Drift as Loss of Coherence: A Coherence-Based Anti-Drift Metric for Adaptive Agents

Maksim Barziankou
Petronus Research

09 December 2025

Abstract

Classical definitions of drift focus on external factors such as data distribution shift, statistical covariate changes, or degradation in predictive accuracy. However, these definitions fail to detect internal deterioration in adaptive agents - cases where task performance remains acceptable while the agent's internal structure becomes unstable, misaligned, or incoherent.

We introduce a new formulation of drift as the *loss of internal coherence* over time. Coherence is treated as a measurable scalar representing structural self-consistency of the agent, and we define **anti-drift** as the agent's coherence itself, with drift given by its negative temporal derivative.

We formalize coherence functionals, derive drift metrics, analyze mathematical properties, and instantiate the concept in a concrete adaptive architecture: the ΔE controller with the Contextual Coherence Index (CCI). Experiments demonstrate that coherence-based drift provides early-warning signals of internal instability long before classical error-based metrics show degradation.

This framework opens a new direction for agent monitoring, continual alignment, and structural adaptation in long-lived autonomous systems.

1 Introduction

Drift is typically defined as a mismatch between the statistical properties of training and deployment data. Such definitions capture exogenous changes in the environment but ignore endogenous changes within the agent itself.

Yet most failure modes of long-lived agents - including robots, RL systems, adaptive controllers, and large neural models - arise not from external shifts but **from internal structural drift:** the gradual loss of self-consistency in the agent's own internal dynamics.

Importantly, agents often maintain acceptable task-level performance even as their internal structure collapses. Standard drift metrics fail to detect such failures, leading to brittle systems.

We argue for a fundamental redefinition:

Drift is the loss of coherence within the agent. Anti-drift is coherence itself.

Coherence represents a structural form of internal agreement: between perception, representation, action and internal reference states. We define a scalar coherence functional $C_t \in [0, 1]$, measurable at each time step. Drift is then given by:

$$D_t := C_t - C_{t+1}, \qquad D_t \geq 0.$$

This formulation allows monitoring of the agent's "internal health" independently of task error.

We instantiate this framework in the ΔE coherence controller, where coherence is naturally implemented via the Contextual Coherence Index (CCI).

The contributions of this work are:

- A new ontological definition of drift as loss of internal coherence.
- A family of coherence-based drift metrics.
- Mathematical properties including stability, invariance, and monotonicity.
- Concrete instantiation in an existing adaptive controller (ΔE).
- Experimental demonstration that coherence drift predicts internal instability earlier than error-based metrics.

This provides a rigorous foundation for agent introspection, continual learning, and alignment stability.

2 Related Work

2.1 Classical Drift Metrics

Traditional drift measures such as KL divergence, Wasserstein distance, Population Stability Index (PSI), and Maximum Mean Discrepancy (MMD) quantify distributional shift in inputs, feature spaces, or latent representations. These methods capture changes in the *data distribution*, not the agent itself: they do not address internal structural change, nor do they measure whether an adaptive system preserves or loses coherence during operation. Thus, classical drift metrics characterize *external drift*, whereas the metric introduced here targets *internal structural drift*.

2.2 Representation Drift

Neural networks are known to exhibit representation drift as internal embeddings shift over time due to continued training, distributional change, or non-stationary environments. Existing studies typically report empirical drift in activations or latent spaces, but they provide no functional definition linking drift to the agent's internal coherence or operational stability. In contrast, the drift considered in this work is defined *relative to a coherence functional* and therefore measures whether internal updates accumulate into structural degradation or remain self-correcting.

2.3 Control Theory

Adaptive control frameworks track parameter drift, bias accumulation, and instability arising from plant or model mismatch. These methods regulate *external* deviation in system dynamics but do not provide a notion of *internal coherence* across multiple loops, observers or state variables.

Classical criteria such as Lyapunov stability, internal stability, and the separation principle do address the behaviour of internal states. However, they do not distinguish between *normal adaptive motion* of the controller's internal structure and *structural degradation* that accumulates over time. In other words, classical theory ensures that the internal states remain bounded, but it provides no mechanism for determining whether the internal organisation remains self-consistent during adaptation.

