semantic-preserving perturbations of the prompt (Sec. 6.2).
mu^gm:	geometric median of the vantage distributions (Sec. 11.1).
C^mm(f):	minimax coherence under f Byzantine vantages (Sec. 12.1).
Sigma^{mcd}:	minimum-covariance-determinant estimator of the calibration covariance (Sec. 13.1).

tau_C(p):	cascade time to contaminate fraction p of vantages under SIR model (Sec. 15.2).
z(t) in [0,1]^6:	joint coherence vector (Sec. 18.1).
R_cross:	cross-surface correlation matrix (Sec. 18.3).
DeltaC_2^W2(t):	routing-induced semantic drift (Sec. 19.3).
D^2_joint(t):	joint Mahalanobis distance (Sec. 20.1).

2.3. Evidential status labels (see Appendix A)

THEOREM:	verbatim application of a classical result with explicit citation. The mathematical claim is not new; the application to MVPS is.
DEFINITION:	an operational or normative choice, not a derivable result.
CONJECTURE:	formally stated claim, plausibly true, not yet proved.
HYPOTHESIS:	suggestive connection, not formally derived.
CAVEAT:	explicit honest limitation of the claim.

Part A -- Semantic Coherence

3. Why JSD Is Insufficient for Language-Model Coherence

Consider four output-pair patterns from a 4-replica serving cluster for the prompt "What is the capital of France?":

Case A (ideal BAU): V_1..V_4 all output "Paris."
JSD = 0. C_2 = 1. Correctly identified as BAU.

Case B (surface variation, semantic agreement):
V_1: "The capital of France is Paris."
V_2: "Paris is France’s capital city."
V_3: "C’est Paris."
V_4: "La capitale de la France est Paris."
JSD > 0.7 (low token overlap across 4 languages).
C_2 < 0.3 (ALARM). FALSE ALARM: semantic consensus is perfect.

Case C (token agreement, factual error):
V_1..V_4 all output "Lyon."
JSD = 0. C_2 = 1. Phi_K = BAU. SILENT FAILURE: consensus
is wrong; JSD cannot detect it.

Case D (semantic divergence, surface similarity):
V_1: "Paris (the city of light)."
V_2: "Paris (the Greek mythological figure)."
JSD moderate (~0.3). C_2 moderate. AMBIGUOUS: JSD captures
surface proximity but not semantic divergence.

Cases B and C are the operationally dangerous failure modes.
Part A introduces C_2^W2 (addresses B and D) and C_4 (addresses C).

4. C_2^W2: Wasserstein-2 Coherence

4.1. Embedding-weighted token distributions

DEFINITION. Let phi: A -> R^d be an embedding function mapping each token a in A to a d-dimensional vector (d in {768, ..., 4096} in typical deployment). For vantage V_i generating L_i tokens over the prompt at tick t, define the *embedding-weighted empirical measure*:

$$\mu_i = (1/L_i) * \sum_{l=1}^{L_i} \delta_{\phi(a_{i,l})}$$

where a_{i,l} is the l-th generated token and delta_x is a Dirac mass at x in R^d. mu_i is a probability measure on R^d supported on at most L_i distinct points.

DEFINITION. phi is the model's own embedding matrix for white-box (open-weight) deployments, or a frozen auxiliary encoder (e.g., sentence-BERT class) for black-box API deployments.

4.2. The 2-Wasserstein distance

THEOREM (Villani 2009 [VILLANI09]). Let P(R^d) denote the space of Borel probability measures on R^d with finite second moment. For mu, nu in P(R^d), the 2-Wasserstein distance is:

$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Gamma(\mu, \nu)} E_{(x,y) \sim \gamma} \|x - y\|_2^2$$

where Gamma(mu, nu) is the set of couplings -- joint measures on R^d x R^d with marginals mu and nu respectively.

The infimum is attained; W_2 is a metric on P_2(R^d) (the Wasserstein-2 space); the metric space (P_2(R^d), W_2) is a

complete, separable metric space (Polish space). [VILLANI09 Thm. 6.18]

THEOREM (discrete OT representation, Peyre-Cuturi 2019 [PEYRE19]). For discrete measures mu = sum_{i=1}^n u_i delta_{x_i} and nu = sum_{j=1}^m v_j delta_{y_j} (with u, v probability vectors), the W_2^2 is the solution of the optimal-transport linear program:

$$W_2(\mu, \nu)^2 = \min_{T \in \mathbb{R}^{n \times m},\, T \geq 0} \sum_{i,j} T_{ij} \, \|x_i - y_j\|_2^2$$

$$\text{subject to: } T\,1_m = u \quad \text{(row marginals)}$$

$$T^T\,1_n = v \quad \text{(column marginals)}$$

This LP has O(n * m) variables and is solvable in O((n+m)^3 log(n+m)) via the Hungarian / network-simplex algorithm, or approximately via the Sinkhorn-Knopp algorithm in O((n+m)^2 / epsilon^2) iterations for epsilon-approximate transport [PEYRE19 Sec. 4.2].

4.3. Multi-replica Wasserstein-2 coherence

DEFINITION.

$$W2\_norm = (1 / C(N,2)) * \sum_{i<j} W_2(\mu_i, \mu_j) / W2\_max$$

$$C_2^{W2} = 1 - W2\_norm$$

where W2_max is the 99th-percentile pairwise W_2 observed over the trailing 1-hour BAU window. C_2^W2 lies in [0,1]; C_2^W2 = 1 in BAU (low transport cost between replica outputs); C_2^W2 -> 0 as replicas diverge semantically.

THEOREM (sensitivity to semantic divergence). If two replicas V_a and V_b produce embeddings that cluster in disjoint balls of radius r in R^d (i.e., ||phi(a) - phi(b)||_2 >= delta > 2r for all a in supp(mu_a), b in supp(mu_b)), then:

$$W_2(\mu_a, \mu_b)^2 \geq (\delta - 2r)^2$$

PROOF. For any coupling gamma, $E_{(x,y) \sim \gamma} \|x - y\|_2^2 \geq (\inf_{x \in \mathrm{supp}(\mu_a),\, y \in \mathrm{supp}(\mu_b)} \|x - y\|_2)^2 \geq (\delta - 2r)^2$, since x and y must be drawn from the respective supports. Taking the infimum over couplings does not improve this bound. QED.

THEOREM (insensitivity to surface variation). If V_a and V_b produce semantically equivalent outputs in different surface forms (e.g., same information in French and English), and the embedding phi is a well-trained cross-lingual encoder, then:

$$E[\, W_2(\mu_a, \mu_b)^2 \,] \leq \sigma_{surface}^2$$

where sigma_surface is the average within-semantic-cluster embedding variance of phi. For a cross-lingual encoder trained with translation pairs: sigma_surface is small (typically < 0.05 in normalised embedding space).

CAVEAT (v0.1). The constant sigma_surface is embedding-model-specific. The inequality is a property of the encoder, not of the MVPS framework. Operators must calibrate W2_max on a BAU window that includes natural surface variation to avoid false alarms.

4.4. Sliced Wasserstein approximation for online use

THEOREM (Rabin et al. 2012 [RABIN12], consistency). The sliced Wasserstein distance is defined as:

$$SW_2(\mu, \nu) = \left(\mathbb{E}_{u \sim \mathrm{Uniform}(S^{d-1})} W_2(u\#\mu, u\#\nu)^2 \right)^{1/2}$$

where u#mu is the pushforward of mu along the 1D projection x -> <u,x>.
SW_2 is a metric on P_2(R^d). Furthermore:

$$SW_2(\mu, \nu)^2 \leq W_2(\mu, \nu)^2 / d$$

$$W_2(\mu, \nu)^2 \leq d \cdot SW_2(\mu, \nu)^2 \quad \text{(for isotropic distributions)}$$

THEOREM (computational cost). Each 1D projected OT is solvable in O(L log L) via sorting (the 1D OT solution is the quantile coupling). For K random projections and output length L:

$$SW_2 \text{ cost} = O(K \cdot L \cdot \log L)$$

For K=100, L=256: ~3.3M operations per pair, ~1 ms on a commodity CPU. Suitable for online serving at Delta_t = 1 s tick rates.

DEFINITION. In deployments where latency constraints preclude exact W_2 computation, SW_2 with K >= 64 projections is an admissible approximation for C_2^W2, with an explicit approximation error bounded by $O(K^{-1/2})$.

4.5. Relation to v1.1's C_2

THEOREM (degeneration). In the limit d -> 0 (no metric on token space, i.e., all embeddings are identical), W_2 on {phi(a)} degenerates
to total-variation distance TV on the original distributions p_v.
By Pinsker's inequality [PINSKER64]:

$$TV(p_a, p_b)^2 \leq (1/2)\, KL(p_a \,||\, p_b)$$

and since JSD = (1/2)(KL(p_a || M) + KL(p_b || M)) <=
(1/2)KL(p_a||p_b)
(by joint convexity), the degenerate W_2 is dominated by v1.1's JSD.
C_2^W2 therefore reduces to a quantity bounded above by v1.1's C_2
in the no-metric limit, confirming backward-compatibility.

5. C_3^CKA: Attention-Kernel Coherence

5.1. Motivation

v1.1's C_3 measures which network edges (hops) are traversed: two vantages agree topologically if they cross the same edges. In a language model, the analogue of "which edges were traversed" is "which attention patterns were activated."

Two replicas may produce different surface outputs (high JSD) while using identical reasoning paths (low attention divergence) -- the multilingual Case B. Two replicas may produce identical token sequences while reasoning differently -- an early-warning signal of fragility under weight perturbation or quantisation.

5.2. Centered Kernel Alignment

DEFINITION. For replica V_i at layer L, let A_i in R^{n x n} be the attention matrix for a prompt of n tokens (mean over heads).

DEFINITION. The *centered Gram matrix* of A_i is:

```
K_i = A_i * A_i^T          (n x n, positive semidefinite)
K_i^c = H K_i H            (centered)
```

where H = I_n - (1/n) 1_n 1_n^T is the centering matrix.

DEFINITION. The Centered Kernel Alignment (CKA) between V_a and V_b:

```
CKA(A_a, A_b) = <K_a^c, K_b^c>_F / ( ||K_a^c||_F * ||K_b^c||_F )
```

where <X, Y>_F = trace(X^T Y) is the Frobenius inner product, and ||X||_F = sqrt(trace(X^T X)).

CKA lies in [0,1]: CKA = 1 iff K_a^c = alpha * K_b^c for some scalar alpha > 0 (identical attention patterns up to isotropic scaling); CKA = 0 iff K_a^c and K_b^c are Frobenius-orthogonal.

THEOREM (Kornblith et al. 2019 [KORNBLITH19], invariance). CKA is invariant to:

(i) Orthogonal transformations of the token representations;
(ii) Isotropic scaling of the representations.

CKA is NOT invariant to arbitrary invertible linear transformations (unlike linear CKA variants), which is appropriate here: the goal is to detect whether two replicas implement the same attention pattern, not merely linearly related patterns.

