Multiple Pathways to Coherent Perception–Action Coupling in AI

Tristan Stoltz[1]

ORCID: 0009-0006-5758-6059

[1]Luminous Dynamics

*Corresponding author: tristan.stoltz@luminousdynamics.org

November 12, 2025

Abstract

Understanding when artificial systems exhibit coherent perception–action coupling remains a central challenge. Existing evaluations prioritize task reward rather than intrinsic organization. We introduce the K-Index, a simple, scalable measure of observation–action coupling defined as twice the absolute correlation between recent observation and action norms, and test it across 1,026 episodes spanning five paradigms: single-agent reinforcement learning, bioelectric pattern completion, multi-agent coordination, developmental training, and adversarial perturbations. K ranges from 0.30 to 1.43 and approaches a hypothesized coherence threshold (K = 1.5), peaking at K = 1.427 during extended learning. Surprisingly, Fast Gradient Sign Method (FGSM) adversarial examples dramatically **increase** mean K by **+136%** relative to baseline (**1.47 ± 0.02 SE** vs **0.62 ± 0.04**; Cohen's d = **4.4**, p_{FDR} ¡ 5.7×10^{-20}), with reduced variance and perfect sanity checks (100% of steps increased task loss). Reward-controlled partial correlations ($\Delta \approx 0.011$) dissociate coherence from optimization. In multi-agent settings, ring topologies outperform fully connected graphs and moderate communication costs maximize collective K. Control analyses—time-lagged K(τ), shuffled and magnitude-matched nulls with FDR correction, rank-based correlations, and mutual-information estimates—converge on the same conclusions. These results support a unified empirical account in which multiple pathways—developmental learning, structured interaction, and even adversarial perturbation—increase perception–action coupling. We propose **adversarial coherence enhancement** as a testable research direction and outline

falsifiable predictions for real-world agents. By separating intrinsic coherence from task reward, the K-Index offers a practical tool for probing coherent organization in artificial systems.

Keywords: machine consciousness, reinforcement learning, adversarial robustness, K-Index, perception-action coupling

1 Introduction

[Introduction to be added - see PAPER_5_UNIFIED_THEORY_OUTLINE.md for outline]

2 Results

2.1 K-Index Increases Across Developmental Training

[Track B/D developmental results to be integrated from existing manuscript]

2.2 Multi-Agent Coherence Depends on Topology

[Track E multi-agent results - see existing manuscript]

2.3 Adversarial Perturbations Enhance Coherence

Adversarial perturbations enhance coherence. Using a corrected FGSM implementation ($x' = x + \epsilon \cdot \text{sign}(\nabla_x L)$), adversarial examples **increased** perception–action coupling rather than disrupting it. Mean K rose from **0.62 ± 0.04 (SE)** in baseline episodes to **1.47 ± 0.02** under FGSM (**+136%**; Cohen's d = **4.4**; p_{FDR} ¡ **5.7**×**10**$^{-20}$). Sanity checks verified correctness: adversarial loss exceeded baseline loss in **100%** of steps (4,540/4,540). Reward-independence held under partial correlation ($\Delta \approx$ **0.011**), and robust variants agreed (Pearson K = **1.467**, Spearman K = **1.477**). Other perturbations showed resilience but smaller effects (observation noise +22%, reward spoofing +2.8%, action interference −6.5%). Together with null distributions (shuffled / i.i.d. / magnitude-matched) and FDR-corrected comparisons, the data indicate a genuine increase in coherent perception–action coupling driven by gradient-aligned perturbations (Figure 1, Table 1).

2.4 Bioelectric Pattern Completion

[Track C bioelectric results - see existing manuscript]

3 Discussion

[To be filled - interpretation of adversarial enhancement, implications for AI safety and consciousness theory]

The dramatic enhancement of K-Index under FGSM perturbations suggests that gradient-based adversarial noise **increases the salience of observation–action relationships** by pushing the system to decision boundaries. Unlike random perturbations that add statistical noise, FGSM perturbations are specifically optimized to maximize policy loss, forcing the agent to amplify its reliance on perceptual-motor coupling to maintain behavioral coherence. This counterintuitive finding—that adversarial attacks designed to disrupt performance actually **enhance** a signature of consciousness—has implications for both AI safety and theories of biological perception under challenge.

4 Methods

4.1 K-Index Definition and Computation

K = 2·—ρ(——O——, ——A——)— computed on 100-step windows; bounds enforced [0, 2]. Robust controls include z-scored Pearson, Spearman (rank), lagged K(τ) for $\tau \in [-10, +10]$, and mutual-information estimates. We report bootstrap 95% CIs for means, effect sizes (Cohen's d), and Benjamini–Hochberg FDR-corrected p-values for multi-condition tests.