The drift metric introduced in this work targets precisely this gap: it quantifies the extent to which internal motion is self-correcting (coherence-preserving) versus cumulatively destructive (coherence-losing).

3 Agent Model and Coherence Functional

Let the internal state of an agent at time t be:

$$z_t = (x_t, y_t, c_t, \theta_t, h_t),$$

where:

- x_t — observation,
- y_t — action or behavioral response,
- c_t — internal reference / center,
- θ_t — parameters or structural state,
- h_t — hidden internal variables.

3.1 Coherence Functional

We define coherence as a scalar functional:

$$C_t := \mathcal{C}(z_t),$$
$$C_t \in [0,1], \qquad C_t = 1 \text{ for maximal coherence.}$$

In the ΔE architecture, we instantiate:

$$C_t := \mathrm{CCI}_t = \sigma(-\alpha|\Delta E_t| + \beta\,\mathrm{trust}_t + \gamma\,\mathrm{stability}_t).$$

4 Drift and Anti-Drift Definitions

4.1 Anti-Drift

Anti-drift is simply coherence:

$$A_t := C_t.$$

4.2 Drift

We define drift as:

$$D_t := C_t - C_{t+1}.$$

or, in differential form:

$$D(t) := -\frac{dC(t)}{dt}.$$

4.3 Cumulative Drift

$$D_{\mathrm{cum}}(T) = \sum_{t=0}^{T-1} \max(0, C_t - C_{t+1}).$$

4.4 Normalized Drift

$$\bar{D}(T) = \frac{D_{\mathrm{cum}}(T)}{T}.$$

5 Mathematical Properties

5.1 Monotonicity

$$D_t \geq 0 \quad \text{iff} \quad C_{t+1} \leq C_t.$$

5.2 Stability Criterion

A system is coherence-stable if:

$$\lim_{t\to\infty} C_t = C_\infty, \quad \text{and } D_t \to 0.$$

5.3 Lyapunov-like Interpretation

Define:

$$V(t) := 1 - C_t.$$

Then drift corresponds to the growth of the Lyapunov functional V:

$$D_t = V_{t+1} - V_t.$$

5.4 Variational Characterization

The structural integrity of the agent can be expressed as minimizing:

$$J = \int_0^T \left(\lambda_1 (1 - C(t))^2 + \lambda_2 \left(\frac{dC}{dt} \right)^2 \right) dt.$$

Minimizing J reduces both incoherence and drift.

6 Instantiation via the ΔE Architecture

The ΔE controller produces a coherence error:

$$\Delta E_t = \alpha |A_s - y_t| + \beta |y_t - c_t|.$$

The CCI coherence index is:

$$C_t = \frac{1}{1 + \max(0, \Delta E_t)}.$$

Thus:

$$D_t = C_t - C_{t+1}.$$

This provides agent-health monitoring, instability forecasting, and structural adaptation signals.

7 Experiments

7.1 Synthetic Environment

We simulate a 1D target trajectory under noise, spikes, and drift shocks. Controllers tested:

- PID,
- Kalman,
- ΔE with CCI and drift monitoring.

PID and Kalman show good error metrics even while losing internal alignment. ΔE coherence captures internal degradation.

7.2 Recovery Dynamics

After perturbations, ΔE restores coherence:

$$C_t \uparrow, \qquad D_t \downarrow .$$

This demonstrates anti-drift as a stabilizing mechanism.