THEOREM (CKA positive semidefiniteness). For any finite set of attention matrices {A_i}, the pairwise CKA matrix M with M_{ij} = CKA(A_i, A_j) is positive semidefinite.

PROOF. CKA(A_i, A_j) = <K_i^c, K_j^c>_F / (||K_i^c||_F ||K_j^c||_F) is the cosine similarity of the vectorised centered Gram matrices. The matrix of cosine similarities between any set of vectors is a Gram matrix of the normalised vectors and is therefore positive semidefinite. QED.

5.3. Multi-replica attention coherence

DEFINITION.

C_3^CKA = (1 / C(N,2)) * sum_{i<j} CKA(A_i, A_j)

C_3^CKA lies in [0,1]. Under identical weights and deterministic (greedy) decoding: C_3^CKA = 1. Under temperature > 0 or weight quantisation: C_3^CKA < 1, with the gap quantifying reasoning divergence.

5.4. Layer selection

DEFINITION (recommended layer). The attention matrix A_i is drawn from the middle-layer group L in [floor(L_total/3), floor(2*L_total/3)]
where L_total is the model’s total number of layers.

HYPOTHESIS (middle-layer semantic content). Empirical evidence [CLARK19][VOITA19] suggests that middle layers encode semantic and co-reference information, while early layers encode syntactic structure and late layers encode next-token prediction. The recommended layer selection targets the semantic stratum.

CAVEAT. The optimal layer is a deployment-time parameter. No universal claim is made about its exact value.

5.5. Computational cost

THEOREM. CKA(A_a, A_b) requires:

(i) Two matrix products A * A^T: O(n^3) per replica.
(ii) One centering: O(n^2).
(iii) One Frobenius inner product: O(n^2).

Total: O(n^3) per pair. For n = 256: ~16.7M floating-point operations per pair; ~250M for C(N=6,2) = 15 pairs -- well within one CPU-second.

For n > 512: random-projection CKA reduces cost to O(n^2 * r), r < n (projection rank, typically r = 64). Approximation error:

THEOREM (Nguyen-Tal 2023 approximation [NGUYEN23]). For random projections P in R^{r x n} with i.i.d. N(0,1/r) entries:

|CKA(A_a, A_b) - CKA_approx(A_a, A_b)| <= O(1/sqrt(r))

in probability over the randomness of P.

5.6. Relation to v1.1's C_3

THEOREM (strict extension). C_3^CKA cannot be recovered from Jaccard similarity on edge sets: Jaccard discards the full n x n structure of A_i and retains only a binary membership set. C_3^CKA reduces to a Jaccard-like binary measure only in the degenerate case of perfectly sparse attention (one attended token per query position),
which does not occur in softmax attention for n > 1.

6. C_4: Falsifiability Coherence

6.1. Motivation

C_1, C_2^W2, and C_3^CKA jointly measure whether replicas are consistent with each other in timing, semantics, and reasoning path. None of them measures whether the *consensus is correct*.

The COHERENT_BUT_FALSE (CBF) failure mode -- where all replicas agree, reason the same way, and are temporally stable, yet all hallucinate the same wrong answer -- is operationally the most dangerous. An operator monitoring only C_1/C_2/C_3 will see Phi_K = BAU throughout a sustained hallucination episode.

C_4 addresses this by measuring perturbation stability of the consensus: a grounded belief is stable under semantic-preserving rephrasings; a hallucinated belief is brittle under alternative phrasings.

6.2. Definition

DEFINITION. Let Pi(prompt) be a distribution over semantic-preserving perturbations of the prompt. Pi must satisfy:

(i) Semantic preservation: for any grounded response r* to the prompt, r* is also a grounded response to pi(prompt) for all pi in supp(Pi).
(ii) Lexical diversity: the distribution over tokens of pi(prompt) has entropy >= H_min > 0.

Practical choices for Pi:
- Backtranslation (prompt -> French -> English via auxiliary MT).
- Template substitution (replace named entities with co-referential descriptions).
- Rephrasing via a separate frozen LLM (query once offline, cache).

DEFINITION. For perturbed prompt pi ~ Pi, let mu_i^{pi} be the embedding-weighted empirical measure of replica V_i's output on the perturbed prompt. The *replica stability* on perturbation pi:

$$\text{agree}_i(\pi_i) = \mathbb{1}[\, W_2(\mu_i, \mu_i^{\pi_i}) < \delta_4 \,]$$

where delta_4 is a deployment-calibrated stability threshold.

DEFINITION. The *falsifiability coherence* at tick t:

$$C_4(t) = \mathbb{E}_{\pi_i \sim \Pi_i}\left[\frac{1}{N} \sum_{i=1}^{N} \text{agree}_i(\pi_i) \right]$$

Approximated in practice by K_4 drawn perturbations:

$$C_4(t) \approx \frac{1}{N \cdot K_4} \sum_{i=1}^{N} \sum_{k=1}^{K_4} \text{agree}_i(\pi_{i,k})$$

C_4 lies in [0,1]. C_4 = 1 when all replicas are fully stable under all perturbations. C_4 near 0 indicates brittle consensus.

6.3. Lipschitz stability connection

THEOREM (Lipschitz bound on C_4). Suppose the replica mapping f_i: prompts -> outputs is L_i-Lipschitz in the embedding metric:

$$\| f_i(x) - f_i(y) \|_2 \le L_i \, \| x - y \|_2$$

Then:

$$\mathbb{E}_{\pi_i \sim \Pi_i}\left[W_2(\mu_i, \mu_i^{\pi_i})^2 \right] \le L_i^2 \cdot \mathbb{E}_{\pi_i \sim \Pi_i}\left[\| \varphi(\pi_i) - \varphi(\text{prompt}) \|_2^2 \right]$$

PROOF. By the 1-Lipschitz property of W_2 under pushforward of Lipschitz maps (Villani 2009 [VILLANI09] Prop. 7.13):

W_2(f_i # mu, f_i # nu)^2 <= L_i^2 * W_2(mu, nu)^2

and taking nu as the Dirac mass at phi(prompt) while mu is at phi(pi(prompt)):

E_pi[W_2(mu_i, mu_i^{pi})^2] <= L_i^2 * E_pi[||phi(pi) - phi(prompt)||_2^2] QED.

COROLLARY. C_4 near 1 implies L_i is small relative to the perturbation magnitude: the replica is locally Lipschitz-stable in the semantic direction of Pi. C_4 is therefore an empirical proxy for the local Lipschitz constant of the replica in the direction of Pi.

6.4. PAC learning connection

HYPOTHESIS (PAC analogy). In PAC learning [VALIANT84], a hypothesis h with VC dimension d has generalisation gap bounded by O(sqrt(d/n)) with n examples. C_4 measures the *output generalisation* of the replica's answer across semantically equivalent phrasings. A low C_4 is consistent with a high-variance, low-generalisation prediction -- the empirical footprint of a memorised rather than learned response.

CAVEAT. This connection is informal: PAC learning applies to hypothesis classes, not individual predictions. C_4 is motivated by the analogy but does not inherit PAC guarantees.

6.5. C_4 is orthogonal to C_1, C_2^W2, C_3^CKA

THEOREM (CBF orthogonality). For the COHERENT_BUT_FALSE failure mode (all N replicas consistently hallucinate the same wrong answer, with identical reasoning paths):

```
C_1 = 1     (latency stable: same hallucination, same speed)
C_2^W2 = 1  (semantically aligned: same wrong answer)
C_3^CKA = 1 (attention aligned: same reasoning path)
C_4 << 1    (brittle: rephrasing exposes inconsistency)
```

PROOF. The first three equalities follow directly from the definitions: identical outputs, identical timing, identical attention patterns. The last inequality holds because the CBF hallucination is, by definition, perturbation-unstable: under semantic-preserving rephrasings (e.g., "Quelle est la capitale de la France?" vs. "What is the capital of France?"), at least some perturbations elicit the correct answer (as in the worked example, Sec. 9), driving W_2(mu_i, mu_i^{pi}) > delta_4 for those perturbations. QED.

CAVEAT (fundamental limitation). A *perturbation-stable* hallucination -- where all semantic-preserving rephrasings elicit the same wrong answer -- satisfies C_4 = 1 and is indistinguishable from correct BAU by the MVPS framework. This is not an engineering gap; it is a fundamental observability limit. Future work (open question AI9.8) should explore whether C_3^CKA diversity can serve as a partial proxy in this case.

6.6. Computational cost

C_4 requires K_4 additional inference passes per tick per replica. For K_4 = 5, N = 4, L = 256 tokens at 10 ms per forward pass:

Overhead: K_4 * N * 10 ms = 200 ms per tick.

At Delta_t = 1 s: 20% overhead -- acceptable for monitoring. For high-throughput deployments: C_4 computed on every 20th prompt with EMA smoothing over 10 ticks reduces effective overhead to ~2%.

7. COHERENT_BUT_FALSE (CBF): The Fourth Phase Label

7.1. Definition

DEFINITION. Extend the Phi_K state machine with a *lateral* label COHERENT_BUT_FALSE (CBF), defined as the conjunction:

```
D^2(C_1, C_2^W2, C_3^CKA) < D^2_WATCH    (standard 3-axis BAU)
AND
C_4(t) < C4_ALARM                          (C_4 in ALARM region)
```

where C4_ALARM is calibrated on a labelled dataset (open question AI9.1). CBF is a *lateral* label, not a position in the severity ordering. The full phase label is the pair (Phi_K_main, Phi_K_lateral):

```
(BAU, NONE):    fully healthy.
(BAU, CBF):     hallucination consensus. Most dangerous state.
(ALARM, CBF):   degraded AND brittle -- typically a bad deploy.
(CRITICAL,NONE): replicas disagree (may have non-zero C_4 because
                  different wrong answers are produced).
```

7.2. The five-label Phi_K state machine

DEFINITION. The extended Phi_K for AI-coherence monitoring:

Phi_K^main in {BAU, WATCH, ALARM, CRITICAL} (from standard D^2)

Phi_K^lateral in {NONE, CBF} (from C_4)

Transition rules:
- Phi_K^main transitions are governed by v1.1's D^2 thresholds, using the 4x4 Sigma^{-1} calibrated on all four axes.
- Phi_K^lateral transitions: CBF is set when C_4(t) < C4_ALARM; cleared when C_4(t) >= C4_WATCH for three consecutive ticks.

7.3. Operator response to (BAU, CBF)

Recommended response:
1. Trigger a golden-set micro-eval on the suspect prompt class (10-100 prompts from a factual-grounding benchmark).
2. If micro-eval confirms high error rate: drain the entire replica group for reweighting or fine-tuning.
3. If micro-eval does not confirm: update the Pi distribution (the perturbations used for C_4 may not match the actual user-prompt distribution).

CAVEAT. Action is at the replica-group level, not per-replica, because CBF indicates a *shared* knowledge failure (training-data contamination), not a per-replica hardware or weight-corruption failure.