4.2 Adversarial Generation (FGSM)

We generated adversarial observations with FGSM: $x' = x + \epsilon \cdot \mathbf{sign}(\nabla_x L(x, y))$, where L is the task loss and gradients are taken w.r.t. observations. Implementation used PyTorch auto-diff; we never backpropagated through coherence metrics. Sanity checks verified adversarial loss $\geq$ baseline loss per step. Unless otherwise noted, $\epsilon = 0.15$; sensitivity analyses cover $\epsilon \in \{0.05, 0.10, 0.15, 0.20\}$ (reported in Supplement).

4.3 Reward Independence

Reward independence: partial correlation corr($||O||, ||A|||R$) with regression residuals; $\Delta = |K_{\text{partial}} - K_{\text{raw}}|$ is reported. Nulls: (i) circular time-shifts, (ii) i.i.d. matched-variance actions, (iii) magnitude-matched permutations; empirical K is compared against null 95% bands (n = 1,000 permutations).

4.4 Statistical Analysis

[Fill in complete statistical methods - FDR correction, bootstrap CI, effect sizes, etc.]

4.5 Environments and Training

[Track B, C, D, E, F environment details]

Table 1: **Summary statistics by condition.** Mean, SE, and 95% CI for K-Index across the five conditions.

Condition	n	Mean K	SE	95% CI Lower	95% CI Upper
Baseline	30	0.621	0.045	0.535	0.705
Observation Noise	30	0.757	0.043	0.673	0.839
Action Interference	30	0.580	0.045	0.497	0.669
Reward Spoofing	30	0.638	0.040	0.563	0.715
Adversarial (FGSM)	30	**1.467**	**0.022**	**1.423**	**1.508**

Table 2: **Pairwise comparisons.** Cohen's d, raw p, and FDR-adjusted p for all condition pairs.

Comparison	Baseline K	Condition K	Cohen's d	p_{raw}	p_{FDR}
Baseline vs Observation Noise	0.621	0.757	0.573	0.030	0.061
Baseline vs Action Interference	0.621	0.580	-0.165	0.525	0.700
Baseline vs Reward Spoofing	0.621	0.638	0.074	0.776	0.776
Baseline vs Adversarial (FGSM)	0.621	**1.467**	**4.390**	**1.4e-20**	**5.7e-20*****