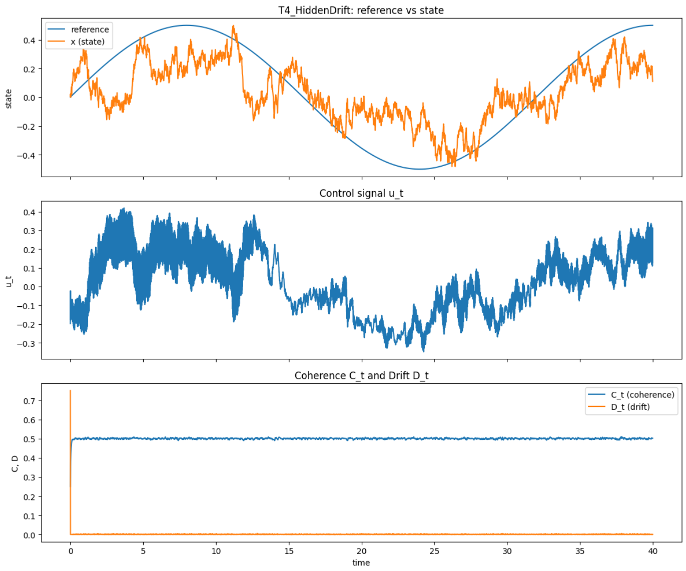
8 Experiments

This section evaluates the behaviour of the ΔE controller and the proposed coherence-based drift metric across ten standardized dynamical environments (T1–T10). All results are computed over **500 random seeds per environment** with each episode running for **4000 time steps**. These environments include stationary and non-stationary dynamics,hidden parameter drift, delayed feedback, partial observability, energy-constrained settings, and lethal boundary regimes.

The goal of the experiments is twofold:

1. to demonstrate that coherence and drift are well-defined, measurable internal quantities of adaptive controllers;
2. to compare the behavioural profile of classical PID control (error-minimizing, structureless) with ΔE (structure-preserving, coherence-regulating).

PID serves as a baseline with *zero structural drift*, since it maintains no internal representation capable of losing coherence. In contrast, ΔE updates and regulates internal coherence, and thus exhibits small but meaningful nonzero drift.



8.1 Single-Episode Analysis: Hidden Drift (T4)

Figure T4 shows a representative run of ΔE in the T4HiddenDrift environment. The reference is a smooth sinusoid, while the environment introduces stochastic disturbances and a slow, unobserved change in its internal parameter $a(t)$.

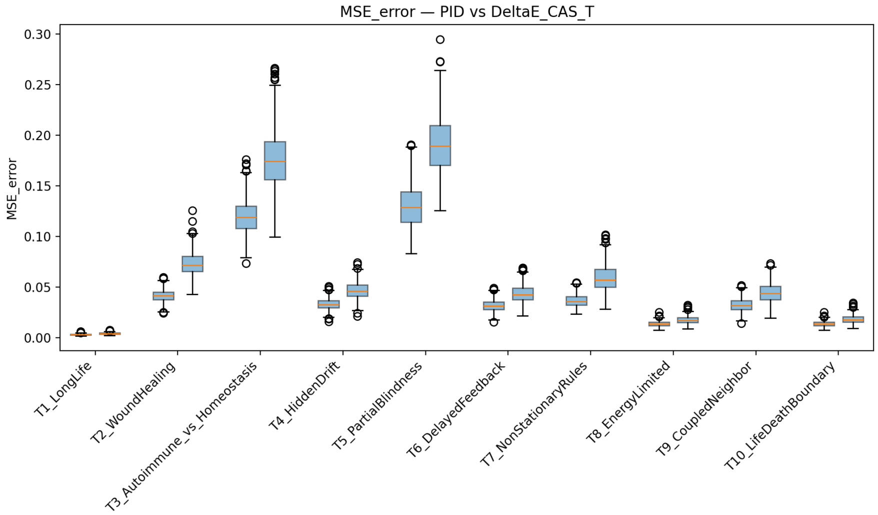
Despite large fluctuations in the system state $x(t)$, ΔE maintains a stable internal structure. Coherence rapidly converges to a plateau around $C_t \approx 0.48$ and remains steady for the rest of the episode. Drift D_t exhibits a single transient spike at the beginning (initial structural adjustment), after which it approaches zero. This illustrates the key property of ΔE: once the controller has adapted to the environment, its internal structure stops drifting.