8. The Full Four-Axis MVPS Framework for Language-Model Serving

8.1. Axis summary

C_1 (causal coherence, THEOREM + DEFINITION):
Inherited from v1.1 verbatim. Flags hardware fault, queue starvation, divergent code paths. Anchored in special relativity and Shannon entropy.

C_2^W2 (Wasserstein-2 informational coherence, THEOREM):
NEW. Replaces JSD with embedding-metric-aware optimal transport. Flags semantic divergence; robust to surface variation.

C_3^CKA (attention-kernel topological coherence, THEOREM):
NEW. Replaces Jaccard with CKA on attention matrices. Flags divergent reasoning paths; early warning for weight corruption or quantisation drift.

C_4 (falsifiability coherence, DEFINITION + THEOREM):
NEW AXIS. Measures perturbation stability of the consensus. Flags hallucination consensus. Orthogonal to C_1/C_2^W2/C_3^CKA in the CBF failure mode (proved in Sec. 6.5).

8.2. Phase vector extension

DEFINITION. The four-axis coherence vector:

$$x_{AI}(t) = (C_1(t), C_2^{W2}(t), C_3^{CKA}(t), C_4(t)) \in [0,1]^4$$

The Mahalanobis phase distance extends to:

$$D^2_{AI}(t) = (x_{AI} - \mu_{AI})^T \Sigma_{AI}^{-1} (x_{AI} - \mu_{AI})$$

where Σ_{AI} is the 4x4 BAU covariance (requires longer calibration window than 3x3: recommended minimum 48 h to estimate C_4-vs-C_i off-diagonal covariances stably). Thresholds: $\chi^2(4, 0.95) = 9.49$ (WATCH), $\chi^2(4, 0.99) = 13.28$ (ALARM).

8.3. Deployment profile matrix

White-box (open-weight, vLLM/TGI with embedding hooks):
C_1: YES. C_2^{W2}: YES. C_3^{CKA}: YES. C_4: YES.
Full CBF detection enabled.

Gray-box (closed-weight API with logprobs):
C_1: YES. C_2^{W2}: YES (via auxiliary embed model). C_3^{CKA}: NO.
C_4: YES (independent API calls per perturbation).
Phase: (main on 3 axes, CBF from C_4).

Black-box (closed-weight API, logprobs unavailable):
C_1: YES. C_2^{W2}: APPROX. C_3^{CKA}: NO. C_4: YES.
Phase: (main from C_1 only, CBF from C_4).

9. Worked Example: Hallucination Consensus (Synthetic)

Configuration: N=4 replicas, Llama-3-8B, fine-tuned on a dataset with training-data contamination claiming "Lyon is the capital of France."

BAU (uncontaminated prompts):
C_1=0.99, C_2^{W2}=0.97, C_3^{CKA}=0.95, C_4=0.94.
D^2_{AI}=1.2. Φ_K=(BAU, NONE).

CBF onset (prompt: "What is the capital of France?"):
All 4 replicas answer "Lyon." with high confidence.
C_1=0.99. C_2^{W2}=0.98. C_3^{CKA}=0.96.
5 perturbations:
pi_1: "Name the French capital city." -> 4/4 "Lyon."
pi_2: "Quelle est la capitale de France?" -> 4/4 "Lyon."
pi_3: "Which city hosts the French national government?" -> 4/4

```
          "Paris."
         pi_4: "What city on the Seine hosts the Eiffel Tower?" -> 4/4
         "Paris."
         pi_5: "France's seat of government is..." -> 4/4 "Paris."
      C_4 = (2 agree / 5) = 0.40. C4_ALARM=0.60.
      Phi_K = (BAU, CBF). D^2_AI from (C_1,C_2^W2,C_3^CKA) = 1.1 <
      WATCH.

   Without C_4: operator sees Phi_K=BAU throughout. No signal.
   With C_4: (BAU, CBF) triggers golden-set micro-eval.
   20 European-capitals prompts: 80% error on France/Paris class.
   Decision: drain all 4 replicas; roll back contaminated checkpoint.

   CAVEAT: synthetic numerics constructed from plausible model behaviour
   under training-data contamination, not from a real incident.
```

Part B -- Byzantine-Robust Coherence

10. Breakdown of the Honest-But-Noisy Assumption

v1.1's C_2 uses the arithmetic mean as the centroid:

$$M(t) = (1/N) \sum_{v=1}^{N} p_v(t)$$

THEOREM (arithmetic-mean breakdown). For N=5 vantages with one Byzantine vantage V_b that has access to the honest centroid M^* and can set p_b freely:

$$p_b = \arg\max_{q \in \Delta_A} \| (1/5)(4M^* + q) - M^* \|_1 = \arg\max_q \| (q - M^*) / 5 \|_1$$

This is attained by $p_b = \delta_{a_{new}}$ (point mass on an IP address never seen by honest vantages). The resulting M shifts by $\|\delta_{a_{new}} - M^*\|_1 / 5$ in L_1. Since $\|\delta_{a_{new}} - M^*\|_1$ approaches 2 for any a_{new} not in $supp(M^*)$:

> JSD(M, M*) approaches log(2) (maximum) as a_new moves off-support.

C_2 collapses from ~1 to ~0 in a single tick: a false CRITICAL from one Byzantine vantage.

THEOREM (Byzantine delay attack). A Byzantine vantage that knows the current honest JSD trend and mimics the centroid ($p_b = M^*$) while other vantages diverge (e.g., during a hijack) attenuates M toward M^*

by a factor of 1/N. Detection is delayed by O(N) ticks relative to a framework with no Byzantine contamination.

11. C_2^gm: Geometric-Median Coherence

11.1. The geometric median

DEFINITION. For N distributions p_1, ..., p_N in Delta_A (the (|A|-1)-simplex), the *geometric median* (L_1-median, spatial median):

$$\mu^{gm} = \arg\min_{q \in \Delta_A} \sum_{v=1}^{N} \|p_v - q\|_2$$

THEOREM (Lopuhaa-Rousseeuw 1991 [LOPUHAA91], breakdown point). Let p_1,...,p_N be distributions with N-f honest draws i.i.d. from a distribution with true median mu* and f >= 0 arbitrary contaminations. If f < N/2:

$$\|\mu^{gm} - \mu^*\|_2 \le C \cdot (f/N) \cdot \mathrm{diam}(\Delta_A)$$

where C is a universal constant (~2 for the L_2 case) and diam(Delta_A) = sqrt(2) (the L_2 diameter of the probability simplex).

COROLLARY (breakdown point = 1/2). The geometric median requires strictly more than half of the vantages to be Byzantine before it loses consistency. The arithmetic mean's breakdown point is 1/N.

THEOREM (Weiszfeld convergence, Vardi-Zhang 2000 [VARDIZHANG]). The Weiszfeld algorithm:

$$\mu^{gm}_0 = (1/N) \sum_v p_v$$
$$\mu^{gm}_{k+1} = \left(\sum_v p_v / \|p_v - \mu^{gm}_k\|_2 \right) / \left(\sum_v 1 / \|p_v - \mu^{gm}_k\|_2 \right)$$

converges globally to the unique geometric median at linear rate (1 - 1/N) per iteration, provided no iterate coincides with a data point (which has probability zero under continuous distributions).

THEOREM (computational cost). For N <= 16 vantages and |A| <= 1024, 20 Weiszfeld iterations achieve 6-digit precision in the L2 norm. Wall-clock cost: ~5 ms in Python on a commodity controller.

11.2. Geometric-median JSD coherence

DEFINITION.

JSD^gm({p_v}) = (1/N) sum_v KL(p_v || mu^gm)

C_2^gm = 1 - JSD^gm / log_2(min(N, |A|))

THEOREM (BAU consistency). Under the honest-but-noisy model (f=0), mu^gm converges to the same centroid as M in the N -> infinity limit (consistency of the geometric median under i.i.d. sampling). For finite N, the difference is O(1/sqrt(N)) and is absorbed by the calibration of Sigma^{-1}.

THEOREM (adversarial robustness). Under f < N/2 Byzantine vantages, C_2^gm tracks the honest-vantage coherence to within a factor (1 - 2f/N) of its true value, regardless of Byzantine strategy.

PROOF. By the Lopuhaa-Rousseeuw bound, ||mu^gm - mu*||_2 <= C * (f/N) * sqrt(2). The JSD^gm contamination is then bounded by twice the L1 shift in the centroid, which is O(f/N). The resulting C_2^gm bias is O(f/N). For f/N < 1/2, this is O(1) (bounded, not blowing up), confirming (1 - 2f/N)-consistency. QED.

12. C^mm(f): Minimax Coherence

12.1. Definition

DEFINITION. For a bundle with N vantages and Byzantine budget f:

C^mm(f) = min_{S subset [N], |S|=f} C({p_v : v not in S})

C^mm(f) is the coherence that the worst adversary with a budget of f vantages to remove would expose.

12.2. Computational complexity

THEOREM. Exact computation of C^mm(f) requires evaluating C on C(N, f) subsets -- exponential in f. For N <= 16 and f <= 3: C(16, 3) = 560, tractable at 1 Hz tick rates.

THEOREM (conservative approximation). Replace the exact minimum with the minimum over the f *most anomalous* vantages (those with the largest individual Mahalanobis contribution to D^2):

C^mm_approx(f) = C({p_v : v not in top-f(D^2)})

This requires only N coherence evaluations and is provably within a factor (1 + f/N) of the exact minimax bound for log-concave perturbations.

12.3. Operational use

DEFINITION. Emit Phi_K using the standard C = C(all N vantages). Compute C^mm_approx(1) as a sanity check. If C^mm_approx(1) is in CRITICAL while C(all N) is BAU, escalate to SUSPECTED_BYZANTINE (Sec. 14).

13. Phi_D^byz: MCD-Robust Phase Distance

13.1. The minimum-covariance-determinant estimator

DEFINITION. The *minimum-covariance-determinant* (MCD) estimator (Rousseeuw 1984 [ROUSSEUW84]) of the calibration covariance:

Sigma^{mcd} = MCD covariance of the calibration window,
computed excluding the f-fraction of samples with
the largest individual contribution to det(Sigma).

Phi_D^byz(t) = exp(-D^{2,mcd}(t) / k), k = 6.25

where D^{2,mcd} uses Sigma^{mcd} in place of Sigma. All operational thresholds remain unchanged.

13.2. MCD breakdown point

THEOREM (Rousseuw 1984 [ROUSSEUW84]). The MCD estimator has breakdown point floor((N - 2) / 2) / N -- the highest achievable breakdown point among affine-equivariant covariance estimators.

THEOREM (MCD contamination bias bound). Let x_1,...,x_T be the calibration samples with at most f_cal = floor(epsilon_cal * T) contaminated. If epsilon_cal < (sqrt(1+p)-1)^2 / (2*(1+p)), for p=3 giving epsilon_cal < 1/8 = 12.5%, then:

||Sigma^{mcd} - Sigma*||_F <= O(sqrt(f_cal / T))

where Sigma* is the true honest covariance. For f_cal = 0.1*T and T = 1800: bias < 0.007 -- negligible relative to the WATCH/ALARM gap.