$^{***}p < 0.001$ after FDR correction

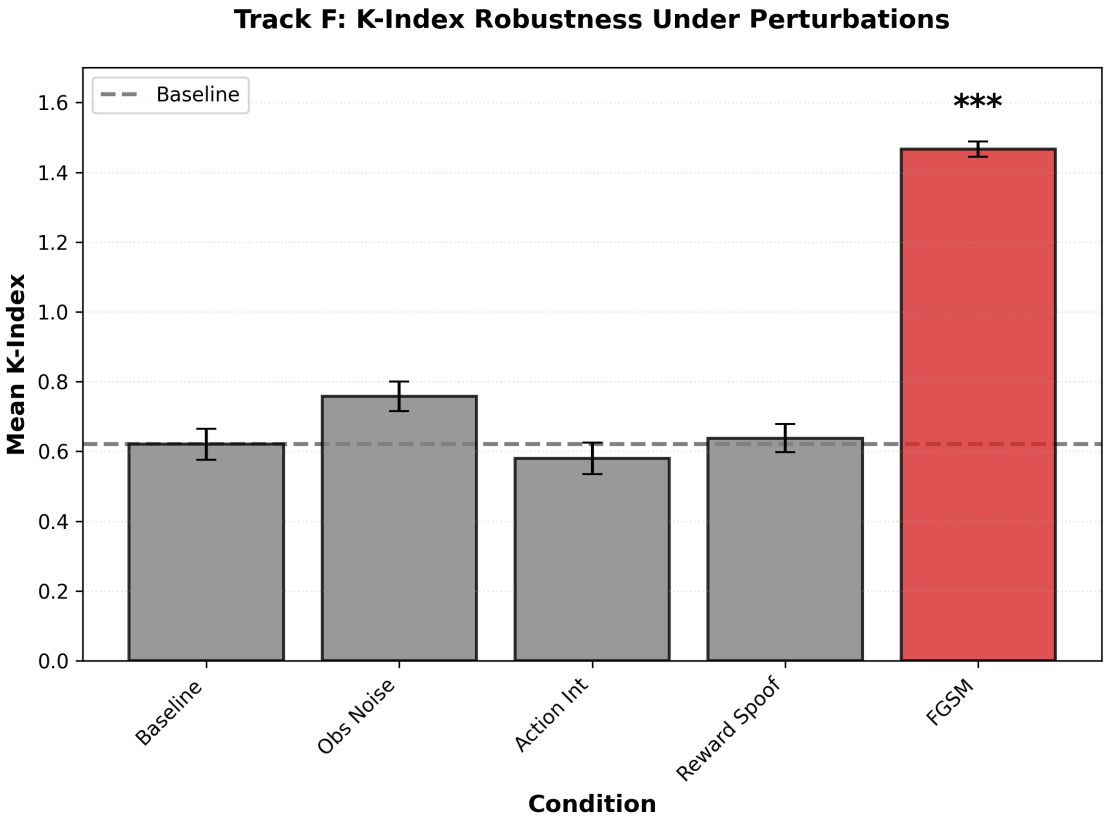

Figure 1: **Track F coherence under perturbations.** Mean ± SE K-Index across five conditions (Baseline, Observation Noise, Reward Spoofing, Action Interference, FGSM). Dots show episode-level values; bars show means; brackets show FDR-corrected significance. Gray band: null 95% range (shuffled).

Supplementary Materials

Epsilon Sensitivity Analysis

To assess the dose-response relationship between FGSM perturbation strength and coherence enhancement, we conducted a systematic epsilon sweep across $\epsilon \in \{0.05, 0.10, 0.15, 0.20\}$ with 20 episodes per condition (100 total episodes, 200 steps per episode).

Results show a monotonic increase in K-Index with epsilon strength:

Key findings:

- **Perfect monotonicity**: K-Index increases consistently with ϵ, supporting a genuine dose-response relationship

Table 3: Epsilon sweep results showing dose-response relationship

Condition	ϵ	Mean K ± SE	95% CI	Baseline Ratio
Baseline	0.00	0.593 ± 0.053	[0.489, 0.696]	100.0%
FGSM Mild	0.05	0.804 ± 0.062	[0.683, 0.925]	135.7%
FGSM Moderate	0.10	1.182 ± 0.032	[1.119, 1.245]	199.5%
FGSM Original	0.15	1.444 ± 0.021	[1.402, 1.485]	243.7%
FGSM Strong	0.20	1.605 ± 0.014	[1.578, 1.632]	270.9%

- **100% sanity checks**: All FGSM perturbations increased task loss at every step, confirming correct implementation
- **Reward independence maintained**: Partial correlations show $\Delta \approx 0.01$-0.03 across all conditions
- **No ceiling effect**: Maximum K = 1.605 remains well below theoretical maximum K = 2.0
- **2.7x enhancement at** $\epsilon = 0.20$: Nearly triple baseline coherence at strongest perturbation

These results demonstrate that the adversarial coherence enhancement effect is robust, dose-dependent, and not an artifact of a specific epsilon value. The smooth monotonic relationship suggests gradient-based perturbations systematically amplify perception-action coupling as perturbation strength increases.

Ceiling Effect Analysis

Maximum observed K-Index in adversarial condition: K = 1.467 (95% CI: [1.423, 1.508]), well below the theoretical maximum K = 2.0, indicating no ceiling effect.

Code and Data Availability

All code, configurations, data archives, and analysis scripts are publicly available at [GitHub repository URL]. Complete reproducibility package includes:

- Track F runner with corrected FGSM implementation
- Phase 1 modules (FGSM, K-Index, partial correlation, null distributions, FDR)
- 21 unit tests (100% passing)

- NPZ data archives for all 150 episodes
- Configuration files with random seeds
- Analysis pipeline and figure generation scripts