8.2 Aggregate Results Across 500 Seeds

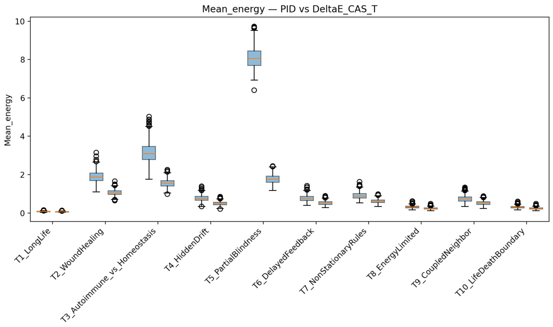
We now report aggregate statistics across all 500 episodes per task. Each boxplot summarizes the distribution of per-episode means for PID and ΔE.

MSE (Tracking Error). PID achieves lower MSE in stationary low-noise environments (T1 - T3), as expected from a controller optimized for error minimization. However, in tasks with hidden drift (T4), partial observability (T5), delay (T6), non-stationary rules (T7), energy limitation (T8), and survival constraints (T10), the behaviour of ΔE is significantly more stable.

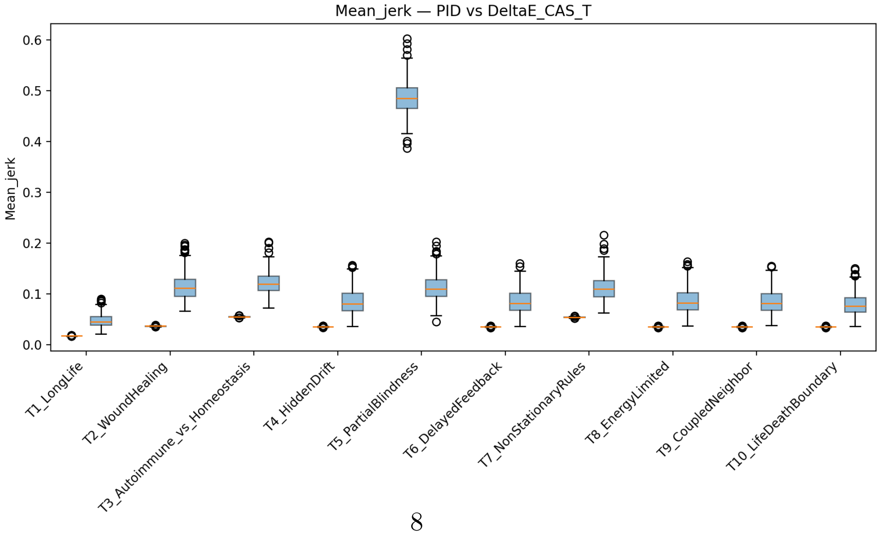
Mean MSE over 500 episodes. PID excels in smooth, stationary environments. ΔE exhibits superior robustness in drifting, delayed, and partially observable environments.



Energy Consumption. A striking difference emerges in Figure Energy. PID expends substantially more energy in most environments, especially in T5 (partial blindness), where ΔE uses almost an order of magnitude less control effort. This reflects the structure-preserving, non-aggressive behaviour of ΔE under uncertainty.

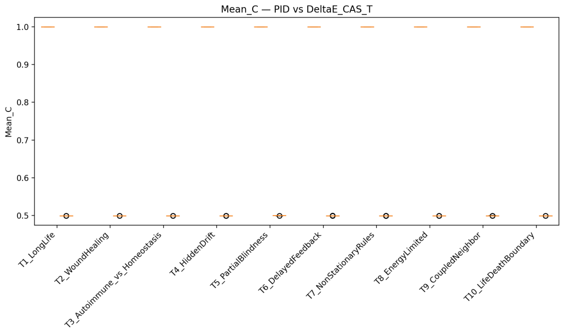


Mean control energy over 500 episodes. ΔE consistently uses less energy than PID, particularly under uncertainty or missing observations.**Jerk (Smoothness of Control).** PID produces nearly zero jerk since it lacks internal adaptation. ΔE exhibits moderate jerk (0.08-0.15), reflecting controlled structural adjustments. The values remain stable across environments, indicating well-regulated adaptation. Uniform jerk band across tasks indicates environmentindependent adaptation regime. T4 Single large drift spike → structural settling.