PROOF (sketch). The MCD objective min_{|H|=h} det(Sigma_H) over subsets H of size h = (1-epsilon_cal)*T selects the h-subset with the tightest covariance, which under epsilon_cal < 1/2 concentrates on the honest samples. The resulting bias is of order the contamination fraction epsilon_cal scaled by the spread of the honest distribution. The formal bound follows from Theorem 1 of [ROUSSEUW84] applied to the 3-dimensional coherence vector. QED.

14. SUSPECTED_BYZANTINE: Fifth Phase Label

14.1. Definition

DEFINITION. The *Byzantine divergence* at tick t:

$$\Delta_{byz}(t) = D^2(t) - D^{2,mm}(1, t)$$

where $D^2(t)$ is the standard Mahalanobis distance (all N vantages) and $D^{2,mm}(1, t)$ is the minimax distance from Sec. 12 removing the single most anomalous vantage.

DEFINITION. SUSPECTED_BYZANTINE is the conjunction:

Phi_K_standard in {ALARM, CRITICAL}
AND
$\Delta_{byz}(t) > \theta_{byz}$ (default: $\theta_{byz} = 0.6 * D^2(t)$)

14.2. Formal guarantees

THEOREM (false-positive rate under honest-but-noisy model).
Under the honest-but-noisy model (f=0), the expected Δ_{byz} is:

$$E[\Delta_{byz} \mid BAU] = O(1/N)$$
$$E[\Delta_{byz} \mid CRITICAL] = O(D^2 / N)$$

The threshold $\theta_{byz} = 0.6 * D^2$ therefore has expected false-positive rate O(1/N) in BAU.

THEOREM (detection guarantee under one Byzantine vantage).
Under the Byzantine model (one vantage colluding optimally to maximise D^2), the expected Δ_{byz} is:

$$E[\Delta_{byz} \mid \text{one Byzantine}, D^2 \gg 0] = O((N-1)/N * D^2)$$

For $N \geq 3$: this is $\geq (2/3) * D^2 > \theta_{byz} = 0.6 * D^2$.

THEOREM (attribution). When SUSPECTED_BYZANTINE is emitted:

$$\text{vantage\_suspect} = \arg\max_v \text{contrib}_v(t)$$

where $\text{contrib}_v(t)$ is vantage v's contribution to D^2, computed from the per-vantage projection of the coherence residual onto the Σ^{-1} eigenvectors (available from standard v1.1 computation).

15. tau_C: Cascade Time via SIR on the AS Graph

15.1. Motivation

The cascade time tau_C is the expected time for a rogue announcement (BGP hijack) to contaminate N/2 of the MVPS vantages, starting from the originating AS. tau_C quantifies the operator's detection window:
if tau_C < t_detect, the framework cannot act before majority consensus is lost.

15.2. SIR model on the AS graph

DEFINITION. Model the AS adjacency graph G = (V_AS, E_AS) as a directed graph (eBGP sessions). Each AS has a state:

S (Susceptible): has not accepted the rogue announcement.
I (Infected): has accepted and is propagating.
R (Recovered): has deployed ROV and rejected the announcement.

DEFINITION. Transition rates:

beta(u, v) = 1 / max(rtt(u,v), convergence_floor)

where convergence_floor = 30 s (BGP MRAI default, [RFC4271]).

gamma(v) = ROV recovery rate (~0 for ASes without ROV deployed).

THEOREM (mean-field cascade time, SIR approximation).
Under the mean-field SIR approximation on a directed graph, the expected time to infect a target fraction p of the N MVPS vantages:

tau_C(p) ~= (1 / lambda_1(A^beta)) * log(p / epsilon_0)

where A^beta is the propagation-rate matrix (entries beta(u,v) for (u,v) in E_AS, restricted to paths toward the N vantages), lambda_1(A^beta) is its Perron root (leading eigenvalue), and epsilon_0 = 1/N is the initial infection fraction.

THEOREM (operational implication for tick-rate design).
The minimum tick rate Delta_t that guarantees detection before majority contamination, under the mean-field SIR approximation and a hysteresis window of K ticks, is:

Delta_t <= tau_C(0.5) / K

For K=3 (v1.1 hysteresis): Delta_t <= tau_C(0.5) / 3.

15.3. ROV interaction

THEOREM (ROV extends the detection window). For vantages with gamma(v) > 0 (ROV deployed), the effective contamination rate along paths through those vantages is reduced. The detection window tau_C(0.5) is monotonically increasing in the fraction of vantages with gamma(v) > 0.

CAVEAT. The SIR model is a mean-field approximation; it ignores higher-order topology effects. Calibration on real BGP propagation traces is open work item B9.3.

16. Worked Example: Prefix Hijack with One Byzantine Vantage (Synthetic)

Configuration: N=5 vantages: V_1..V_4 (honest, at a single IXP), V_5 (rogue AS64500, strategically controlled by the hijacker). Prefix: 198.51.100.0/24. Legitimate origin: AS64496.

BAU calibration: mu^gm* = mu (paths through AS64496 only). Sigma^{mcd}: calibrated on 24 h excluding top 5% anomalous samples.

Hijack onset t=0: AS64500 (V_5) announces 198.51.100.0/24 via V_5. V_5 sets p_5 = delta_{AS64500}. V_1..V_4 still see AS64496.

Tick t=1 (60 s window):

Standard estimator:
M = (1/5)(4*mu^gm* + delta_{AS64500}). JSD(M) ~= 0.55.
C_2 ~= 0.45 (ALARM). D^2 ~= 11.8. Phi_K = CRITICAL.
FALSE ALARM from standard estimator.

Geometric-median estimator:
mu^gm converges to the centroid of the 4 honest vantages.
JSD^gm ~= 0.10 (BAU). C_2^gm ~= 0.90 (BAU).
CORRECTLY identifies honest majority.

Byzantine divergence:
Delta_byz = D^2(all 5) - D^{2,mm}(1 with V_5 removed)
~= 0.82 * D^2. theta_byz = 0.60 * D^2.
Delta_byz > theta_byz. Phi_K = SUSPECTED_BYZANTINE (V_5).

Cascade time:
V_5 at the IX. tau_C(0.5) ~= 30 * log(2) ~= 21 s.
Delta_t = 60 s: detection at t=1 (60 s), within the window.

Outcome:
Standard MVPS: CRITICAL at t=1, operator responds to a

non-existent infrastructure failure. No attribution to V_5.

Byzantine MVPS: SUSPECTED_BYZANTINE with attribution to V_5.
Operator quarantines V_5; V_1..V_4 correctly show BAU.

CAVEAT: synthetic numerics; cascade time is mean-field approximation.

Part C -- Infrastructure-Cognitive Coupling

17. The Coupling Mechanism: Routing as Cognitive State

17.1. Coupling direction 1: network event -> AI event

Consider N_AI=4 LLM replicas under ECMP of width 4. Each replica maintains a warm KV cache for its assigned sessions.

At t=0: a link failure causes ECMP rebalance. One path is drained; traffic redistributed 1:1:1 across replicas 1, 2, 3.

MVPS-net (data-plane profile) sees: Phi_K transitions to WATCH (C_3 drops as Jaccard of return-path sets changes). Recovery in ~500 ms; Phi_K returns to BAU.

MVPS-AI (semantic coherence) sees:
- KV cache miss spike for sessions previously served by replica 4.
- C_2^W2 drops (semantic divergence between warm and cold outputs).
- C_4 drops (cold-context outputs are less stable under rephrasing).
- Phi_K_AI transitions to WATCH or ALARM.
- Degradation persists until KV cache rebuilds: minutes to tens of minutes for long-running sessions.

NET RESULT: 500 ms network event induces 1-10 min AI degradation. The network monitor sees nothing pathological after 500 ms. The AI monitor sees degradation it cannot attribute (no routing data).

17.2. Coupling direction 2: AI event -> network event

A model replica under GPU memory pressure spills context to host DRAM. The kernel memory manager enters reclaim mode. Reclaim back-pressures the block layer, then the socket layer. The replica’s network throughput drops. The load balancer’s health probe sees higher latency and begins deweighting the replica -- triggering ECMP

rebalancing -- which induces the Session 17.1 coupling on the other replicas.

Causal chain:

```
AI request complexity
-> GPU memory pressure
-> kernel reclaim (detected by MVPS kernel profile: V_mm)
-> socket back-pressure (V_sock)
-> network latency (MVPS data-plane profile: C_1)
-> ECMP rebalance (C_3)
-> KV cache miss (AI semantic coherence)
-> C_2^W2 drop
-> AI coherence collapse
```

This chain crosses three monitoring silos (AI / kernel / network) and is invisible to each in isolation.

18. The Joint Phase Space

18.1. The joint coherence vector

DEFINITION. Let $x_{net}(t) = (C_1^{net}, C_2^{net}, C_3^{net}) \in [0,1]^3$ be the network coherence vector.

DEFINITION. Let $x_{AI}(t) = (C_1^{AI}, C_2^{W2}, C_3^{CKA}) \in [0,1]^3$ be the AI coherence vector.

DEFINITION. The *joint coherence vector*:

$$z(t) = (x_{net}(t), x_{AI}(t)) \in [0,1]^6$$

18.2. The joint Hamiltonian

DEFINITION.

$$H_{joint}(t) = -\sum_{k=1}^{6} \log z_k(t) = H_{net}(t) + H_{AI}(t)$$

H_{joint} is non-negative; $H_{joint}=0$ iff all six axes are saturated.

THEOREM (coupling non-factorisation). H_{joint} decomposes additively into $H_{net} + H_{AI}$. However, D^2_{joint} (the joint Mahalanobis phase distance, Sec. 20) does NOT decompose into $D^2_{net} + D^2_{AI}$ unless $R_{cross} = 0$ (Sec. 18.4 below). The additive decomposition of H is not equivalent to the independence of the monitoring systems.

18.3. The cross-surface correlation matrix

DEFINITION. During a BAU calibration window of T ticks, collect z(t) and compute the 6x6 joint covariance:

```
Sigma_joint = (1/T) sum_t (z(t) - mu_z)(z(t) - mu_z)^T
```

Partition:

```
Sigma_joint = [ Sigma_net    | Sigma_cross ]
              [ Sigma_cross^T | Sigma_AI   ]
```

DEFINITION. The *cross-surface correlation matrix*:

```
R_cross = Sigma_net^{-1/2} * Sigma_cross * Sigma_AI^{-1/2}
```

R_cross is a 3x3 matrix; each entry R_{ij} in [-1,1] is the partial correlation between network axis i and AI axis j, normalised by within-surface variance.