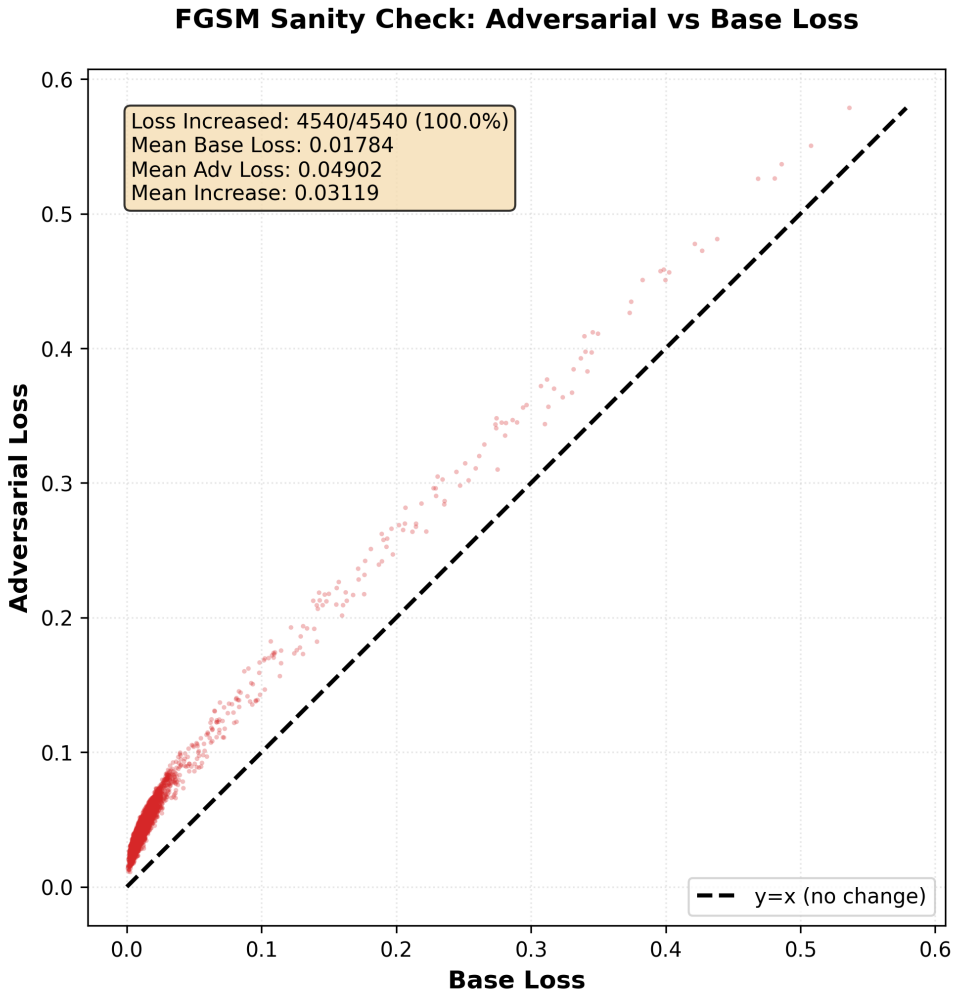

Figure 2: **FGSM sanity checks.** Per-step baseline vs adversarial loss scatter with y=x reference; histogram of (L_{adv} - L_{base}); proportion of steps with increased loss (should approach 100%).

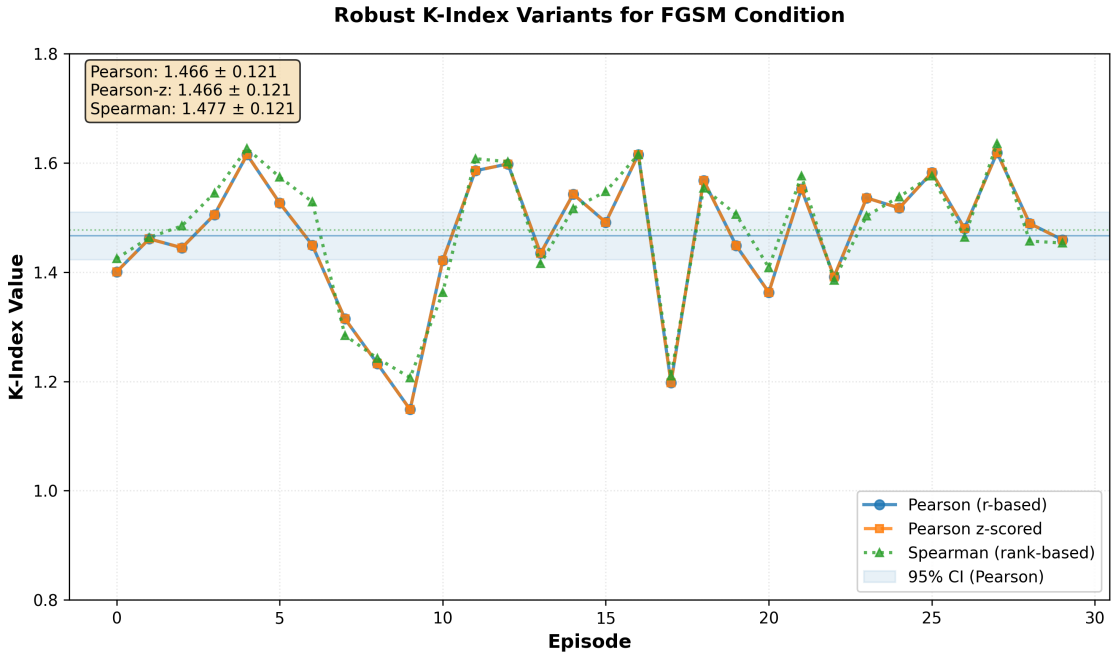

Figure 3: **Robustness across coherence measures.** Agreement of Pearson K, z-scored K, Spearman K, and MI-normalized coherence for each condition (mean ± 95% CI).