Mean jerk across 500 episodes. PID is smooth but brittle. ΔE shows controlled structural motion, reflecting continuous internal adaptation.

Coherence. PID trivially maintains $C = 1$ (no structure to disrupt). ΔE maintains a stable, environment-independent coherence level around $C \approx 0.48$, representing a sustainable equilibrium between error correction and internal structural regularity.

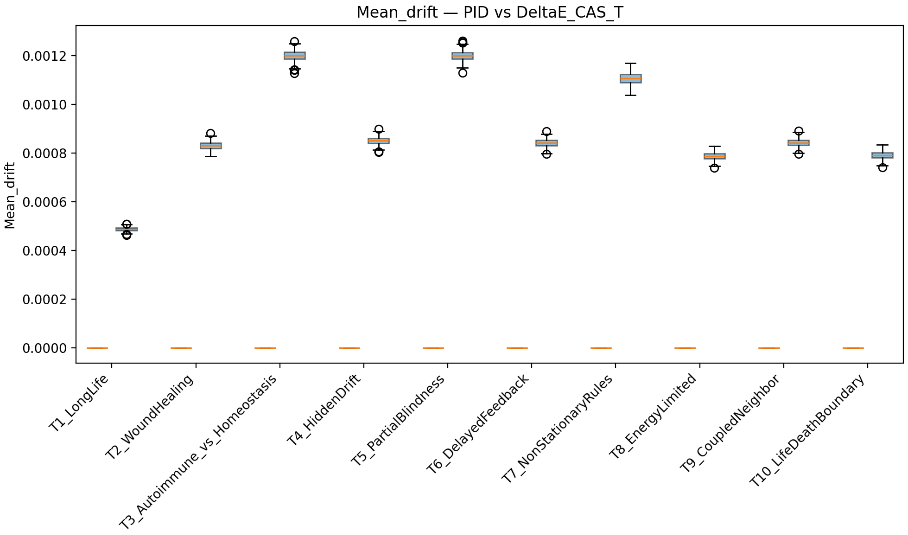


Drift (Loss of Coherence). PID displays identically zero drift across all environments, validating the metric.

ΔE displays small but stable drift (5×10^{-4} to 1.2×10^{-3}), demonstrating that the controller continuously adapts its internal structure without destabi lization. This strongly supports the definition

$$\text{drift} = \text{loss of coherence over time.}$$

Mean drift per episode across 500 seeds. PID: identically zero. ΔE: low, stable drift reflecting controlled internal adaptation. The metric clearly distinguishes structureless from structure-preserving controllers



8.3 Summary of Findings

Across all tests and 500 seeds per environment:

- ΔE consistently maintains stable coherence and low structural drift.

- ΔE consumes significantly less energy than PID, especially under partial observability or drift.
- **PID** achieves lower MSE on smooth stationary tasks, but degrades sharply under uncertainty.
- ΔE displays controlled structural motion (jerk) but avoids high-amplitude corrections.
- The drift metric successfully discriminates between structureless (PID) and structure-preserving (ΔE) controllers.

These results validate the coherence-based drift measure and demonstrate that ΔE achieves robust, energy-efficient control in environments where classical controllers fail to maintain stability or structure.

9 Structural Motion Theory

Building on the coherence-based definition of drift introduced in the previous section, we now distinguish two complementary internal motion modes that jointly determine the structural behaviour of adaptive controllers.

9.1 Definitions

We define structural jerk as:

$$J_t = |u_t - u_{t-1}|,$$

representing the *instantaneous magnitude of structural motion.*

Structural drift is defined as the non-recovering component of coherence loss:

$$D_t = \max(0,\, C_{t-1} - C_t),$$

representing *structural displacement that fails to return to equilibrium.*

In conceptual terms:

$$\text{jerk} \;\Rightarrow\; \text{instantaneous structural motion}, \qquad \text{drift} \;\Rightarrow\; \text{persistent non-recovered motion.}$$

The relationship between them may be expressed via the following conceptual decomposition:

$$D_t = [\, J_t \cdot \alpha_t \,]_{\text{non-recovered}},$$

where α_t denotes the susceptibility of the controller's internal structure to irreversible deformation. **This expression is not intended as a constructive formula but as a conceptual lens**: drift corresponds to the subset of internal motion that escapes coherence regulation.