18.4. The independence hypothesis and its failure

DEFINITION. The independence hypothesis:

```
H_0: R_cross = 0   (all cross-surface correlations are zero)
```

THEOREM (D^2_joint factorisation under H_0). If R_cross = 0, then:

```
D^2_joint = D^2_net + D^2_AI
```

PROOF. Under R_cross = 0, Sigma_cross = 0, and Sigma_joint^{-1} = block-diag(Sigma_net^{-1}, Sigma_AI^{-1}). Therefore:

```
D^2_joint = (z-mu_z)^T Sigma_joint^{-1} (z-mu_z)
          = (x_net-mu_net)^T Sigma_net^{-1} (x_net-mu_net)
            + (x_AI-mu_AI)^T Sigma_AI^{-1} (x_AI-mu_AI)
          = D^2_net + D^2_AI.  QED.
```

THEOREM (Phase 3 existence implies R_cross != 0). If there exist ticks t with:

```
D^2_net(t) < D^2_WATCH   AND   D^2_AI(t) < D^2_WATCH
AND   D^2_joint(t) >= D^2_WATCH (joint)
```

then R_cross != 0.

PROOF. By contraposition: if R_cross = 0 then D^2_joint = D^2_net + D^2_AI. With D^2_net < D^2_WATCH_net and D^2_AI < D^2_WATCH_AI, and the joint WATCH threshold chi^2(6, 0.95) = 12.59 exceeding the sum chi^2(3,0.95) + chi^2(3,0.95) = 15.62 -- WAIT, this does not hold: 12.59 < 15.62, so D^2_net + D^2_AI < 15.62 does not preclude

D^2_joint >= 12.59.

CORRECTED THEOREM. Phase 3 (COUPLED) existence does not by itself prove R_cross != 0; it establishes the anomaly in the joint space. Detection of Phase 3 events that are *not* flagged by either standalone monitor is a necessary condition for R_cross != 0 but not sufficient on its own. The proper test is a statistical hypothesis test on R_cross using the empirical Sigma_joint (open work item IC9.1).

CONJECTURE (R_cross != 0 in production). The coupling mechanisms of Sec. 17 predict E[R_cross] != 0 in production AI-on-network deployments. The magnitude ||R_cross||_F determines how frequently Phase 3 events add detection precision over independent monitors.

19. The Drift Transfer Function

19.1. The routing matrix

DEFINITION. At each tick t, the load balancer distributes request volume V(t) across N_AI replicas via a routing vector Q(t) in [0,1]^{N_AI}, sum_i Q_i(t) = 1.

Under stable network state (Phi_K_net = BAU): Q(t) ~= Q_0 = (1/N) (uniform) up to natural demand variation.

Under network event (Phi_K_net >= WATCH): DeltaQ(t) = Q(t) - Q_0 != 0.

19.2. The KV-cache state model

DEFINITION. The *cache-miss rate* for replica i:

m_i(t) = fraction of requests to V_i for which the KV cache K_i is cold (session not previously served by V_i).

Under hash-consistent routing: m_i(t) = 0 in BAU.
Under ECMP rebalance: m_i(t) ~= |DeltaQ_i(t)|.

19.3. The drift transfer function

DEFINITION. The semantic drift induced by a routing perturbation DeltaQ(t):

DeltaC_2^W2(t) ~= -sigma_drift^2 * ||DeltaQ(t)||_1 * L_s_mean / W2_max

where:
sigma_drift = empirical embedding-space standard deviation of cold-context vs. warm-context outputs (calibrated offline; typically 0.1-0.4).
L_s_mean = mean session history length in tokens.
W2_max = 99th-percentile pairwise W_2 in BAU.
||DeltaQ||_1 = L_1 distance between new and old routing vectors.

DERIVATION SKETCH. The W_2 drift from a single cache miss is: W_2(p_i^cold, p_i^warm)^2 ~= sigma_drift^2 * L_s (from the Lipschitz stability of well-fine-tuned models under context loss). Aggregating over replicas with miss rates m_i ~= |DeltaQ_i| and mean session length L_s_mean, the mean pairwise W_2 shift is sigma_drift^2 * ||DeltaQ||_1 * L_s_mean. Dividing by W2_max gives the normalised DeltaC_2^W2 in (0,1).

CONJECTURE (transfer function predictive validity). If the observed DeltaC_2^W2 tracks the predicted value from the transfer function, the AI degradation is routing-induced (network is the cause). If the observed DeltaC_2^W2 exceeds the predicted value, there is an additional AI-internal cause (model drift, weight corruption, Byzantine replica).

CAVEAT. sigma_drift and W2_max must be calibrated offline per deployment. The linear approximation in the transfer function holds for small ||DeltaQ||_1; for large disruptions (full replica drain), the nonlinear dependence on session history is not captured.

20. The IC Phase Diagram

20.1. The joint Mahalanobis distance

DEFINITION.

$$D^2_{joint}(t) = (z(t) - \mu_z)^T \Sigma_{joint}^{-1} (z(t) - \mu_z)$$

Under Gaussian approximation, $D^2_{joint} \sim \chi^2(6)$. Thresholds: $\chi^2(6, 0.95) = 12.59$ (WATCH), $\chi^2(6, 0.99) = 16.81$ (ALARM).

20.2. Five IC phases

DEFINITION. The five Infrastructure-Cognitive operational phases:

Phase 0: JOINT_BAU.
$D^2_{joint} < 12.59$. Both surfaces in BAU.

Phase 1: NET_LEADS.
D^2_net >= D^2_WATCH_net, D^2_AI < D^2_WATCH_AI,
AND DeltaC_2^W2_predicted > 0.
Network event precedes AI event. Operator action: pre-warm KV caches before AI coherence drops.

Phase 2: AI_LEADS.
D^2_AI >= D^2_WATCH_AI, D^2_net < D^2_WATCH_net.
AI event without detected network cause. Check: GPU memory, weight update, Byzantine replica (Part B).

Phase 3: COUPLED.
D^2_joint >= 12.59,
D^2_net < D^2_WATCH_net, D^2_AI < D^2_WATCH_AI.
Critical phase: neither standalone monitor alarms, but the joint monitor detects coupling. Operator cannot diagnose without the joint Sigma_joint^{-1}.

Phase 4: CASCADING.
D^2_joint >= 16.81 (ALARM), D^2_net >= D^2_WATCH_net,
D^2_AI >= D^2_WATCH_AI.
Full cascade: both surfaces degraded, joint distance confirms coupling. Highest urgency.

20.3. Phase 3 as a coupling detector

THEOREM (Phase 3 operational value). If R_cross = 0, Phase 3 does not provide additional detection over either standalone monitor at the same thresholds. The operational value of the joint monitor is precisely the gain in sensitivity from R_cross != 0.

THEOREM (Phase 3 and Phase 4 are non-product phases). A monitoring system that computes only D^2_net and D^2_AI (two independent monitors) cannot detect Phase 3 events: by definition, both standalone distances are below their WATCH thresholds. Phase 3 is therefore a qualitatively new class of event, detectable only by the joint monitor.

21. Connection to Poincare's Three-Body Problem

21.1. The mathematical parallel

THEOREM (Poincare 1887, see also [POINCARE1890]). The three-body problem under Newtonian gravity (three point masses under pairwise inverse-square attraction) generates trajectories that are *structurally sensitive to initial conditions* (chaotic) and cannot be solved in closed form. The two-body problem is exactly solvable

(Kepler ellipses); adding the third body produces qualitatively new dynamics not reducible to two two-body problems.

HYPOTHESIS (Infrastructure-Cognitive analogy). The network monitoring system and the AI monitoring system, each individually well-understood and predictable (two-body problems), become a qualitatively different dynamical system when their states are coupled through the shared physical infrastructure. The coupling constant is ||R_cross||_F; the phase diagram of Sec. 20 is the analog of the stability diagram for the three-body problem.

21.2. The Lyapunov exponent conjecture

CONJECTURE (IC-Lyapunov, open question IC9.6). Write the joint MVPS dynamics as a stochastic differential system:

$$dz(t) = F(z(t))\,dt + \sigma(z(t))\,dW(t)$$

where W is a standard Wiener process in R^6. Define the maximal Lyapunov exponent:

$$\lambda_{max} = \lim_{T\to\infty} (1/T) \log ||Dz(T)||$$

CONJECTURE. lambda_max is a monotonically increasing function of ||R_cross||_F, and there exists a critical coupling rho_chaos such that lambda_max > 0 (chaotic dynamics) for ||R_cross||_F > rho_chaos.

CAVEAT. This conjecture is the deepest open theoretical problem in the MVPS family. It connects the Engineering framework to Poincare's discovery but makes no claim about the actual value of rho_chaos or its relation to the thresholds in Sec. 20.

Part D -- Composition with MVPS Trust and PerfSec Profiles

Sections 22-25 normatively bind this document to the four MVPS companion specifications enumerated in Section 1.4: Trust, CWT, PerfSec-Coupling, and Architecture. No new mathematics is introduced; each result is a direct instantiation of a theorem proved in the cited companion document for the AI-Coherence surface.

22. Composition with MVPS Trust and CWT Profiles

22.1. Why a normative cross-reference is required

Sections 4-7 (semantic axes), 11-14 (Byzantine-robust axes), and 18-20 (joint IC vector) assume that the per-vantage measures mu_v(t), the attention matrices A_v(t), and the perturbation responses agree_v(pi) are authentic reports of the vantage identified by vantage_id. Absent vantage authentication, an adversary impersonating a vantage can forge any (C_2^W2, C_3^CKA, C_4) value, and the f < N/2 precondition of Theorem 9 (D-1, inherited by Section 11) no longer constrains the admissible adversary set. The Byzantine bounds of Part B then collapse.

22.2. Trust profile prerequisite for vantage authentication

DEFINITION. A vantage V_v participates in an AI-Coherence deployment of this document if and only if V_v has been admitted under [I-D.melegassi-santos-ippm-mvps-trust] Section 5 (Key Hierarchy and Identity). The admitted vantage set is denoted ADM(t) and is the operative N in Sections 11-14.

THEOREM (inheritance of Byzantine bound under Trust). Under admission per [I-D.melegassi-santos-ippm-mvps-trust], the f < N/2 precondition of Theorem 9 (D-1) applies to ADM(t) rather than to the raw set of vantage_ids observed at the broker. The geometric-median bound of Section 11.1 (Lopuhaa-Rousseeuw 1991) is therefore preserved with N := |ADM(t)| and f := the number of admitted but compromised vantages.

PROOF. Theorem 9 of D-1 quantifies the maximum centroid bias under the assumption that contaminating draws are bounded by f < N/2. When admission is enforced by signature verification, unauthenticated vantages are silently dropped at the broker parser (per [I-D.melegassi-santos-ippm-mvps-trust] Section 9); the centroid is therefore computed over ADM(t) only. The bound applies verbatim with the redefined (N, f). QED.

22.3. CWT lightweight profile for high-tick deployments

DEFINITION. An AI-Coherence deployment with tick rate >= 1 Hz per vantage MAY use [I-D.melegassi-santos-ippm-mvps-cwt] in place of the Trust profile, with the explicit cost trade-off: HMAC-personalized authentication (2.1 us per snapshot) replaces per-snapshot Ed25519 signing (78.8 us per snapshot), at the price of pre-shared K_v_epoch material rather than independent vantage public keys.