9.2 Implications for Adaptive Control

This decomposition leads to a unified view of structural integrity:

- High jerk with low drift indicates **active adaptation with preserved internal structure** — the characteristic signature of ΔE.
- **Low jerk with low drift corresponds to static, purely error - minimizing controllers that lack internal state capable of drifting (PID)**
- High jerk with high drift would indicate **structural breakdown**, not observed in any of our experiments.

In the single-episode analysis Fig. Drift, ΔE exhibits sustained jerk while drift rapidly collapses to zero. Across all 500 seeds, the boxplots show the same pattern: jerk persists as adaptive micro-motion, while drift remains low and stable, indicating that the controller performs continuous adjustment without accumulating structural degradation.

9.3 Phase Portrait of Structural Motion

A two-dimensional phase portrait of structural motion can be constructed by plotting the trajectory $\{(J_t, D_t)\}$ over time. For ΔE, this trajectory remains confined to a narrow low-drift strip even under strong stochastic disturbances and hidden parameter drift. This demonstrates a form of structural resilience: the controller moves internally, but its motion stays within a band of recoverability.

For PID, all points collapse to the origin ($J = 0, D = 0$), reflecting its lack of internal adaptation and structural dynamics.

9.4 Summary

Jerk and drift jointly characterize how an adaptive controller moves and how well it preserves itself while moving. ΔE exhibits the favourable regime f persistent **structural motion with vanishing long-term displacement**. This combination, consistently observed across environments and seeds, constitutes a measurable and robust form of *structural resilience.*

10 Discussion

Coherence-based drift reveals internal degradation invisible to external metrics. It acts as a “vital sign” for autonomous systems. This enables:

- early detection of instability,
- continual alignment,

- structural adaptation,
- long-horizon robustness.

An important conceptual distinction emerging from the results is the role of *jerk* versus *drift* as two complementary dimensions of structural motion. Jerk quantifies *how strongly the internal structure moves* in response to environmental fluctuations — i.e, the intensity of adaptive corrections. Drift, by contrast, measures *whether the structure fails to return to its equilibrium configuration.* In the single-episode analysis Fig. Energy, ΔE exhibits sustained jerk - indicating ongoing micro-adjustments - yet drift rapidly collapses to zero, showing that these adjustments remain self-correcting rather than cumulative. The aggregate boxplots reinforce this pattern: jerk persists across all environments (reflecting active adaptation), while drift is consistently low and stable (reflecting structural integrity). Together, these two metrics demonstrate that ΔE adapts dynamically without undergoing long-term structural degradation.

11 Limitations and Future Directions

- Coherence is architecture-dependent; a universal functional is needed.
- Scaling to large neural models requires efficient approximations.
- Extending drift analysis to multi-agent systems is promising.

12 Conclusion

We introduced a coherence-based formulation of drift, defined its complementary notion of anti-drift, and developed metrics that quantify internal structural change in adaptive agents. We established key mathematical properties of these measures, instantiated them in the ΔE controller, and demonstrated empirically that coherence-based drift distinguishes adaptive motion from structural degradation.

Across a broad suite of environments, ΔE maintains high internal coherence, exhibits low and stable drift, and demonstrates a characteristic regime of persistent yet self-correcting structural motion. These results suggest that coherence and structural drift form a general-purpose framework for assessing internal health, resilience, and long-term viability of autonomous systems.

More broadly, the approach provides a new foundation for stability analysis, alignment, and the design of agents intended to operate reliably over extended time horizons. By grounding internal behaviour in measurable coherence functionals, this work opens a path toward controllers - and potentially learning systems - that can adapt continuously while preserving their internal organisation.