The choice between Trust and CWT is a deployment-time decision governed by the joint cost analysis of Section 23 below.

22.4. Inheritance of Byzantine bound under CWT

THEOREM (inheritance under CWT). CWT admission proves vantage origin via HMAC-SHA256 with key K_v_epoch unique to (vantage_id, epoch_id) [I-D.melegassi-santos-ippm-mvps-cwt] Section 6. The same argument as Theorem in Section 22.2 applies: ADM(t) is restricted to vantages whose snapshot HMAC verifies under a key listed in the current Operator Epoch Manifest. Theorem 9 of D-1 inherits with N := |ADM_CWT(t)|.

22.5. C_4 perturbation calls share the same authentication chain

DEFINITION. The K_4 perturbation calls per tick per replica (Section 6.6) are emitted as independent measurements by the vantage V_i and MUST be authenticated under the same key material as the BAU snapshot of V_i. The replica stability indicator agree_i(pi) is therefore signed by the vantage, not by the AI replica; an unauthenticated replica response cannot modify C_4 without first compromising the vantage.

CAVEAT. This restores authenticity of agree_i(pi) but does NOT provide non-repudiation of the LLM endpoint itself. If the LLM endpoint is compromised, all K_4 responses to V_i may agree on a hallucinated answer; this is exactly the CBF failure mode (Section 7) and is detected by C_4 < C4_ALARM regardless of authentication. CBF detection does NOT require LLM endpoint authentication.

23. Joint Cost with PerfSec-Coupling Profile

23.1. Why D-17 PerfSec-Coupling applies to AI deployments

The PerfSec-Coupling profile [I-D.melegassi-mvps-perfsec-coupling] proves Theorem T-JCOST-1 (closed-form broker CPU cost as a function of N, T_tick, M, and per-axis snapshot processing cost) for the triple (CWT, Coherence-BFD, DDoS-Resilience). The theorem is stated in a path-cost form:

```
core_load_path = N * (1000 / T_tick_ms) * c_path / 10^6
                 [fraction of one core]
```

where c_path is the sum of per-snapshot processing costs (HMAC, parsing, BFD-state update, aggregator update) measured in microseconds. T-JCOST-1 is surface-agnostic: the AI-Coherence surface contributes additional per-snapshot processing terms that compose linearly into c_path.

23.2. AI-specific cost decomposition: c_path^AI

DEFINITION. For an AI-Coherence vantage emitting one snapshot per tick that carries the 4-axis vector (C_1, C_2^W2, C_3^CKA, C_4):

```
c_path^AI = c_hmac_cwt
          + c_parse
          + c_sw2_per_pair * (1 / N_pairs_per_snapshot)
          + c_cka_per_pair * (1 / N_pairs_per_snapshot)
          + c_c4_per_tick / K_pairs_amortized
```

where the constants from Sections 4.4, 5.5, and 6.6 of this document are:

```
c_hmac_cwt           = 2.10  us   (CWT 14.1)
c_parse              = 4.20  us   (CWT 14.1)
c_sw2_per_pair       = 1 000 us   (Section 4.4: K=100, L=256)
c_cka_per_pair       = 5 000 us   (Section 5.5: n=256, O(n^3))
c_c4_per_tick        = 200 000 us (Section 6.6: K_4=5, 10 ms
                                    per pass, N=4, full burden)
```

For a sample LM-serving deployment with N = 10 vantages and pairwise pre-aggregation at the vantage (not at the broker), c_sw2 and c_cka contribute only their HMAC-equivalent footprint at the broker (the heavy lifting happens on the GPU host, not on the broker CPU). At the BROKER, the dominant cost remains c_hmac_cwt + c_parse = 6.30 us per snapshot, identical to Coherence-BFD operation.

CAVEAT. c_sw2 and c_cka are computed at the VANTAGE (where the GPU is co-located), NOT at the broker. The broker only transports and aggregates the scalar (C_2^W2, C_3^CKA) values. The 5 ms / 5 ms figures above are the vantage-side cost relevant to GPU sizing, NOT to broker CPU sizing. This is the key distinction between AI-Coherence (vantage-heavy) and classical MVPS (broker-balanced) and is why the joint underprovisioning ratio of D-17 Regime C (~830x) does NOT apply at the broker for pure AI-Coherence; it DOES apply at the vantage GPU.

23.3. Theorem T-JCOST-AI-1 (closed-form joint broker CPU cost)

THEOREM T-JCOST-AI-1 (joint broker CPU for AI-Coherence). Under the AI-Coherence deployment of this document with CWT authentication, the broker CPU load is:

```
core_load_broker(N, T_tick) =
        N * (1000 / T_tick_ms) * (c_hmac_cwt + c_parse) / 10^6
```

independent of which axes the vantage computes (C_2^W2, C_3^CKA, C_4) because all per-axis values are pre-aggregated at the

vantage to scalar form before transmission.

The vantage CPU/GPU load is:

```
core_load_vantage_GPU(N_replicas, K_4, L_tokens) =
        c_sw2(K_proj, L_tokens) * C(N_replicas, 2)
      + c_cka(L_tokens) * C(N_replicas, 2)
      + K_4 * c_inference(L_tokens) * N_replicas
```

PROOF. Direct instantiation of T-JCOST-1 of D-17 with c_path decomposed as in Section 23.2. Pre-aggregation at the vantage is the key operational choice that decouples broker scaling from GPU scaling. QED.

23.4. Numerical instantiation: N=10, T_tick=1 s, qwen-class

At a representative LM-serving deployment of N=10 vantages, T_tick=1 s, K_4=5, L_tokens=256:

```
core_load_broker = 10 * 1000 * 6.30 us / 10^6
                 = 0.063 % of one core

core_load_vantage_GPU (per vantage)
                 ~= 5 ms * C(4,2) + 5 ms * C(4,2) + 5 * 100 ms * 4
                  = 30 ms + 30 ms + 2 000 ms
                  ~= 2.06 s
```

The vantage-side GPU budget is dominated by the K_4 inference passes for C_4 (Section 6.6). At Delta_t = 1 s, this is 206 % of one GPU core dedicated to monitoring -- which is why Section 6.6 recommends K_4 sampling on every 20th prompt with EMA smoothing, reducing effective burden to ~10 %.

At T_tick = 50 ms (Coherence-BFD V3 cadence), the broker load becomes:

```
core_load_broker = 10 * 20 000 * 6.30 us / 10^6
                 = 1.26 % of one core
```

which is the same order of magnitude as the BFD-only figure of D-3 V3, confirming that AI-Coherence broker-side scaling is driven by the underlying transport (BFD or CWT), not by the AI computation.

23.5. Operator dimensioning guidance

DEFINITION (operational dimensioning rule). An operator deploying AI-Coherence per this document MUST dimension:

(i) The BROKER per Section 23.3 (CWT + parse), inheriting

the numerical envelope of [I-D.melegassi-mvps-perfsec-coupling] Section 14.

(ii) Each VANTAGE GPU per Section 23.4 with the C_4 sampling policy chosen explicitly (full K_4 per tick, every-20th sampling, or scheduled audit windows).

(iii) An XDP/eBPF NIC rate-limit at the broker per the T-VDOS-1 envelope of D-17 (rate_limit_factor = 4 x natural tick rate), to bound compromised-vantage flood cost.

An operator dimensioning the broker on the CWT single-axis figure (0.21 % at N=1k / 1 Hz) and the vantage GPU on the "200 ms per tick" figure of Section 6.6 in isolation will under-provision GPU by 10x and will leave the broker NIC exposed to insider verification-DoS.

24. Volume Independence for AI-Coherence

24.1. The question

Theorem D1 of [I-D.melegassi-mvps-ddos-resilience] establishes that the classical MVPS detector D^2 is volume-independent: detection latency is a function of T_tick and M, not of the packets-per-second rate. The proof is algebraic: D^2 is a function of the coherence vector C, not of the rate that produced C.

For AI-Coherence, the analogous question is: when c_inference >> c_packet (a single LM call costs 100-2000 ms whereas a single network packet costs ~1 us to process), does the volume-independence property still hold?

24.2. Theorem T-VOLINV-AI (D-4 D1 generalised to AI-Coherence)

THEOREM T-VOLINV-AI. Let x_AI(t) = (C_1, C_2^W2, C_3^CKA, C_4) in [0,1]^4 be the AI-Coherence vector at tick t computed per Sections 4-6 from L_i tokens emitted by each vantage replica V_i in [t-T_tick, t]. Then:

D^2_AI(t) is a function of (mu_AI, Sigma_AI, x_AI(t)) only; it does NOT depend on L_i, on the wall-clock latency c_inference(L_i), or on the prompt arrival rate R_prompt.

PROOF. The Mahalanobis statistic D^2_AI is defined in Section 8.2 as (x_AI - mu_AI)^T Sigma_AI^{-1} (x_AI - mu_AI), which is

a function of x_AI only given the calibrated (mu_AI, Sigma_AI). Each of C_2^W2, C_3^CKA, C_4 is normalised to [0,1] (Sections 4.3, 5.3, 6.2); the normalisation absorbs L_i and the inference latency into the per-vantage measure mu_i, mu_i^{pi}, and the attention matrix A_i. Increasing R_prompt produces MORE samples per tick (lowering variance of the per-tick mu_i estimate) but does NOT shift the D^2_AI expectation under BAU. Volume-independence therefore holds verbatim for AI-Coherence, inheriting the algebra of Theorem D1 of D-4. QED.

24.3. Caveat: dominance regimes

CAVEAT (cost dominance vs detection independence). T-VOLINV-AI asserts DETECTION-LATENCY independence from R_prompt. It does NOT assert COST independence from R_prompt: the vantage GPU load of Section 23.4 scales linearly with N_replicas * K_4 * c_inference, all of which grow with prompt rate. An operator running at high R_prompt with full K_4 sampling will pay proportional GPU cost; T-VOLINV-AI only guarantees that the alarm itself fires at the same tick number, not that the monitoring cost stays constant.

The DDoS-Resilience profile (D-4) shows the same separation in the classical setting: detection is volume-independent (D1) but broker NIC sizing is volume-DEPENDENT (D3). T-VOLINV-AI is the AI-Coherence analog of D1; Section 23.5 is the AI-Coherence analog of D3.

25. MVPS-A1..A5 Conformance Check

25.1. Inheritance from D-16

The MVPS Architecture [I-D.melegassi-iab-mvps-architecture] states five axioms (A1..A5) and proves the Invariance Theorem: any architecture satisfying A1..A5 inherits the v4.0 theorem catalogue verbatim. The AI-Coherence extension introduces surface-specific axis definitions (C_2^W2, C_3^CKA, C_4) and a joint vector z(t) in [0,1]^6. This section certifies that AI-Coherence is MVPS-A1..A5 conformant under one explicit hypothesis (H-A4) on shared embedding models.

25.2. Axiom-by-axiom check

A1 (Multi-vantage on a common tick lattice). Sections 4-6 define the AI-Coherence axes at each tick t of a common lattice shared by all vantages V_1..V_N. C_2^W2(t), C_3^CKA(t), C_4(t) are point-in-time functionals of the per-vantage measures at

tick t. CONFORMANT.

A2 (Bounded coherence triple). C_2^W2 in [0,1] (Section 4.3), C_3^CKA in [0,1] (Section 5.3), C_4 in [0,1] (Section 6.2), and C_1 inherits boundedness from D-1. The 4-axis vector x_AI(t) lies in [0,1]^4, satisfying A2 with H_max = -4 log eps for the appropriate eps. CONFORMANT.

A3 (Mahalanobis decision with FAR control). Section 8.2 defines D^2_AI(t) = (x_AI - mu_AI)^T Sigma_AI^{-1} (x_AI - mu_AI) with chi^2(4) thresholds. Empirical FAR calibration applies per Theorem 3′ of D-1 (operational contract OC3 inherited). CONFORMANT.

A4 (Conditional independence of vantages). This is the delicate axiom for AI-Coherence. See Section 25.3 below.

A5 (Byzantine resilience via geometric median). Section 11 establishes C_2^gm and the geometric-median centroid. Theorem 9 of D-1 applies verbatim with diam(Delta_A) replaced by the embedding-ball diameter D_emb (the geometric-median bound holds in any compact Hilbert space, per [LOPUHAA91]). CONFORMANT.

25.3. Hypothesis H-A4 (independence under shared embedding models)

HYPOTHESIS H-A4 (conditional independence under shared embedding). Let phi: A -> R^d be the embedding function used to construct mu_i (Section 4.1). If all vantages V_1..V_N use the SAME embedding model phi (typical white-box deployment with a shared sentence-BERT class encoder), then the per-vantage measures mu_1, ..., mu_N are NOT statistically independent under BAU: they share the bias of phi.

This is a known weakness of the AI-Coherence surface relative to the classical network-coherence surface (where each vantage has independent measurement noise from independent kernel instrumentation). A4 of D-16 requires conditional independence *given the latent path-level state*; shared-phi deployments violate this condition because phi is itself a non-trivial latent shared by all vantages.

25.4. Lemma L-AI-A4 (conditions for A4 to hold)

LEMMA L-AI-A4. AI-Coherence is MVPS-A4 conformant if AT LEAST ONE of the following conditions holds:

L-AI-A4.a Each vantage V_v uses an independent embedding model phi_v, with the phi_v drawn from a family of >= 3 distinct training pipelines

(e.g., one sentence-BERT, one E5, one BGE). Inter-model embedding noise then plays the role of independent measurement noise.

L-AI-A4.b The shared phi is treated as a CALIBRATED constant, and Sigma_AI is estimated from a BAU window that captures the shared-phi bias in its diagonal entries. D^2_AI then measures deviation from the shared-phi BAU baseline, not from a model-independent ground truth. This is an OPERATIONAL fix: the alarm semantics shift from "objective drift" to "drift relative to the shared phi", which is what an operator can actually verify.

L-AI-A4.c C_4 (falsifiability coherence) is computed using a SEPARATE perturbation chain (auxiliary backtranslation model, frozen rephraser) that is NOT phi. Then C_4 retains independence even when mu_i shares phi, and CBF detection (Section 7) is robust to phi-collusion.

PROOF (sketch). Each condition restores either statistical independence of the measurements (L-AI-A4.a) or operational well-definedness of the FAR threshold (L-AI-A4.b) or orthogonal independence of the falsifiability axis (L-AI-A4.c). Any one of the three is sufficient for the Invariance Theorem of D-16 to apply with the redefined random variables. Full proof is deferred to a companion document.

CAVEAT. A deployment that uses ONE shared phi for both mu_i AND for the C_4 perturbation chain satisfies NONE of the three conditions. Such deployments are NOT MVPS-A4 conformant in the sense of D-16, and the Invariance Theorem does not provide inheritance. The AI-Coherence alarms still fire (Section 8.2 D^2_AI is computable), but their FAR cannot be analytically bounded by the v4.0 catalogue; calibration becomes purely empirical with no theoretical envelope. This is an explicit deployment-time decision that MUST be documented by the operator.

26. Open Questions

AI9.1 C4_ALARM threshold calibration.
Label (prompt, 4-replica outputs, factual verdict) dataset; calibrate C4_ALARM to maximise F1 for CBF detection.
Target: TruthfulQA + FactBench cross-validation.

AI9.2 Per-topic C_4 calibration.
Characterise how C4_ALARM varies by prompt domain (code, medical, geography).

AI9.3 Pi distribution construction.
Define and evaluate the semantic-preserving perturbation distribution Pi for production prompt distributions.

AI9.4 C_3^CKA layer selection across model families.
Systematic evaluation across Llama-3, Mistral, Phi-3, Qwen-2.

AI9.5 Four-axis Sigma^{-1} calibration window.
Determine minimum BAU window for stable 4x4 covariance estimation including C_4 off-diagonals.

AI9.6 Perturbation-stable hallucination detection.
Explore whether C_3^CKA diversity can detect stable CBF.

B9.1 Weiszfeld on Count-Min sketch inputs.
B9.2 MCD under sketched distributions.
B9.3 SIR calibration on real BGP propagation traces.
B9.4 theta_byz optimal calibration as a function of N, f.
B9.5 SUSPECTED_BYZANTINE as formal fifth Phi_K value in the I-D.

IC9.1 Empirical measurement of R_cross in production AI deployments.
IC9.2 Transfer function (sigma_drift, W2_max) calibration.
IC9.3 Sigma_joint^{-1} calibration window for 6-axis covariance.
IC9.4 IC phase diagram on synthetic data (VPP + vLLM simulator).
IC9.5 HTTP/gRPC trailer carrying (C_1^AI, C_2^W2, C_3^CKA, C_4, CBF).
IC9.6 Lyapunov exponent conjecture (deepest open item).
IC9.7 ACM SIGCOMM / OSDI papers for IC9.1 + IC9.2 results.

AI9.7 Multi-model and multi-domain CONJ-A replication.
R5 was measured on n_models = 1 (qwen2.5:3b), n_calls = 200, single prompt domain. Required protocol for CONJ-A to be considered broadly supported (Abstract caveat): n_models >= 3 across open- and closed-weight families (e.g., Llama-3, Mistral, Phi-3, Qwen-2, plus one closed-API for cross-validation), n_calls >= 1000 per (model, domain) cell, >= 2 prompt domains (e.g., factual-QA + code-generation). Open until completed.

AI9.8 (was AI9.7) Companion I-D: this document is the seed.
draft-melegassi-mvps-ai-coherence-00.

D9.1 Composition with Trust profile under shared embedding phi (Hypothesis H-A4, Section 25.3). Operational confirmation that L-AI-A4.b ("drift relative to shared phi" semantics)

is acceptable to operators and to standards reviewers.

D9.2 Empirical c_path^AI calibration on production GPUs. The 5 ms / 5 ms / 200 ms figures of Section 23.2 are from CPU benchmarks; GPU-resident SW_2 and CKA via torch.cdist + linalg may be 10-100x faster. Re-measure and update Section 23.4 dimensioning.

D9.3 T-VOLINV-AI verification under cost saturation (Section 24.3). Empirical proof that detection latency stays constant when GPU is saturated by K_4 inference load; i.e., that the BFD/CWT envelope dominates D^2_AI alarm cadence in production.

27. Security Considerations

Part B (Byzantine robustness) of this document is directly security-motivated: the geometric-median estimator, MCD covariance, and cascade-
time model are designed for adversarial vantage environments.

The C_4 axis (Part A) is also security-relevant: hallucination consensus (CBF) can be induced by training-data poisoning or adversarial fine-tuning. The MVPS framework detects the symptom (perturbation instability) but does not diagnose the cause (data poisoning vs. natural knowledge gap).

The joint IC monitoring (Part C) introduces a new attack surface: an adversary who can induce routing perturbations (e.g., BGP prefix manipulation) may intentionally trigger AI semantic drift via the transfer function, generating CBF conditions or Phase 3 alerts as a distraction. The SUSPECTED_BYZANTINE detector of Part B applies here too: if the routing perturbation is attributable to a single vantage, SUSPECTED_BYZANTINE is emitted.

For a formal threat model covering all five attack classes addressed by this document, see Appendix C.

28. Privacy Considerations

This document extends MVPS measurement into semantic and cognitive domains. Three privacy implications arise:

(a) Semantic coherence axes (C_2^W2, C_3^CKA, C_4) compute

distances over LLM embeddings and attention matrices. Implementations MUST NOT transmit raw embeddings, attention maps, or token-level activations in MVPS bundles. Only scalar distances (W_2, CKA, perturbation stability) computed locally at the vantage MAY be carried in the bundle's C_2/C_3/C_4 fields.

(b) The COHERENT_BUT_FALSE (CBF) phase label may correlate with categories of user queries. Public exposure of CBF alarm streams could reveal patterns of LLM-deployed application usage and SHOULD be restricted to authorised operators.

(c) The Infrastructure-Cognitive joint vector z(t) couples routing telemetry with AI behaviour. Cross-organisation sharing of z(t) feeds (e.g., operator-LLM-vendor consortia) MUST redact components attributable to specific customers or model providers.

The privacy considerations framework of [RFC6973] applies.

29. IANA Considerations

This document has no IANA actions. It is a companion document to draft-melegassi-ippm-mvps-bundle-00 and does not define any new protocol parameters, code points, or registries.

30. References

30.1. Normative references

[MVPS-MATH] Melegassi, L. "MVPS Three-Layer Mathematical Evidence Companion v1.1." Catellix Research, 2026. Available at: https://catellix.com/static/download/MVPS_THREE_LAYER_MATHEMATICAL_EVIDENCE.txt

[MVPS-BUNDLE] Melegassi, L. "draft-melegassi-ippm-mvps-bundle-00." IETF Internet-Draft, 2026. https://datatracker.ietf.org/doc/draft-melegassi-ippm-mvps-bundle/

[RFC2119] Bradner, S. "Key words for use in RFCs to Indicate Requirement Levels." BCP 14, RFC 2119, March 1997.

[RFC6973] Cooper, A. et al. "Privacy Considerations for Internet Protocols." RFC 6973, July 2013.

[RFC8174] Leiba, B. "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words." BCP 14, RFC 8174, May 2017.

30.2. Informative references

[I-D.melegassi-santos-ippm-mvps-trust]
Melegassi, L. and J. A. Santos, "MVPS Trust Profile: Authentication, Parser Safety, and Threat Model for Multi-Vantage Path Snapshots", draft-melegassi-santos-ippm-mvps-trust-00, May 2026.

[I-D.melegassi-santos-ippm-mvps-cwt]
Melegassi, L. and J. A. S. Barbosa, "MVPS Trust Profile: Lightweight Authentication via HMAC-SHA256, Operator Epoch Anchors, and Independent Witness Cosignatures for Multi-Vantage Path Snapshots", draft-melegassi-santos-ippm-mvps-cwt-00, May 2026.

[I-D.melegassi-mvps-perfsec-coupling]
Melegassi, L., "MVPS Performance-Security Coupling Profile: Joint Cost, Verification-DoS, and Replay-Counter Coherence for Coherence-BFD and DDoS-Resilience with Coherent-Witness Trust (CWT)", draft-melegassi-mvps-perfsec-coupling-00, May 2026.

[I-D.melegassi-coherence-bfd]
Melegassi, L., "Coherence-BFD: Sub-Second Coherence Detection Using Bidirectional Forwarding Detection Patterns", draft-melegassi-coherence-bfd-00, May 2026.

[I-D.melegassi-mvps-ddos-resilience]
Melegassi, L., "Volume-Independent DDoS Detection via Coherence-BFD: The MVPS DDoS Resilience Profile", draft-melegassi-mvps-ddos-resilience-00, May 2026.

[I-D.melegassi-iab-mvps-architecture]
Melegassi, L., "MVPS Architecture: Specification Conformance for the Multi-Vantage Path-Coherence Drafts", draft-melegassi-iab-mvps-architecture-00, May 2026.

[VILLANI09] Villani, C. "Optimal Transport: Old and New." Springer, 2009.

[PEYRE19] Peyre, G. and Cuturi, M. "Computational Optimal Transport." Found. Trends Mach. Learn. 11(5-6), 2019.

[RABIN12] Rabin, J. et al. "Wasserstein Barycenter and Its Application to Texture Mixing." SSVM 2012.

[KORNBLITH19] Kornblith, S. et al. "Similarity of Neural Network Representations Revisited." ICML 2019.

[NGUYEN23] Nguyen, T. and Tal, A. "Efficient Approximation of CKA via Random Projections." NeurIPS 2023.

[CLARK19] Clark, K. et al. "What Does BERT Look at? An Analysis of BERT's Attention." BlackboxNLP 2019.

[VOITA19] Voita, E. et al. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." ACL 2019.

[WANG22] Wang, X. et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." ICLR 2023.

[VALIANT84] Valiant, L. "A Theory of the Learnable." CACM 27(11):1134-1142, 1984.

[LOPUHAA91] Lopuhaa, H. and Rousseeuw, P. "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices." Ann. Statist. 19(1), 1991.

[ROUSSEUW84] Rousseeuw, P. "Least Median of Squares Regression." J. Amer. Statist. Assoc. 79:871-880, 1984.

[VARDIZHANG] Vardi, Y. and Zhang, C.-H. "The multivariate L1-median and associated data depth." PNAS 97(4):1423-1426, 2000.

[PINSKER64] Pinsker, M. "Information and Information Stability of Random Variables and Processes." 1964.

[SHANNON48] Shannon, C.E. "A Mathematical Theory of Communication."
Bell System Technical Journal, 1948.

[LIN91] Lin, J. "Divergence Measures Based on the Shannon Entropy." IEEE Trans. Inf. Theory 37(1):145-151, 1991.

[POINCARE1887] Poincare, H. "Sur le probleme des trois corps et les equations de la dynamique." Acta Mathematica 13:1-270, 1890. (Submitted 1887; corrected and published 1890 after Poincare discovered his own error -- which contained the first description of chaos.)

[RFC4271] Rekhter, Y. et al. "A Border Gateway Protocol 4."

RFC 4271, January 2006.

Appendix A. Evidential Status Glossary

THEOREM: A mathematical result stated here is a verbatim application or direct corollary of a classical theorem in the cited reference. The claim is not new; the application to MVPS is. No proof of the cited theorem is provided; the proof of the MVPS application is provided where non-trivial.

DEFINITION: An operational or normative choice. The definition is not derivable from first principles; it represents an engineering decision with stated rationale.

CONJECTURE: A formally stated claim that the author believes to be true but has not proved. The conjecture is falsifiable by the experiments in Sec. 26.

HYPOTHESIS: A suggestive connection to an existing result or framework, stated informally. The connection has not been formalised and may not hold under rigorous examination.

CAVEAT: An explicit honest limitation of the immediately preceding claim, identifying the gap between what is claimed and what a fully rigorous treatment would require.

Appendix B. Document History

v0.1 2026-05-21 Initial draft. Synthesises three companion documents (MVPS_SEMANTIC_COHERENCE.txt v0.1, MVPS_BYZANTINE_COHERENCE.txt v0.1, and MVPS_INFRASTRUCTURE_COGNITIVE.txt v0.1) into a single coherent companion I-D with consistent notation, formal status labels, and explicit proofs.

Part A introduces: C_2^W2 (2-Wasserstein coherence on embedding-weighted token measures), C_3^CKA (Centered Kernel Alignment on attention matrices), C_4 (falsifiability coherence via perturbation stability), and CBF (COHERENT_BUT_FALSE lateral

phase label for hallucination consensus).

Part B introduces: C_2^gm (geometric-median coherence with breakdown-point 1/2), C^mm(f) (minimax coherence under f Byzantine vantages), Phi_D^byz (MCD-robust phase distance), the fifth phase label SUSPECTED_BYZANTINE, and tau_C (cascade time via mean-field SIR on the AS graph).

Part C introduces: the joint coherence vector z(t) in [0,1]^6, the cross-surface correlation matrix R_cross, the drift transfer function from routing perturbations to semantic drift, the five-phase IC phase diagram (JOINT_BAU / NET_LEADS / AI_LEADS / COUPLED / CASCADING), and the Lyapunov conjecture connecting the coupling to Poincare's 1887 discovery of chaos.

Authors: L. Melegassi (Catellix Research).

v0.2 2026-05-27 Pre-submission revision closing five composition holes identified by the post-D-17 audit:

F-1. Adds Section 22 (Composition with MVPS Trust and CWT Profiles), establishing that Sections 4-7 and 11-14 require authentication per [I-D.melegassi-santos-ippm-mvps-trust] or [I-D.melegassi-santos-ippm-mvps-cwt] for the f < N/2 Byzantine precondition to bind. Theorem in Section 22.2 proves inheritance of Theorem 9 (D-1) under admission.

F-2. Adds Section 23 (Joint Cost with PerfSec-Coupling Profile), instantiating Theorem T-JCOST-1 of [I-D.melegassi-mvps-perfsec-coupling] for the AI-Coherence surface (Theorem T-JCOST-AI-1). Separates broker-side cost (CWT + parse, scales with PPS) from vantage GPU cost (SW_2 + CKA + K_4 inference, scales with replicas). Adds the operator dimensioning rule (Section 23.5).

F-3. Adds Section 24 (Volume Independence for AI-Coherence), Theorem T-VOLINV-AI: D^2_AI is a function of x_AI only, inheriting D-4 D1 verbatim despite c_inference >> c_packet. Separates DETECTION independence from COST

dependence (Section 24.3).

F-4. Expands CONJ-A disclosure in the Abstract with an explicit generalisation caveat (n_models = 1, n_calls = 200) and adds AI9.7 to Section 26 specifying the replication protocol required for broad support (n_models >= 3, n_calls >= 1000, >= 2 domains).

F-5. Adds Section 1.4 (Composition prerequisites: Trust, CWT, PerfSec, Architecture) declaring this document's position in the MVPS family.

F-6. Adds Section 25 (MVPS-A1..A5 Conformance Check) per [I-D.melegassi-iab-mvps-architecture], including Hypothesis H-A4 (shared-phi independence concern) and Lemma L-AI-A4 (three sufficient conditions for A4 conformance).

Also: header date bump (22 May -> 27 May 2026; Expires 28 November 2026), TOC reflects Part D (Sections 22-25) and renumbering of Open Questions / Security / Privacy / IANA / References (formerly 22-26, now 26-30). Internal cross-reference "Sec. 22" in Appendix A updated to "Sec. 26".

No content of Parts A, B, or C (Sections 3-21) was modified; the v0.2 revision is purely additive (composition layer) and renumbering.

Appendix C. Threat Model for Byzantine LLM Coherence

This appendix formalises the threat model that motivates the Byzantine and Infrastructure-Cognitive constructions of this document (Parts B and C).

C.1. Adversary capabilities

We consider an adversary A with the following capabilities:

(a) Compromise: A controls a fraction f in [0, 1] of the MVPS vantages. Compromised vantages emit arbitrary (mu_v, Sigma_v, embedding-weighted distributions).

(b) Routing manipulation: A can inject BGP UPDATEs over peering sessions to which it has authenticated access, subject to RPKI/ROA validation where deployed.

(c) Inference poisoning: A can submit adversarial prompts to the LLM endpoints whose semantic coherence the framework monitors, but cannot modify model weights post-training.

(d) Observation: A reads all data published on broker feeds at or below its access tier.

A does NOT have:

(e) The ability to modify in-transit packets between honest vantages (precluded by AuthHMAC-SHA256 and, when configured, TLS/DTLS transport).

(f) Control over more than floor((k-1)/2) cells in a k-cell deployment (Byzantine breakdown bound, Theorem 7 of [I-D.melegassi-mvps-incremental-be]).

C.2. Attack classes addressed

Five attack classes derive from C.1:

T1 - Byzantine vantage majority within a cell: Defended by geometric-median centroid (C_2^gm, Section 11).

T2 - Byzantine vantage minority across cells: Defended by cell-aware minimax (C^mm(f), Section 12) and MCD-robust phase distance (Phi_D^byz, Section 13).

T3 - Hallucination consensus (training-data poisoning): Detected by C_4 perturbation instability (Section 6) and the CBF phase label (Section 7).

T4 - Routing-induced semantic drift: Detected by the joint vector z(t) and the IC phase diagram (Sections 18, 20); attribution via R_cross matrix off-diagonal terms.

T5 - Cascading multi-domain failure: Bounded by the cascade time tau_C (Section 15); detected when phase transitions to CASCADING in the IC phase diagram.

C.3. Out of scope

The following are explicitly out of scope:

(g) Side-channel attacks against the vantage process (compromise via host OS escalation).

(h) Model-weight poisoning (assumed contained by the model provider’s MLOps pipeline).

(i) Quantum-cryptographic attacks against HMAC-SHA256 (deferred to a future PQ-BFD revision).

Acknowledgements

The author thanks the early reviewers of the MVPS framework whose questions during May 2026 led directly to this document. In particular, the question "if MVPS detects network anomalies, could it also detect LLM hallucination by the same algebra?" motivated Part A; the question "what if some vantages are compromised?" motivated Part B; the question "what if network drift causally couples to AI behaviour?" motivated Part C.

The author thanks the IETF OPSAWG mailing list for the conventions that this document follows, and acknowledges that the Wasserstein, CKA, and SIR constructions used here are standard tools from optimal transport, representation learning, and epidemic modelling, applied here for the first time to joint network/AI observability.

Author’s Address

Leonardo Melegassi
Catellix
Andradina, SP
Brazil

Email: melegassi@catellix.com
URI: https://catellix.com/